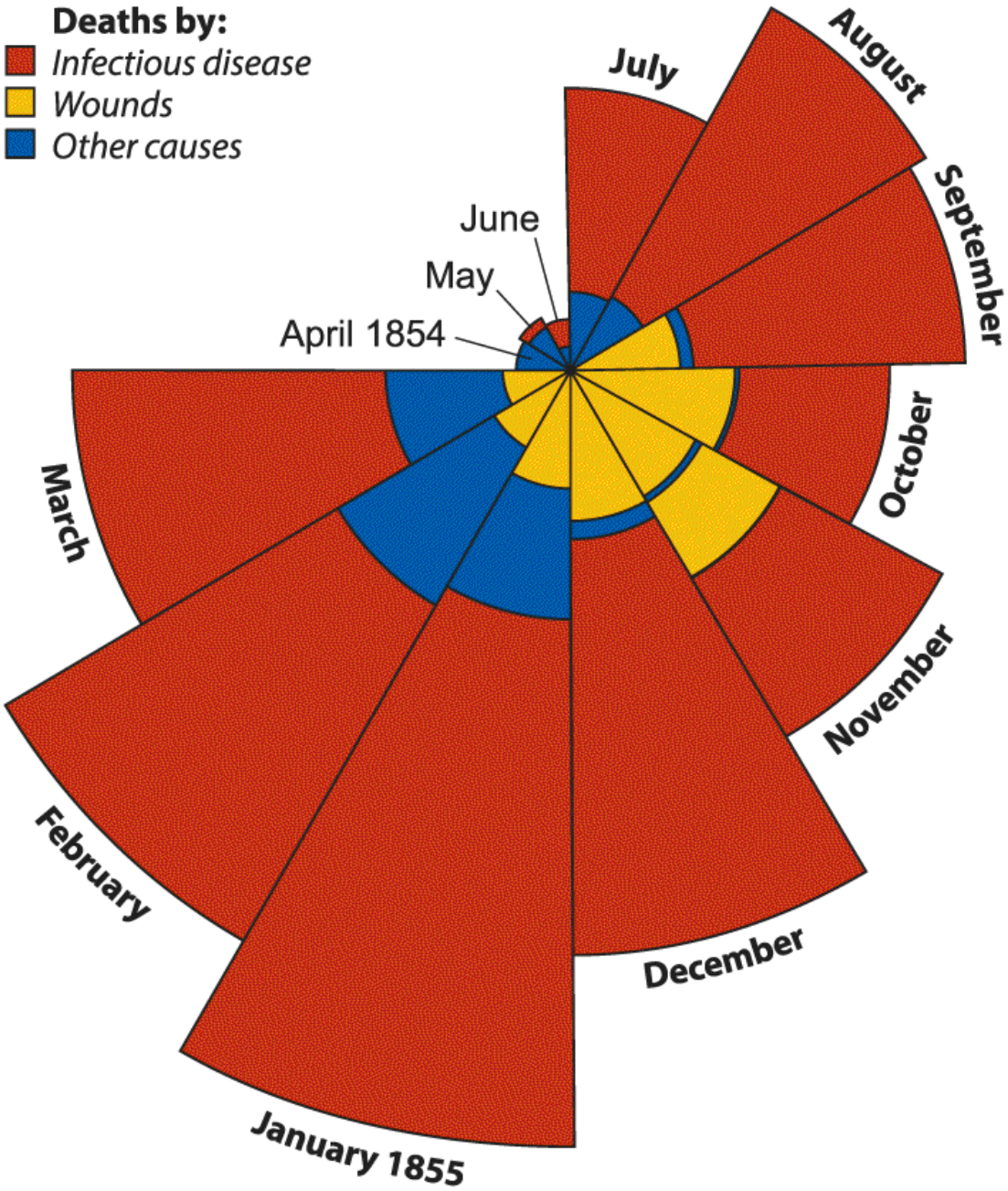
Chapter 2 Displaying data



Whitlock & Schluter, *The Analysis of Biological Data*, 3e

Description

Data from the chart is as follows:

April 18 54: All deaths are caused by other causes.

May: Some deaths are caused by infectious disease and majority of deaths are caused by other causes.

June: Some deaths are caused by other causes and majority of deaths are caused by infectious disease.

July: Some deaths are caused by other causes and majority of deaths are caused by infectious disease.

August: Some deaths are caused by other causes and majority of deaths are caused by infectious disease.

September: Majority of deaths is caused by infectious disease, some are caused by wounds, and few are caused by other causes.

October: Majority of deaths is caused by infectious disease, some are caused by wounds, and a few are caused by other causes.

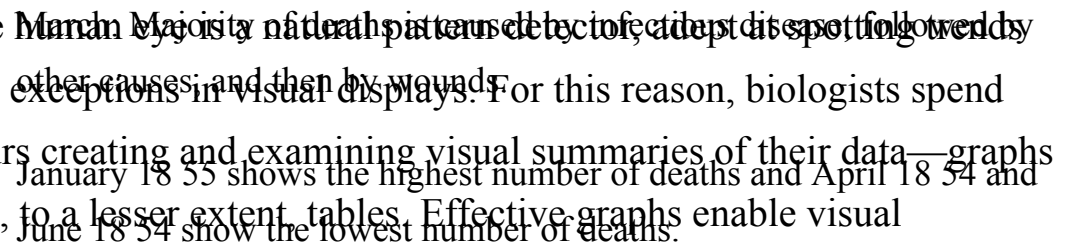
November: A few deaths are caused by other causes and almost equal deaths are caused by infectious disease and wounds.

December: Majority of deaths is caused by infectious disease, some are caused by wounds, and few are caused by other causes.

January 18 55: Majority of deaths is caused by infectious disease, followed by other causes, and then by wounds.

February: Majority of deaths is caused by infectious disease, followed by other causes, and then by wounds.

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 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 show. 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 then review which types of graphs best show the data. The top choices 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, whose layouts should also be optimized to show patterns in the data.

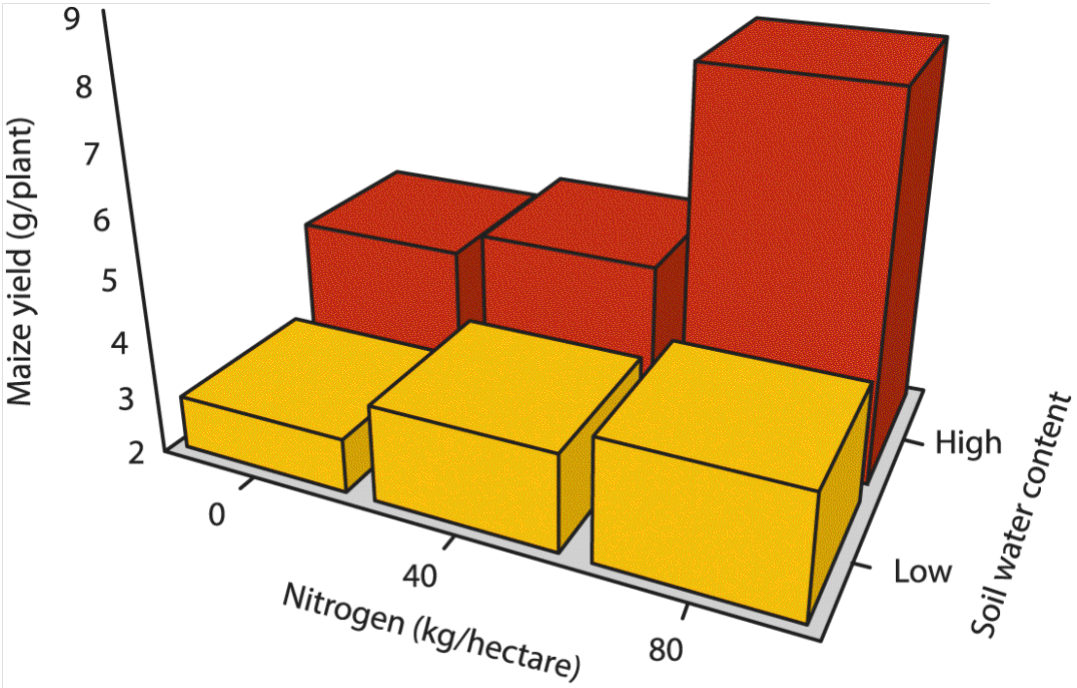
2.1 Guidelines for effective graphs

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. These two purposes, analysis and presentation, are largely coincident because the most revealing displays will be the best for both identifying patterns in the data and 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 graph intentionally bad to highlight four common defects. Examine the graph before reading farther and try to recognize some of the errors. 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.



Whitlock & Schluter, *The Analysis of Biological Data*, 3e © 2020 W. H. Freeman and Company

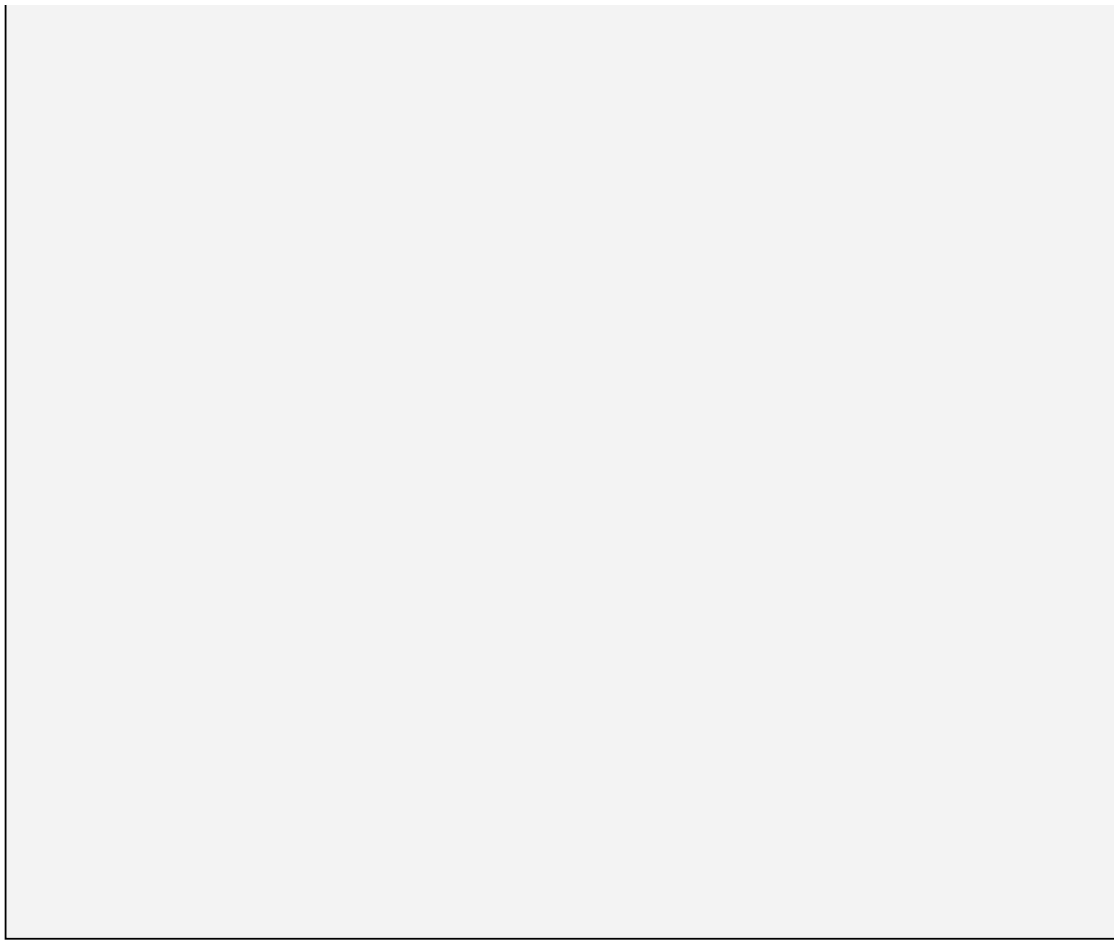
FIGURE 2.1-1

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

Description

The horizontal axis is labeled Nitrogen in kilogram per hectare, with points 0, 40, and 80. The vertical axis is labeled Maize yield in gram per plant, with points ranging from 2 to 9 with increment of 1. Another horizontal axis is labeled Soil water content, with qualitative data low and high. 6 rectangular bars in 2 rows of low and high are plotted.

The approximate data are as follows. 0 kilogram per hectare nitrogen, 3 gram per plant maize yield, low soil water content; 40, 4, low; 80, 5, low; 0, 6, high; 40, six point five, high; 80, 9, high.



Mistake #1: The graph hides 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. As a result, 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 style of 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 ensure 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). In other words, show the individual data points or a frequency distribution to show their shape. A good graph allows you to see the data to help the eye detect patterns. 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 skew and outliers, 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 the 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.

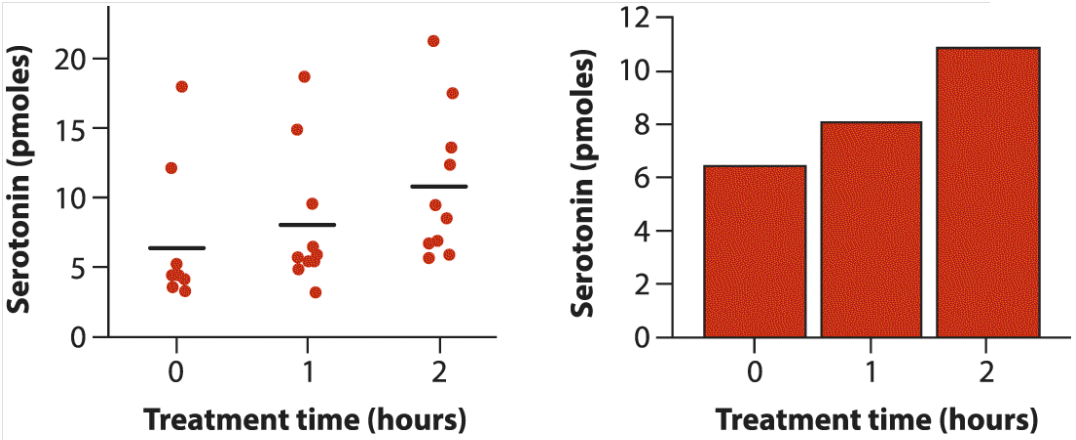


FIGURE 2.1-2

A graph that shows the data (*left*) and a graph that hides the same data (*right*). 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.

Description

The horizontal axis is labeled Treatment time in hours, with points 0, 1, and 2. The vertical axis is labeled Serotonin in pmoles, with points 0, 5, 10, 15, and 20.

The approximate data in the scatter plot are as follows. The clusters of points are most dense between (0, 3) and (0, 5) when treatment time is 0; (1, 3) and (1, 7) when treatment time is 1; (2, 5) and (2, 10) when treatment time is 2.

The approximate data in the bar chart are as follows. 0 treatment hours, 6 serotonin (pmoles); 1, 8; 2, 11.

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 of 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 communicate the findings most clearly. 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 chartjunk, which tend to obscure the patterns in the data. In the rest of this chapter, we'll compare alternative ways of graphing the same data sets to test 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 the details 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, depicting government spending on

education each year since 1998 in British Columbia, shows an example. 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 (the revised graph also removed the 3-D effects and the numbers above the bars to make the pattern easier to see).

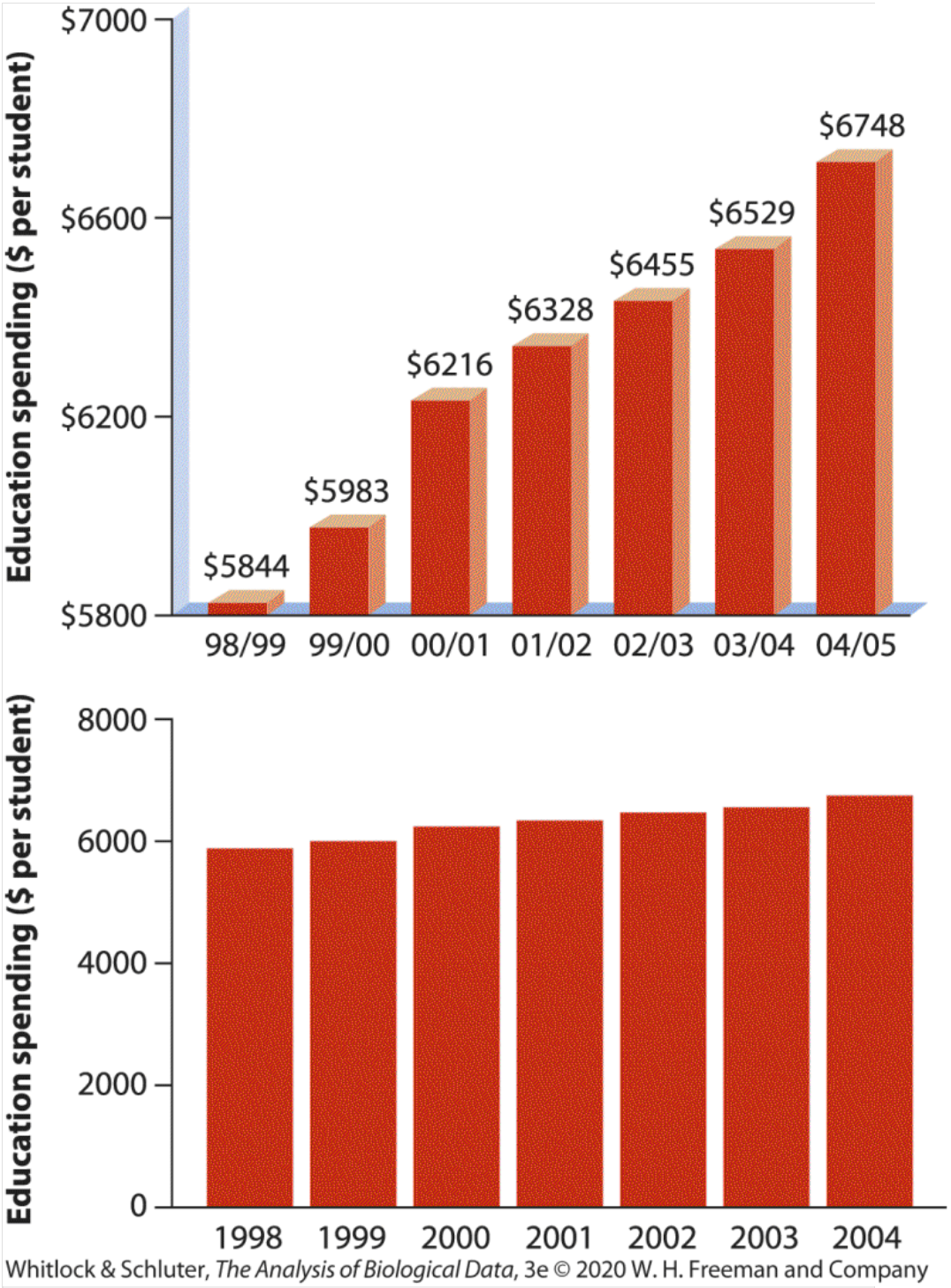


FIGURE 2.1-3

Upper graph: A bar graph, taken from data from a British Columbia government brochure, indicating 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 the bars

to make the pattern easier to see. The upper graph is modified from data from British Columbia Ministry of Education (2004).

Description

The first bar chart shows the horizontal axis with years 98-99, 99-00, 00-01, 01-02, 02-03, 03-04, 04-05. The vertical axis is labeled Education spending in dollars per student, with points 5800, 6200, 6600, and 7000.

The data in the first bar chart are as follows. 98-99 year, 5844 dollars; 99-00, 5983; 00-01, 6216; 01-02, 6328; 02-03, 6455; 03-04, 6529; 04-05, 6748.

The second bar chart shows the horizontal axis with years 1998, 1999, 2000, 2001, 2002, 2003, 2004. The vertical axis is labeled Education spending in dollars per student, with points ranging from 0 to 8000 with increments of 2000. The data in the second bar chart are as follows. 1998 year, 5844 dollars; 1999, 5983; 2000, 6216; 2001, 6328; 2002, 6455; 2003, 6529; 2004, 6748.

Graphs without bars, 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.

Next, we show graphical methods for the most common types of biological data and show how specific features can ensure that they meet these four principles of good graph design. Typically, more than one graph type is available for displaying a frequency distribution or showing an association between variables. Try more than one, and decide which is best according to which shows the pattern most clearly.

2.2 Showing data for one variable

To visualize data for a single variable, 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 gives the number of occurrences for 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.

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

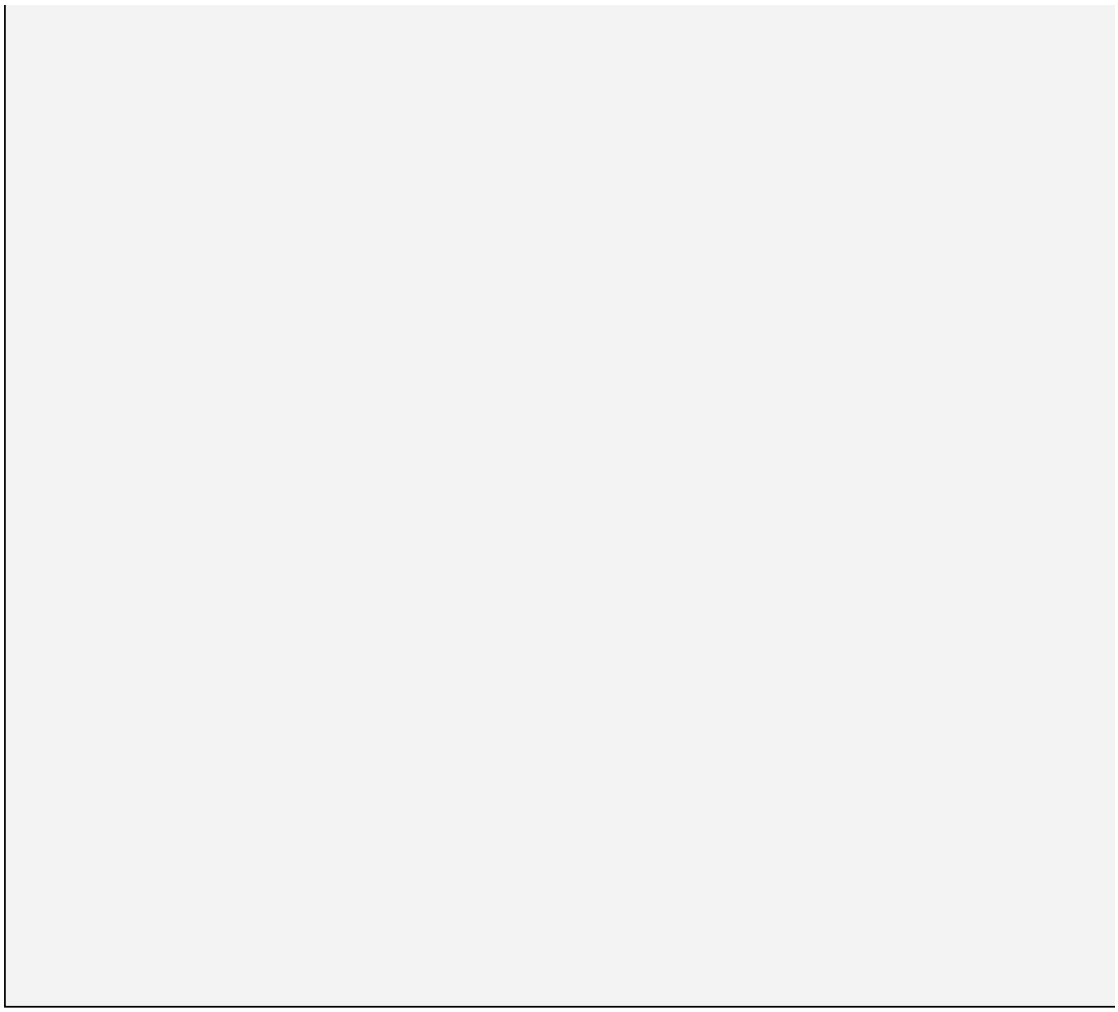
Example 2.2A illustrates both kinds of displays.

EXAMPLE 2.2A: Crouching tiger

kevdog818/Getty Images

Description

-



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 2.2-1 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 livestockCollecting grass or fodder for livestock	4444
Collecting non-timber forest productsCollecting non-timber forest products	1111
FishingFishing	88
Herding livestockHerding livestock	77
Disturbing tiger at its killDisturbing tiger at its kill	55
Collecting fuelwood or timberCollecting fuelwood or timber	55
Sleeping in a houseSleeping in a house	33
Walking in forestWalking in forest	33
Using an outside toiletUsing an outside toilet	22
TotalTotal	8888

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.

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 spent 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.

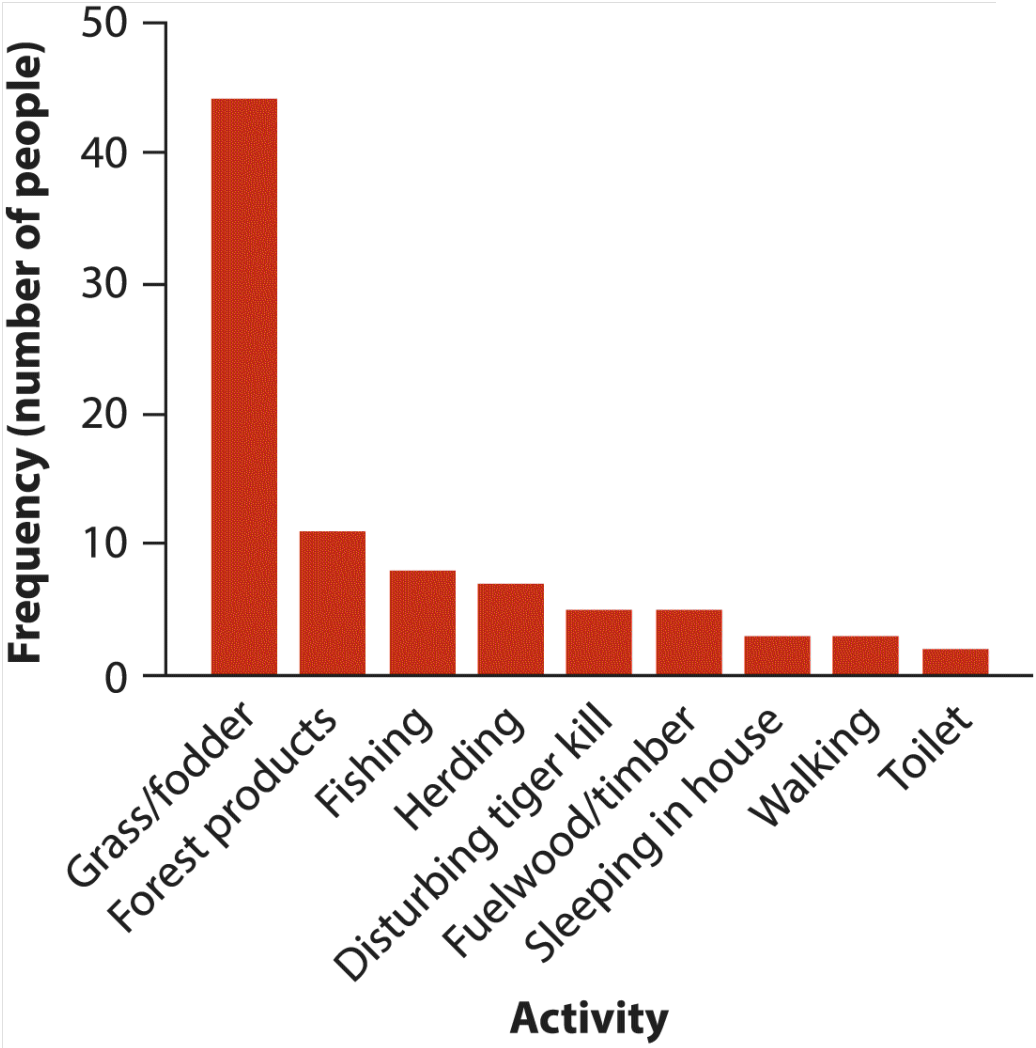


FIGURE 2.2-1

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.$n = 88.$ The frequencies are taken from Table 2.2-1, which also gives more detailed labels of activities.

Description

The horizontal axis is labeled Activity, with data Grass or fodder, Forest products, Fishing, Herding, Disturbing tiger kill, Fuelwood of timber, Sleeping

in house, Walking, and Toilet. The vertical axis is labeled Frequency, n number of people, ranging from 0 to 50 with increments of 10. The approximate data are as follows. Grass/fodder, 44; Forest products, 11; Fishing, 8; Herding, 7; Disturbing tiger kill, 5; Fuelwood/timber, 5; Sleeping in house, 3; Walking, 3; Toilet, 2.

Making a good bar graph

A bar graph shows the data by illustrating frequencies in each group. The eye compares the areas of the bars, which must therefore be of equal width. It is crucial that the baseline of the y-axis$^{y\text{-axis}}$ 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 order categories in the graph 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 a 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)$^{(n)}$ 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.

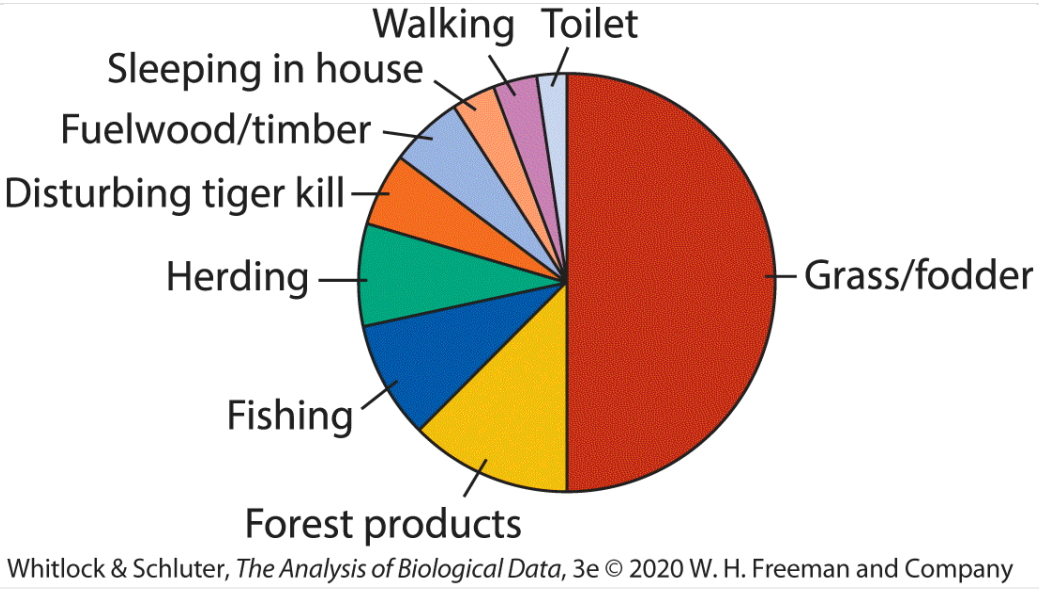
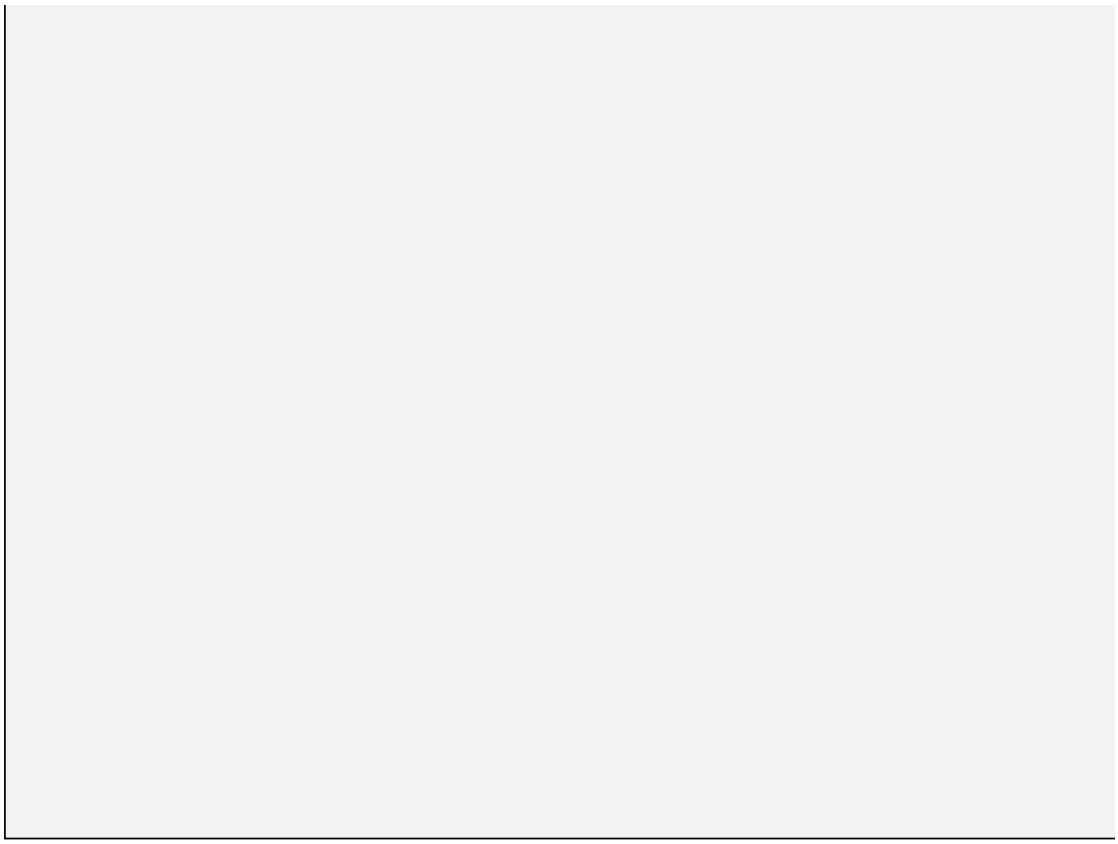


FIGURE 2.2-2

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.$n = 88.$

Description

-



The pie chart has received a lot of criticism from experts in information graphics. One reason is that the eye has a more difficult time comparing frequencies of different groups, compared to a bar graph. 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. We suggest that the bar graph be the first choice when showing frequencies in categorical data.

Showing numerical data: frequency table and histogram

To show the data for a single numerical variable, use a frequency table or a **histogram**. A histogram uses the 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 using Example 2.2B.

EXAMPLE 2.2B: Effects of Zika virus infection on fetuses

The Zika virus can be spread via mosquitoes, sexual contact, or from mother to fetus. In 2015, there was an outbreak in Brazil that then spread to other countries in the Americas. Small head size (microcephaly) associated with abnormal brain development was frequently reported in newborn babies of infected mothers. The following data are 40 measurements of head width (biparietal diameters, in mm) of fetuses in a sample of pregnant women infected with the virus. The measurements were obtained using ultrasound at 33 to 36 weeks of gestational age at a clinic in Rio de Janeiro, Brazil (Brasil et al. 2016).

---

61 69 70 80 80 80 81 82 82 82 82
83 84 84 84 84 85 85 85 85 85 85
85 85 86 86 86 86 87 87 87 88 88
89 89 89 90 90 90 92

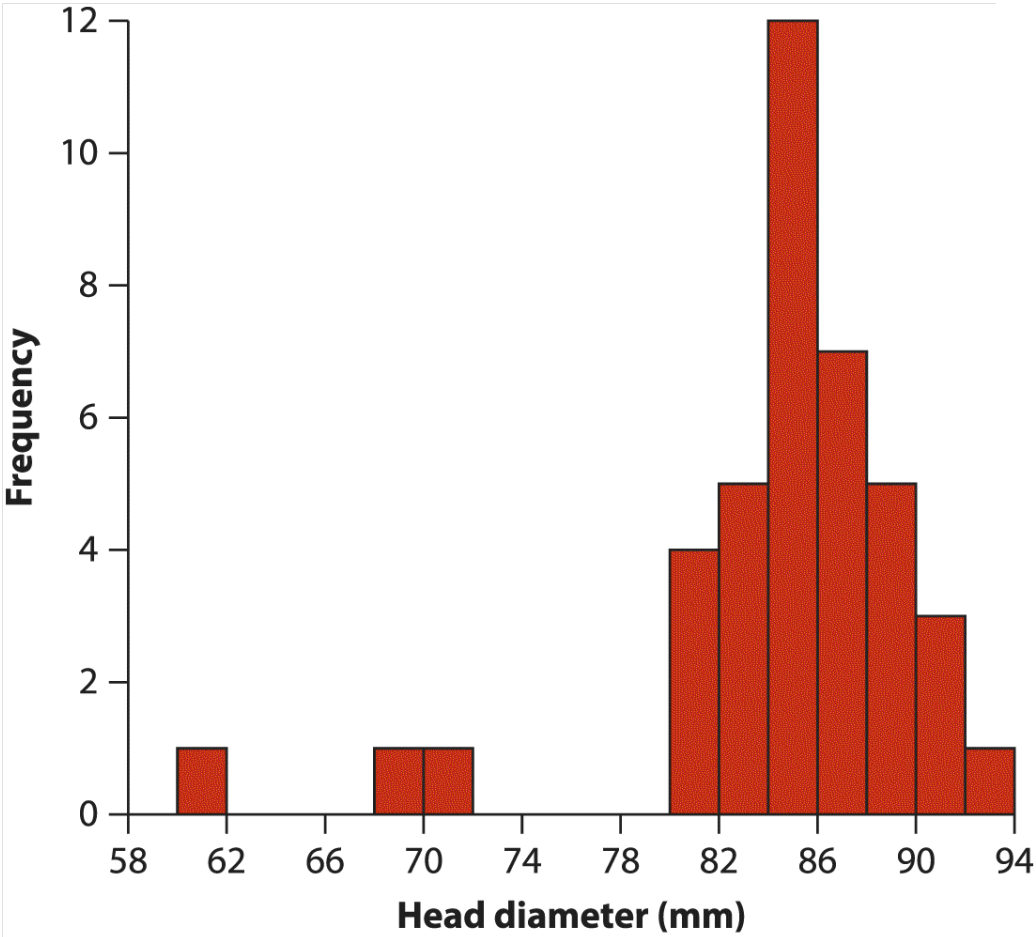
We treated each fetus in the study between 33 and 36 weeks of gestational age as the unit of interest and head width as the measurement. The range of head-width values was divided into 7 intervals of equal size (60–65, 65–70, and so on). The number of individual fetuses within each head-width interval

was counted and presented in a frequency table to help see patterns (Table 2.2-2).

TABLE 2.2-2 Frequency distribution of head widths of fetuses of pregnant women infected with the Zika virus.

Head width (mm)	Frequency (number of fetuses)
60–65	1
65–70	1
70–75	1
75–80	0
80–85	13
85–90	20
90–95	4
Total	40

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 fetuses) in each head-width interval is perceived as bar area. The histogram shows that the shape of the frequency distribution is not symmetric but has a long "tail" extending to the left. The majority of fetuses had a head width between 80 and 94 mm, but three had substantially smaller heads. All three fall outside the norm for fetuses of equivalent age from uninfected mothers, and these are considered microcephalic (Brasil et al. 2016).



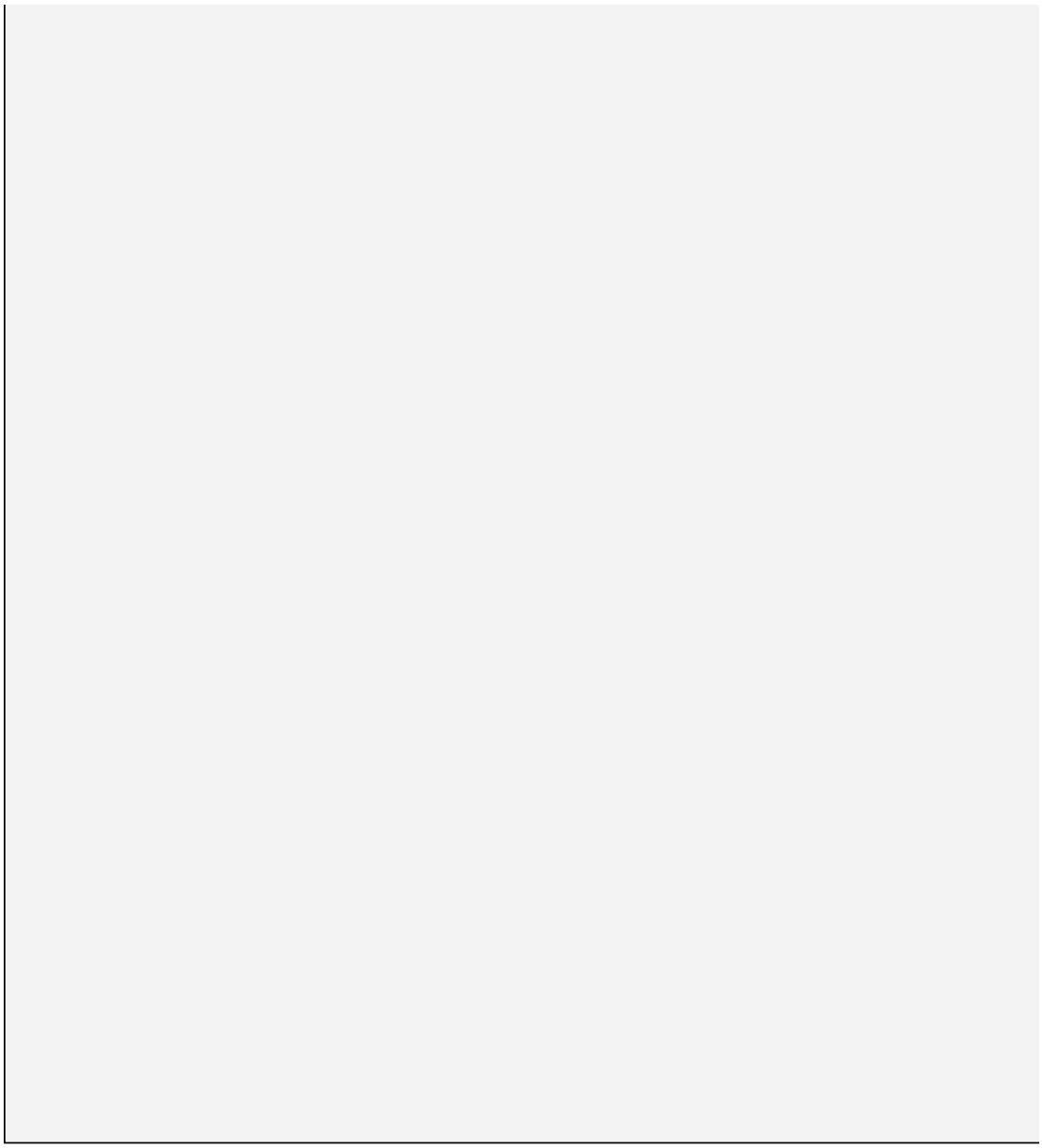
Whitlock & Schluter, *The Analysis of Biological Data*, 3e © 2020 W. H. Freeman and Company

FIGURE 2.2-3

Histogram illustrating the frequency distribution of head sizes of fetuses of pregnant mothers infected with the Zika virus. Total number of individuals: n=40.$n = 40.$

Description

The horizontal axis represents head diameter in millimeters from 58 to 94 in intervals of 4 millimeters. The vertical axis represents frequency from 0 to 12 in intervals of 2. The approximate data from the histogram are as follows: Head diameter 60-62, frequency 1; Head diameter 68-70, frequency 1; Head diameter 70-72, frequency 1; Head diameter 80-82, frequency 4; Head diameter 82-84, frequency 5; Head diameter 84-86, frequency 12; Head diameter 86-88, frequency 7; Head diameter 88-90, frequency 5; Head diameter 90-92, frequency 3; Head diameter 92-94, frequency 1

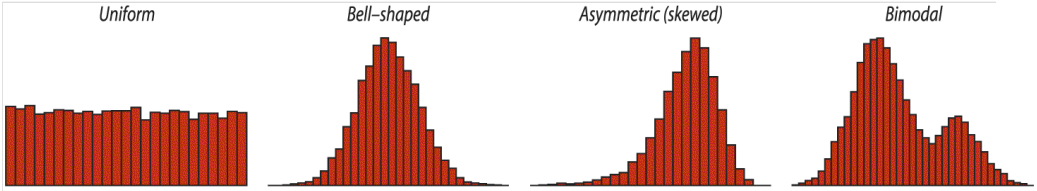


Describing the shape of a histogram

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.



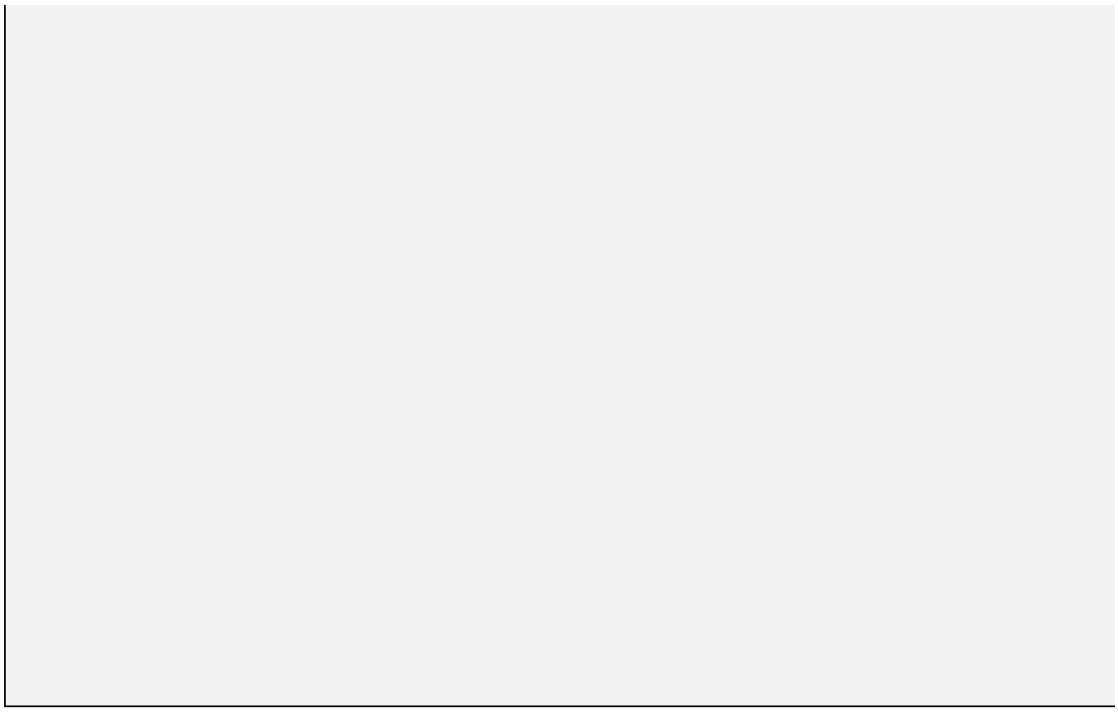
Whitlock & Schluter, *The Analysis of Biological Data*, 3e © 2020 W. H. Freeman and Company

FIGURE 2.2-4

Some possible shapes of frequency distributions.

Description

-



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.[3] The frequency distribution of head widths of fetuses of Zika-infected mothers has negative skew (Figure 2.2-3).

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 fetus head widths (Figure 2.2-3) includes at least one and possibly three extreme observations that fall outside the range of values of most of the sample. Although microcephaly is unusual, outliers are common in biological data generally. Outliers sometimes result from

mistakes in recording the data, in which case they might be removed from the data set. Or, as in the case of the fetus data, outliers may represent real observations from nature and should not be dropped from the data. Outliers should always be investigated so that their cause may be determined.

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

How to draw a good histogram

The choice of interval width must be made carefully in a histogram, 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.

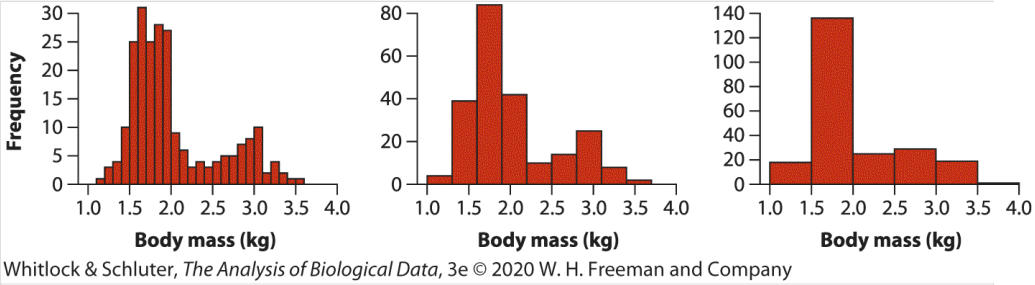


FIGURE 2.2-5

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*).

Description

The horizontal axis is labeled Body mass in kilogram, ranging from 1 to 4 with increments of 0 point 5. The vertical axis is labeled Frequency. The approximate data in the first plot are as follows. 1 point 1 to 1 point 2, 1; 1 point 2 to 1 point 3, 3; 1 point 3 to 1 point 4, 4; 1 point 4 to 1 point 5, 10; 1 point 5 to 1 point 6, 25; 1 point 6 to 1 point 7, 31; 1 point 7 to 1 point 8, 25; 1 point 8 to 1 point 9, 27; 1 point 9 to 2, 26; 2 to 2 point 1, 8; 2 point 1 to 2 point 2, 5; 2 point 2 to 2 point three, 3; 2 point three to 2 point 4, 4; 2 point 4 to 2 point 5, 3; 2 point 5 to 2 point 6, 4; 2 point 6 to 2 point 7, 5; 2 point 7 to 2 point 8, 5; 2 point 8 to 2 point 9, 7; 2 point 9 to 3, 8; 3 to 3 point 1, 10; 3 point 1 to 3 point 2, 2; 3 point 2 to 3 point 3, 4; 3 point 3 to 3 point 4, 2; 3 point 4 to 3 point 5, 1; 3 point 5 to 3 point 6, 1.

The approximate data in the second plot are as follows. 1 to 1 point 3, 2; 1 point 3 to 1 point 6, 40; 1 point 6 to 1 point 9, 82; 1 point 9 to 2 point 2, 42; 2 point 2 to 2 point 5, 10; 2 point 5 to 2 point 8, 14; 2 point 8 to 3 point 1, 24; 3 point 1 to 3 point 4, 8; 3 point 4 to 3 point 7, 1.

The approximate data in the third plot are as follows. 1 to 1 point 5, 17; 1 point 5 to 2, 138; 2 to 2 point 5, 25; 2 point 5 to 3, 30; 3 to 3 point 5, 18; 3 point 5 to 4, 2.

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.[4]

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. We use "left closed" intervals, which means that the value 70 falls in the interval 70–72 rather than in the interval 68–70 (Table 2.2-2).

There are no strict rules about the number of intervals to use 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),$1 + \ln(n)/\ln(2)$, where nn 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. Try several alternatives to 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 *violin plot* and the *strip chart* are alternatives, but these are used most often to show differences when there are data from

two or more groups. We describe these graphs in the next section. Two other types of graph, the *box plot* and the *cumulative frequency distribution*, are explained in Chapter 3.

2.3 Showing association between two variables and differences between groups

Here we illustrate how to use data to show associations between two variables and differences between groups. The most suitable type of graph depends on whether the two variables are categorical, numerical, or one of each data type.

Showing association between categorical variables

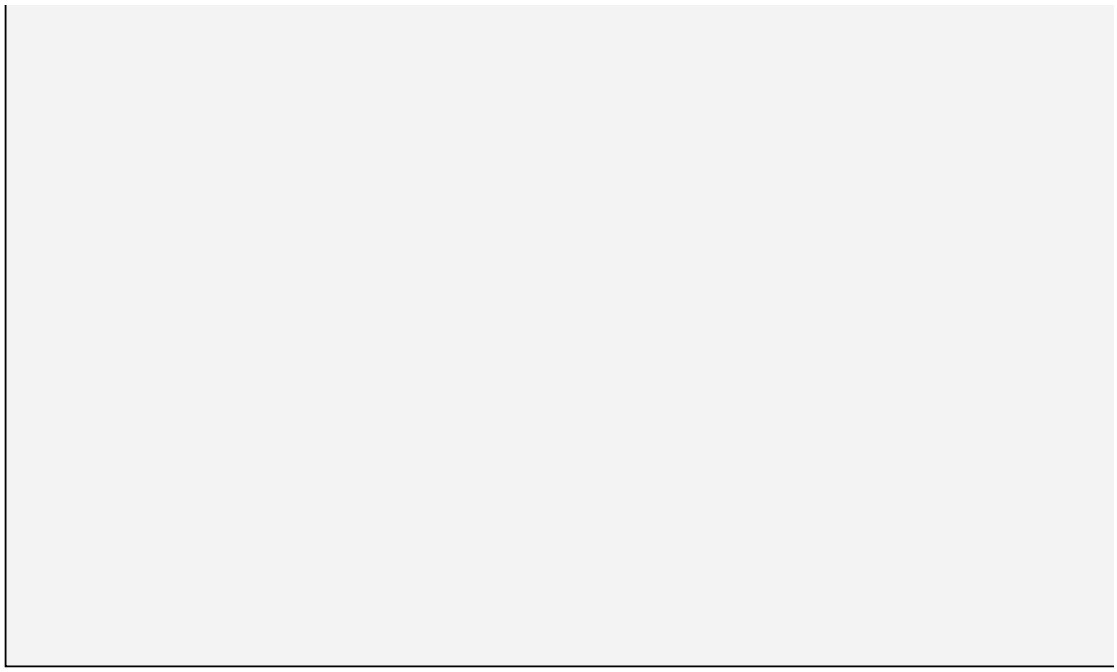
If two categorical variables are associated, the relative frequencies for one variable will differ among categories of the other variable. To reveal such association, show the frequencies using a contingency table, a mosaic plot, or a grouped bar graph. Here's an example.

EXAMPLE 2.3A: Reproductive effort and avian malaria

Karel Gallas/ Shutterstock

Description

-



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 in each attempt. Oppliger et al. (1996) investigated the impact of reproductive effort on the susceptibility to malaria[5] 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

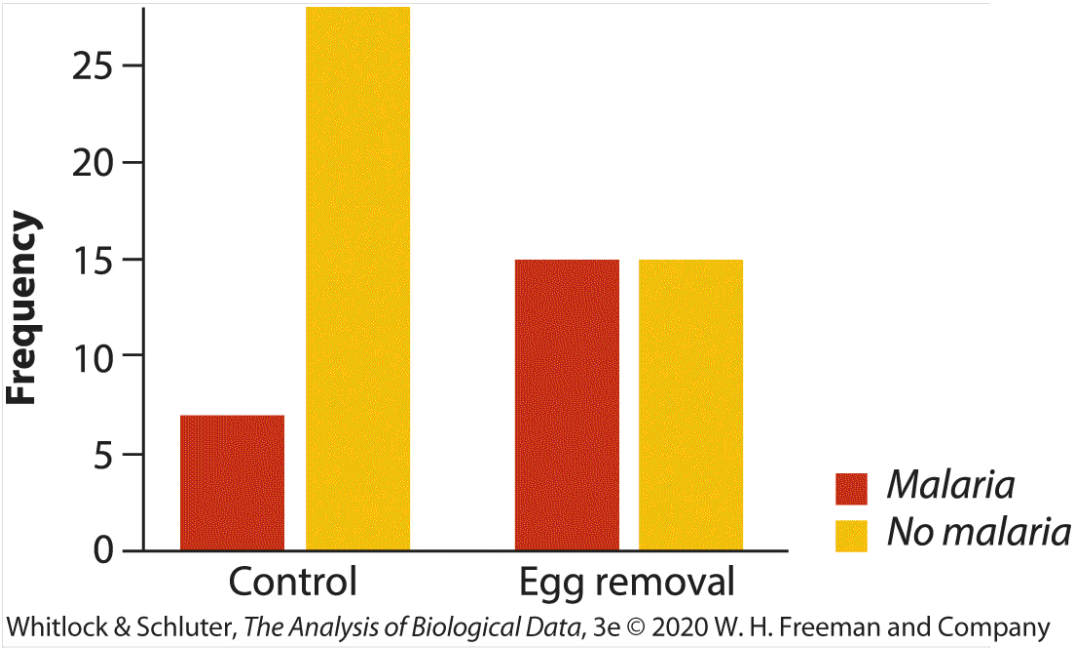
Each experimental unit (bird) is counted exactly once in the four main "cells" of Table 2.3-1, and so the total count (65) is the number of birds in the study. A cell is one combination of categories of the row and column variables in the table. The explanatory variable (experimental treatment) is displayed 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 7 of the 35 control birds. This difference between treatments 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 well 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 the bars for no malaria differ between treatments. Most birds in the control group had no malaria (the gold bar is much taller than the red bar), whereas in the experimental group, the frequency of subjects with and without malaria was equal.



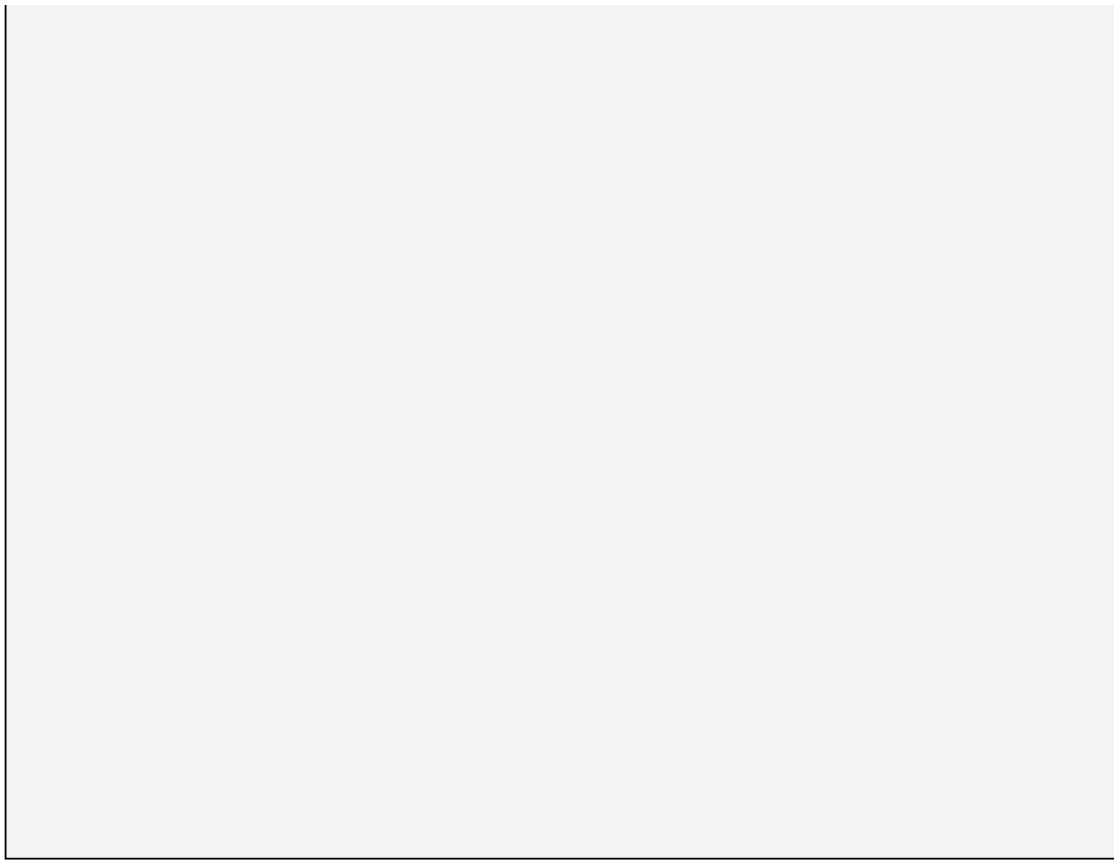
Whitlock & Schluter, *The Analysis of Biological Data*, 3e © 2020 W. H. Freeman and Company

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$n = 65$ birds.

Description

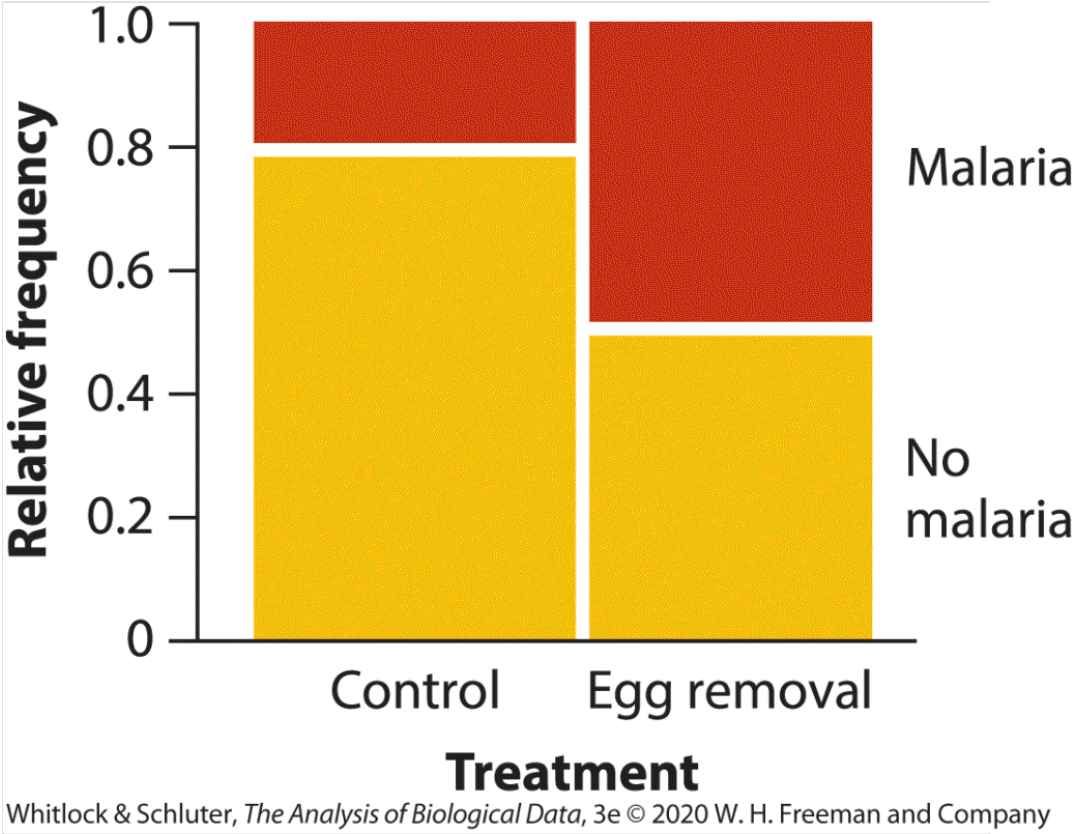
The horizontal axis is shown with data Control and Egg removal. The vertical axis is labeled Frequency, ranging from 0 to 25 with increments of 5. The data are as follows. Control, Malaria, frequency 7; Control, No malaria, 28; Egg removal, Malaria, 15; Egg removal, No malaria, 15.



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 gold bar (no

malaria) at a higher vertical position than in the egg removal stack, where the incidence of malaria was greater.



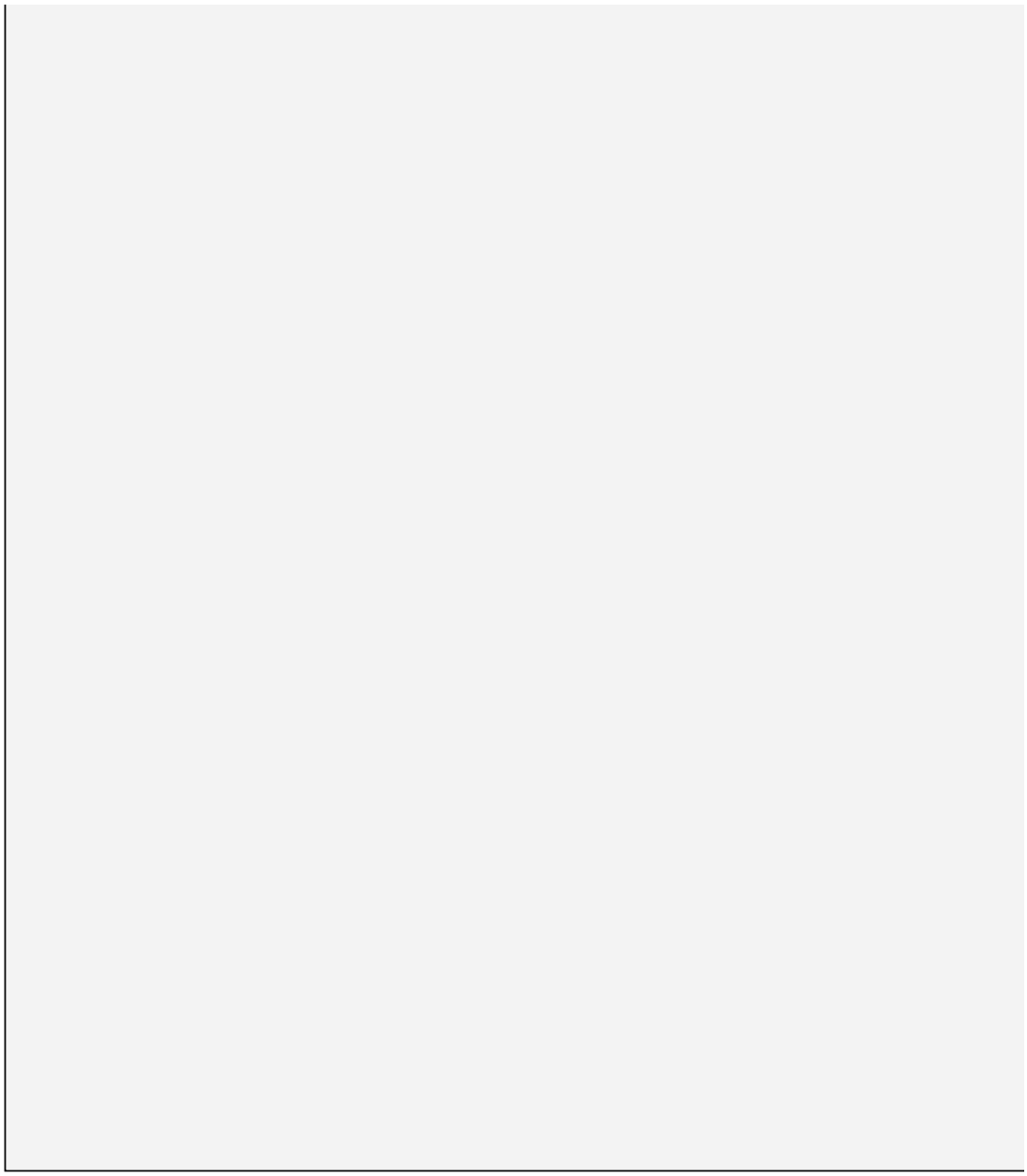
Whitlock & Schluter, *The Analysis of Biological Data*, 3e © 2020 W. H. Freeman and Company

FIGURE 2.3-2

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

Description

The horizontal axis is labeled Treatment, with data Control and Egg removal. The vertical axis is labeled Relative Frequency, ranging from 0 to 1 with increments of zero point two. The data are as follows. Control, No Malaria, relative frequency 0 to zero point eight; Control, Malaria, zero point eight to 1; Egg removal, No Malaria, 0 to zero point five; Egg removal, Malaria, zero point five to 1.



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, such as how different the frequencies are in the different categories, and it is a good idea to try all three. Choose the method that shows the pattern in the data most effectively. 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.

Showing association between numerical variables: scatter plot

You are probably already familiar with the scatter plot, which is used to show association between two numerical variables. The position along the horizontal axis (the x-axis$x$-axis) indicates the measurement of the explanatory variable. The position along the vertical axis (the y-axis$y$-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

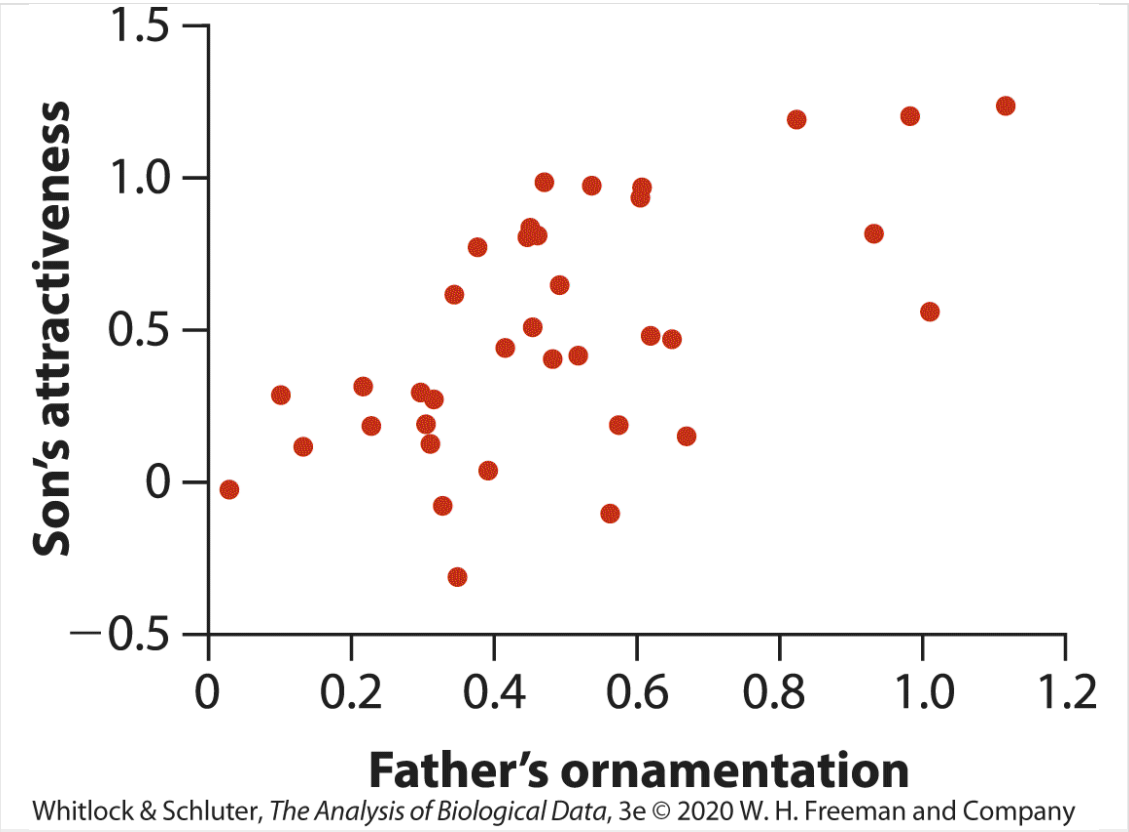
left to the upper right of the graph), negative (the points run from the upper left to the lower right), or absent (no discernible pattern).

Kimberly Hughes

Description

-

For example, Brooks (2000) 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) in 36 father-son pairs. The father's ornamentation is the explanatory variable in the resulting scatter plot of these data (Figure 2.3-3).



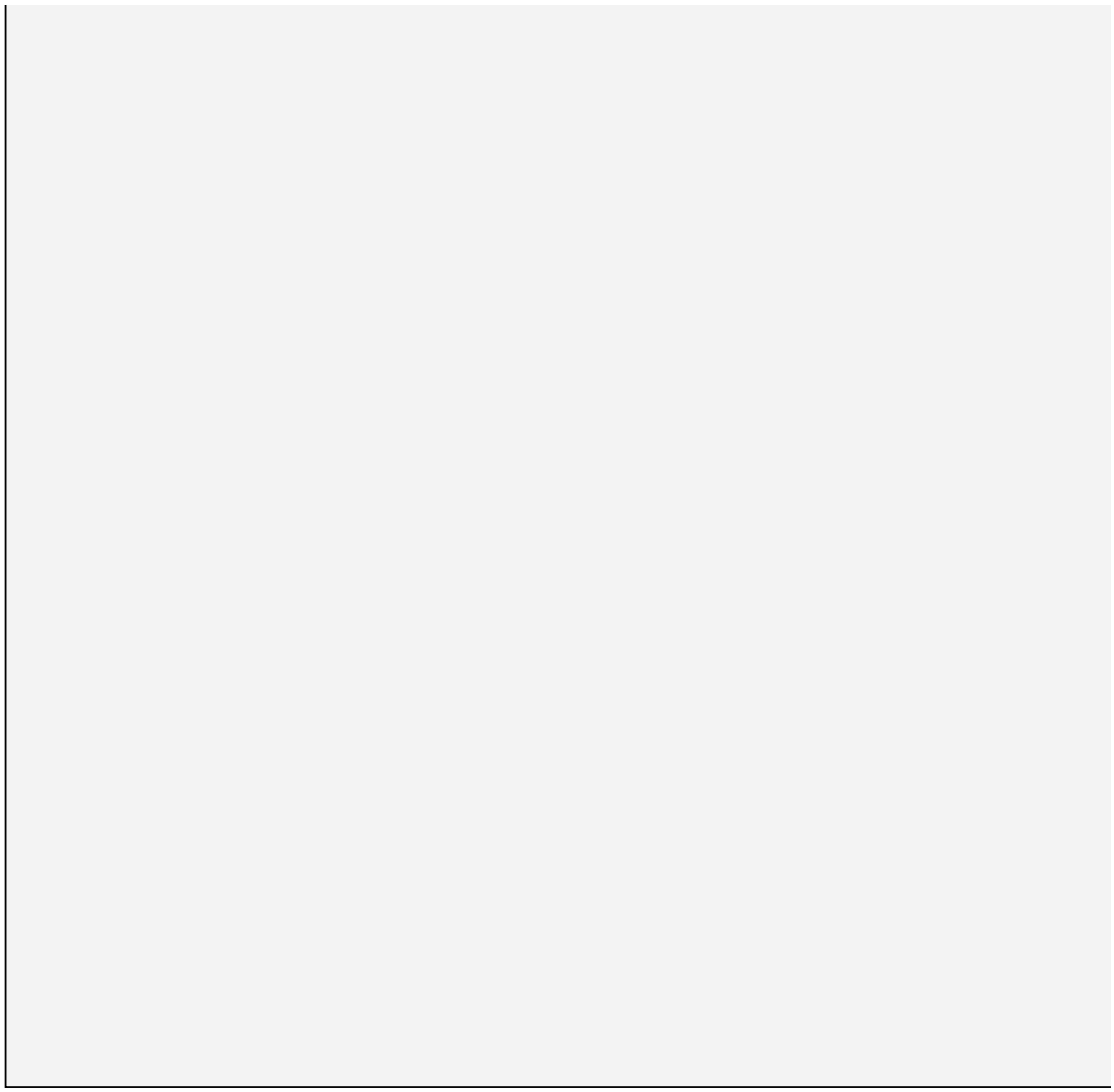
Whitlock & Schluter, *The Analysis of Biological Data*, 3e © 2020 W. H. Freeman and Company

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$n = 36$ father-son pairs.

Description

The horizontal axis is labeled Father's ornamentation, ranging from 0 to 1 point 2 with increments of 0 point 2. The vertical axis is labeled Son's attractiveness, ranging from negative 0 point 5 to 1 point 5 with increments of 0 point 5. All data in the graph are approximate. The cluster of points is most dense between (0, 0) and (0 point 7, 1).



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.

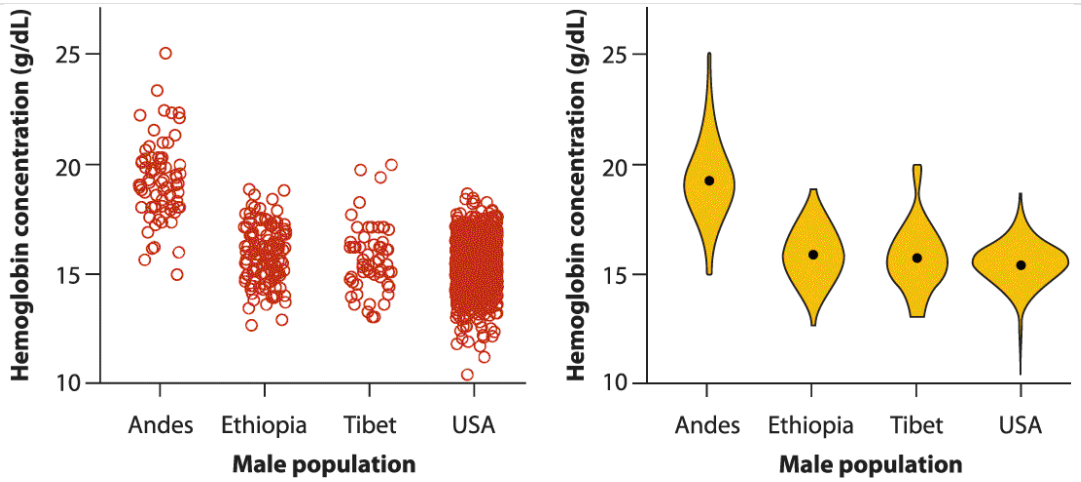
A *scatter plot* is a graphical display of two numerical variables in which each observation is represented as a point on a graph with two axes.

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. (Equivalently, these methods show the differences in the numerical variable between the categories, or groups.) Three that we recommend are the *strip chart* (which we first saw in Figure 2.1-2), the *violin plot*, and the *multiple-histogram* method. Here we compare these methods with an example. We recommend against the common practice of using a bar graph because bars fail to show the data (bar graphs are ideal for frequency data).

EXAMPLE 2.3B: Blood responses to high elevation

The amount of oxygen obtained in each breath at high altitude can be as low as one-third of that obtained at sea level. Do 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 United States (Beall et al. 2002). Results are shown in Figures 2.3-4 and 2.3-5.



Whitlock & Schluter, *The Analysis of Biological Data*, 3e © 2020 W. H. Freeman and Company

FIGURE 2.3-4

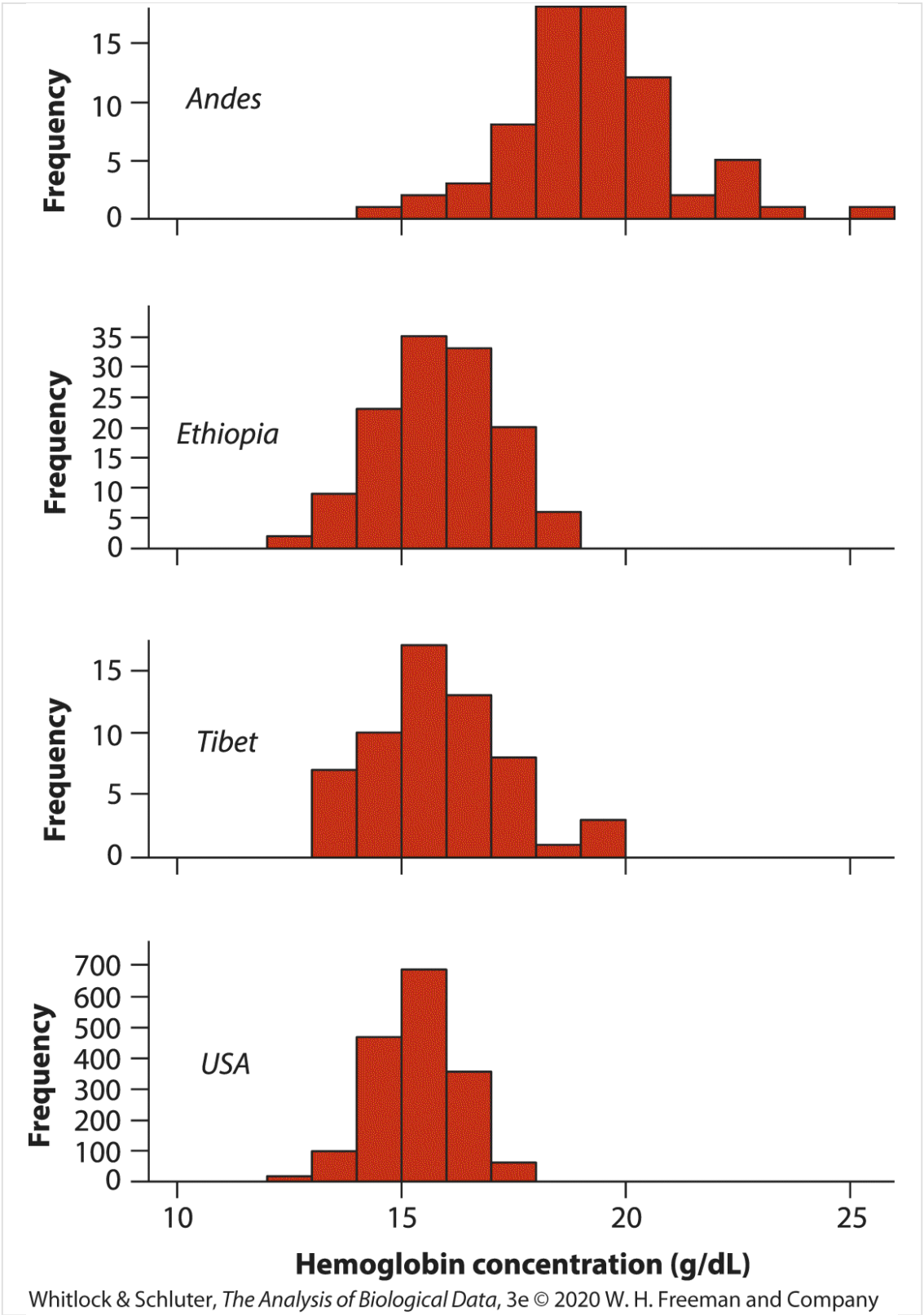
Strip chart (*left*) and violin plot (*right*) showing hemoglobin concentration in males living at high altitude in three different parts of the world: the Andes (71), Ethiopia (128), and Tibet (59). A fourth population of 1704 males living at sea level (USA) is included for comparison.

Description

For the strip chart, horizontal axis represents male population in Andes, Ethiopia, Tibet, and the U S A. The vertical axis represents hemoglobin concentration in grams per deciliter (lowercase G over lowercase D uppercase L) from 10 to 25 with an interval of 5. Approximate data from the strip chart are as follows: For Andes, data points are most dense between 16 and 21 grams per deciliter. For Ethiopia, data points are most dense between 13 and 18 grams per deciliter. For Tibet, data points are most dense between 15 and 17 grams per deciliter. For the U S A, data points are most concentrated between 13 and 17 grams per deciliter and show the highest density with a solid fill.

For the violin plot, the horizontal axis represents male population in Andes, Ethiopia, Tibet, and the U S A. The vertical axis represents hemoglobin concentration in grams per deciliter ranging from 10 to 25 with an interval of 5. Approximate data from the violin plot are as follows: For Andes, most of

the males have Hemoglobin concentration between 16 to 21 grams per deciliter. For Ethiopia, most of the males have Hemoglobin concentration between 13 and 18 grams per deciliter. For Tibet, most of the males have Hemoglobin concentration between 15 and 17 grams per deciliter. For the U S A, most of the males have Hemoglobin concentration between 13 and 17 grams per deciliter. The U S plot has the widest diameter and Andes the thinnest.



Whitlock & Schluter, *The Analysis of Biological Data*, 3e © 2020 W. H. Freeman and Company

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.

Description

The approximate data for Andes are as follows. 14 to 15, 1; 15 to 16, 2; 16 to 17, 3; 17 to 18, 8; 18 to 19, 17; 19 to 20, 17; 20 to 21, 11; 21 to 22, 2; 22 to 23, 5; 23 to 24, 1; 25 to 26, 1.

The approximate data for Ethiopia are as follows. 12 to 13, 2; 13 to 14, 9; 14 to 15, 22; 15 to 16, 35; 16 to 17, 32; 17 to 18, 19; 18 to 19, 5.

The approximate data for Tibet are as follows. 13 to 14, 7; 14 to 15, 10; 15 to 16, 17; 16 to 17, 13; 17 to 18, 8; 18 to 19, 1; 19 to 20, 3.

The approximate data for U S A are as follows. 12 to 13, 10; 13 to 14, 100; 14 to 15, 480; 15 to 16, 680; 16 to 17, 300; 17 to 18, 50.

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$y$-axis) and the category (group) to which it belongs on the other (here, the horizontal or x-axis$x$-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, a couple of the male populations in Example 2.3B 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 is the **violin plot**, which displays the data using a compact visual summary (right panel of Figure 2.3-4). The violin plot approximates the frequency distribution for each group like a histogram, but the distribution is smoothed and is shown with its mirror image. The dot in the center of each violin is the mean. 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 violins and data points. They both summarize the same data, but the location of the peak and where most of the observations

lie is easier to see in the violin plot than in the strip chart when groups have many data points.

> A *violin plot* is a graph that shows an approximation of the frequency distribution of a numerical variable in each group and its mirror image.

The plots in Figure 2.3-4 clearly show that 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.[6]

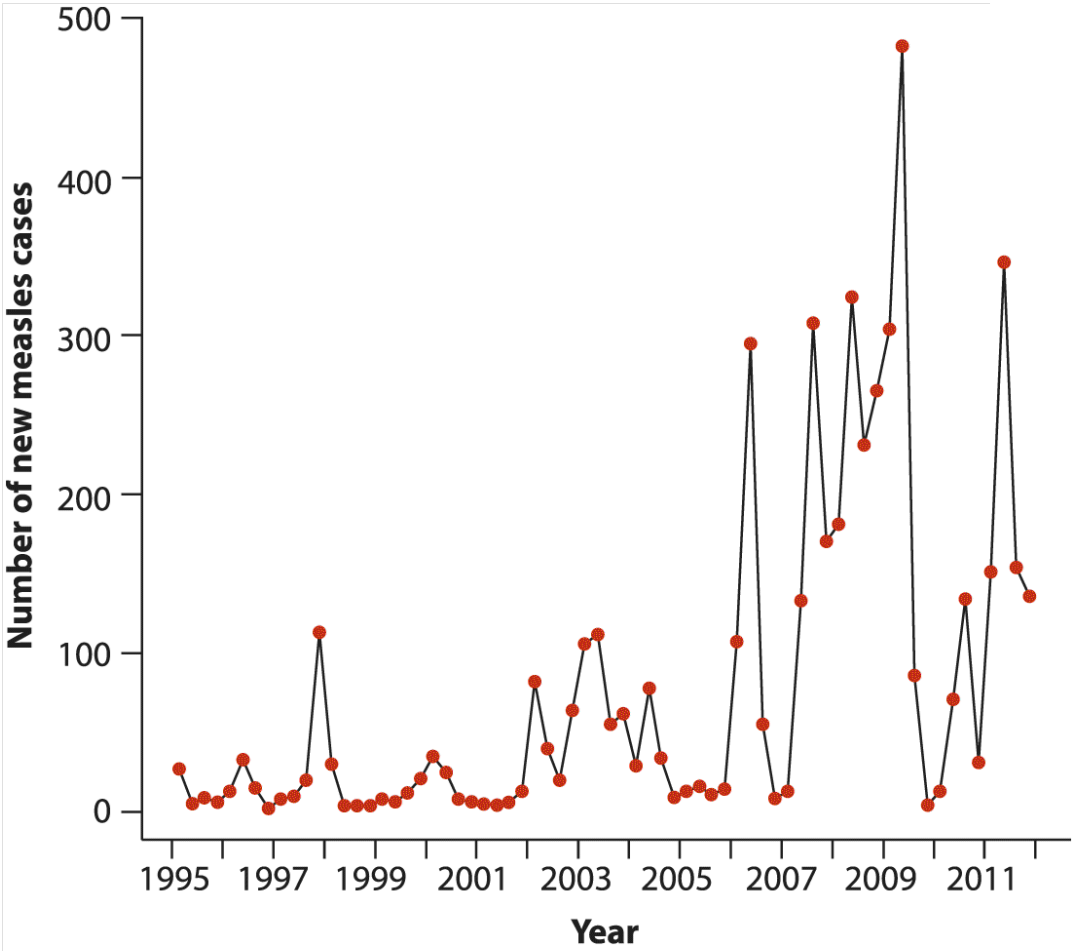
The third method uses multiple histograms, one for each category, to show the data, as shown in Figure 2.3-5. It is important that the histograms be stacked above one another as shown, so that the position and spread of the data are most easily compared. Side-by-side histograms lose most of the advantages 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 numerical 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 violin plot shows the most important features of the frequency distribution 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

Here we briefly introduce two other types of graphs in frequent use in biology, the *line graph* and the *map*. These two methods are often used to plot a summary measurement taken at consecutive points in time or space.

A **line graph** uses dots connected by line segments to display trends over time in a summary measurement, such as a mean, or other ordered series. For example, Figure 2.4-1 shows the number of cases of measles (the summary measurement) for quarterly time intervals between 1995 and 2011. The lines connecting the points help to show the temporal pattern more vividly. The steepness of the line segments reflects the speed of change in the number of cases from one quarter-year to the next. Notice 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 example, the area under the curve between two time points is proportional to the total number of new cases in that period.[7]



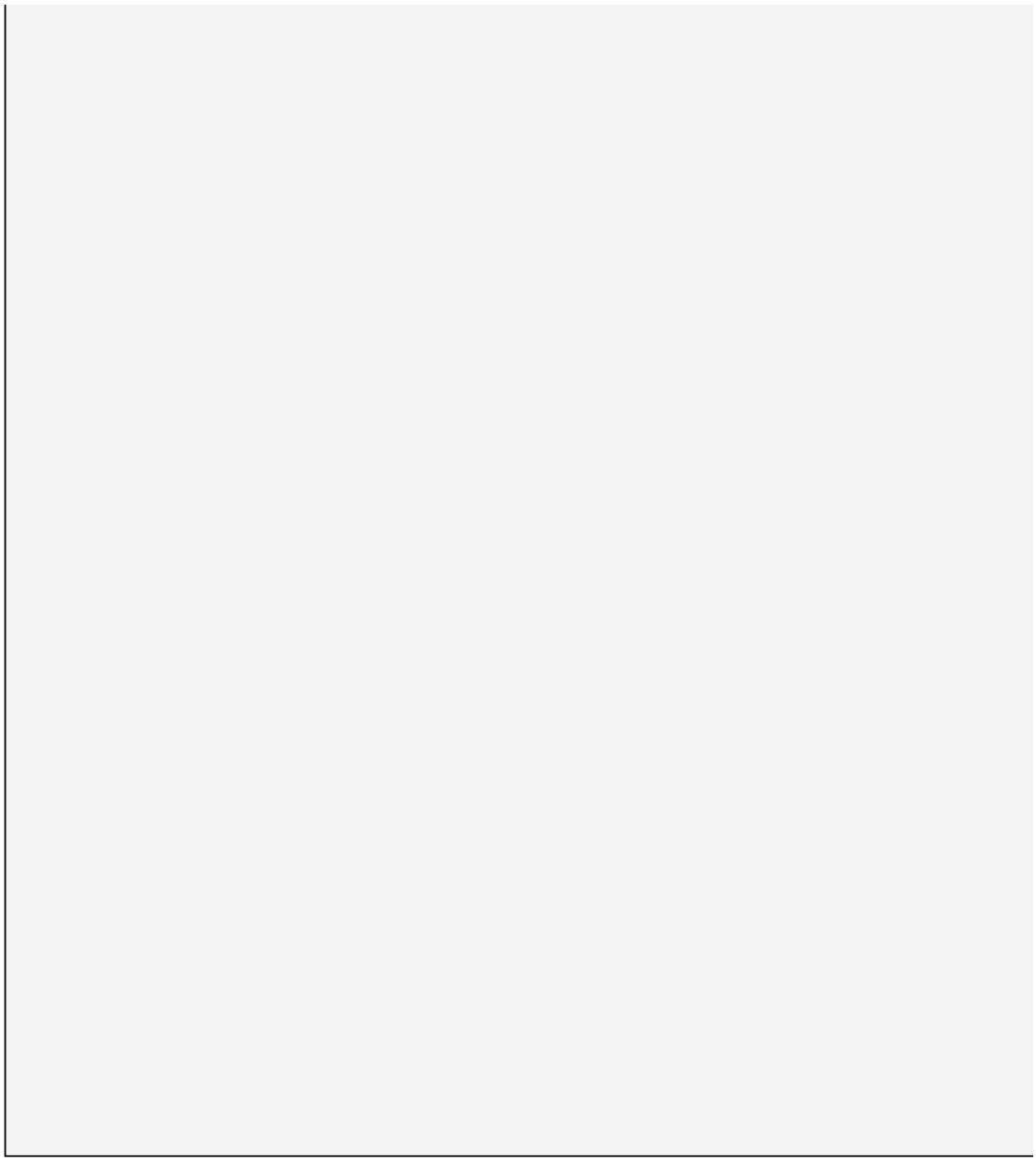
Whitlock & Schluter, *The Analysis of Biological Data*, 3e © 2020 W. H. Freeman and Company

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. Data are from Health Protection Agency (2012).

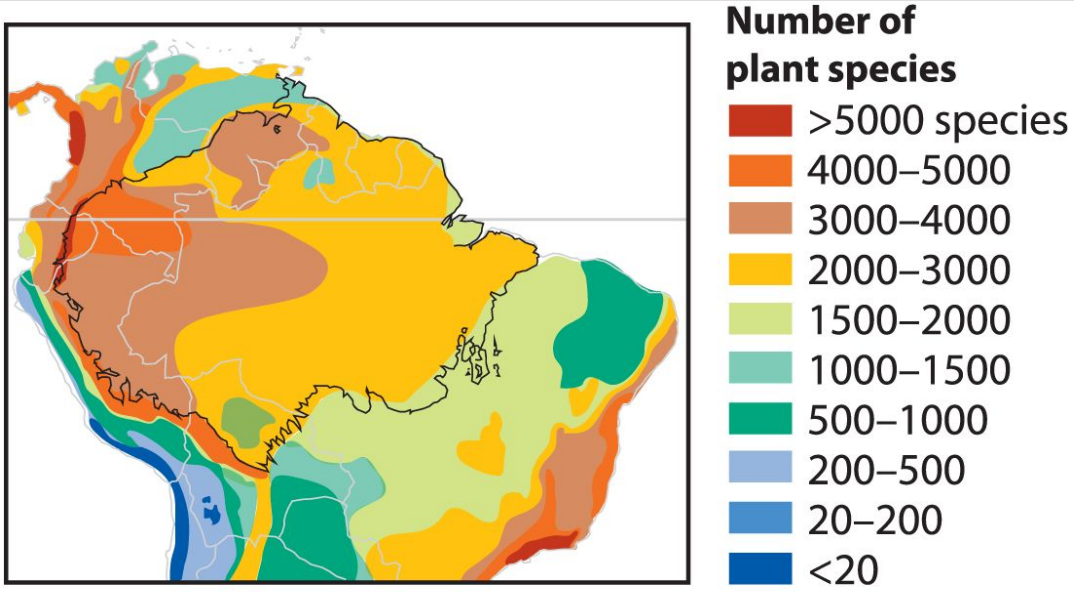
Description

The horizontal axis is labeled Year, marked from 1995 to 2011 with increments of 2. The vertical axis is labeled Number of new measles cases, ranging from 0 to 500 with increments of 100. Most of the points are plotted between 1995 and 2007 with cases about 2. Between 2006 and 2011, the cases range from 0 to 500. All the points are connected by lines.



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, including a spatial grid or at political or geological boundaries on a map. Maps can be used to show measurements at locations on the surface of any two- or three-dimensional objects, such as the brain or the body. A visual representation of a magnetic resonance imaging (MRI) scan is a map.

For example, Figure 2.4-2 shows 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 map summarizes 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.[8]



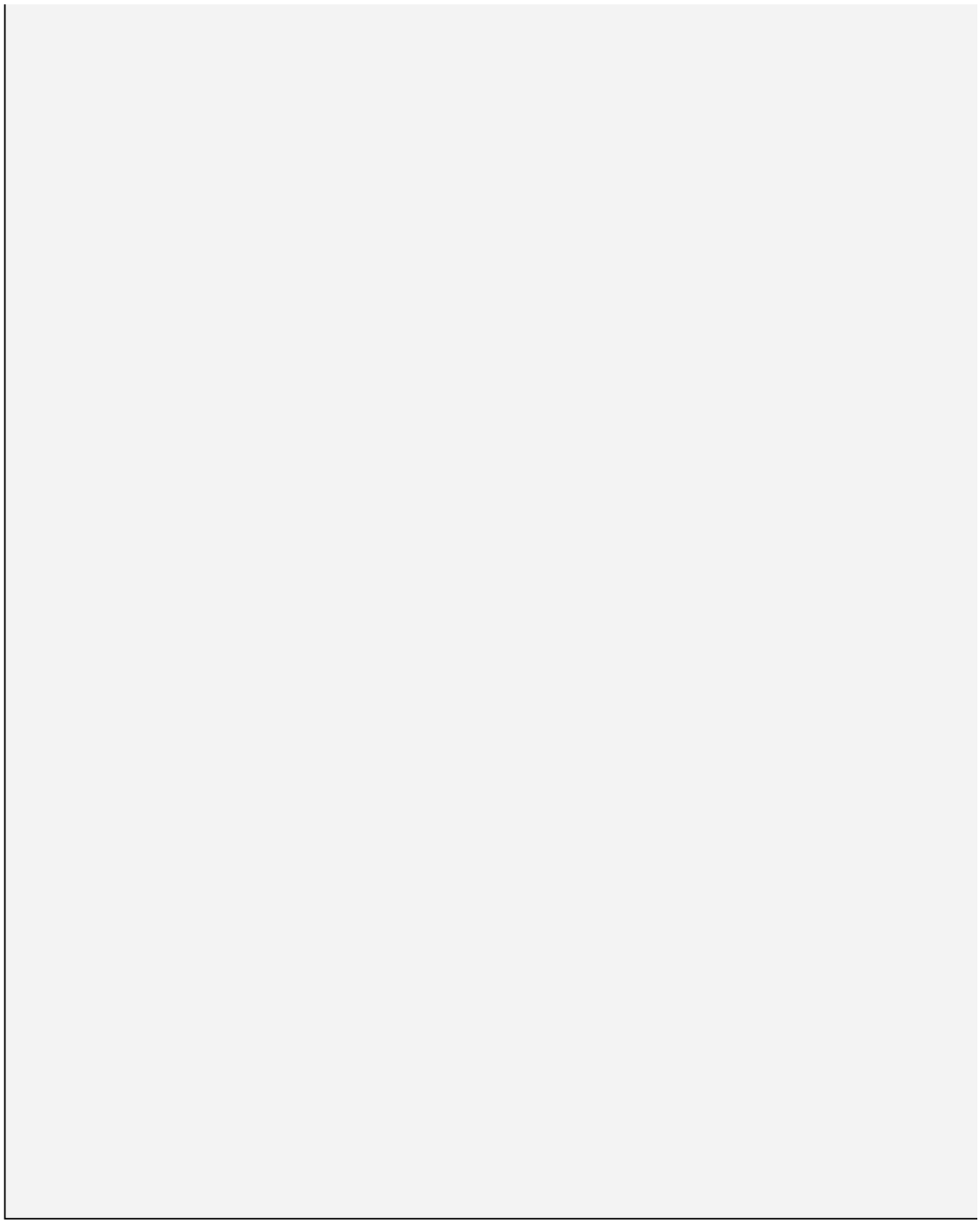
Whitlock & Schluter, *The Analysis of Biological Data*, 3e © 2020 W. H. Freeman and Company
Global Conservation Significance of Ecuador's Yasuní National Park, Bass MS,et al. PLoS ONE 2010.

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×100 km.$100\ \text{km} \times 100\ \text{km}.$ The color scale is on the right, with hotter colors reflecting more species. The horizontal gray line is the equator.

Description

The data are categorized as greater than 5000, 4000 to 5000, 3000 to 4000, 2000 to 3000, 1500 to 2000, 1000 to 1500, 500 to 1000, 200 to 500, 20 to 200, and less than 20.



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 2.2-2, which shows the frequency of fetuses whose measurements lie in a sequence of head-width 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 or detailed 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. 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.

Let's look at an example of a 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, who ruled Spain from 1516 to 1700, contributed to its downfall. The quantity FF is a measure of inbreeding in the offspring. FF is zero if king and queen were unrelated, and FF 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.

TABLE 2.5-1 Inbreeding coefficient (F)(F) of Spanish Habsburg kings and queens a their progeny.

King/Queen	FF	Pregnancies	Miscarriages & stillbirths	Neonatal deaths	Later deaths	Survivors to age 10	S
Ferdinand of AragonFerdinand of Aragon							
Elizabeth of CastileElizabeth of Castile	0.0390.039	77	22	00	00	55	0.7
Philip IPhilip I							
Joanna IJoanna I	0.0370.037	66	00	00	00	66	1.0
Charles ICharles I							
Isabella of PortugalIsabella of Portugal	0.1230.123	77	11	11	22	33	0.4
Philip IIPhilip II							
Elizabeth of ValoisElizabeth of Valois	0.0080.008	44	11	11	00	22	0.5
Anna of AustriaAnna of Austria	0.2180.218	66	11	00	44	11	0.1
Philip IIIPhilip III							
Margaret of AustriaMargaret of Austria	0.1150.115	88	00	00	33	55	0.6
Philip IVPhilip IV							
Elizabeth of BourbonElizabeth of Bourbon	0.0500.050	77	00	33	22	22	0.2
Mariana of AustriaMariana of Austria	0.2540.254	66	00	11	33	22	0.3

Data are from Alvarez et al. (2009).

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 farther, examine the table and make a note of any deficiencies. How might these deficiencies be overcome by modifying the table?

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 FF 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, FF 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 FF value of the offspring, and

survival has been placed in the adjacent column. Blank lines have been eliminated, and decimals have been rounded to two places.

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

King/Queen	F^{F}	Survival (postnatal)	Survival (total)	Number of pregnancies
Philip II/Elizabeth of ValoisPhilip II/Elizabeth of Valois	0.010.01	1.001.00	0.500.50	44
Philip I/Joanna IPhilip I/Joanna I	0.040.04	1.001.00	1.001.00	66
Ferdinand/Elizabeth of CastileFerdinand/Elizabeth of Castile	0.040.04	1.001.00	0.710.71	77
Philip IV/Elizabeth of BourbonPhilip IV/Elizabeth of Bourbon	0.050.05	0.500.50	0.290.29	77
Philip III/Margaret of AustriaPhilip III/Margaret of Austria	0.120.12	0.630.63	0.630.63	88
Charles I/Isabella of PortugalCharles I/Isabella of Portugal	0.120.12	0.600.60	0.430.43	77
Philip II/Anna of AustriaPhilip II/Anna of Austria	0.220.22	0.200.20	0.170.17	66
Philip IV/Mariana of AustriaPhilip IV/Mariana of Austria	0.250.25	0.400.40	0.330.33	66

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.

2.6 How to make data files

Nowadays, graphs are almost always created on a computer, using either a spreadsheet program or a statistics package. How should your data be entered on the computer so that you can graph it and analyze it most readily? Here are a few tips on how to enter and save your data to computer files for easiest use by most statistics packages.

	Population	Hemoglobin Concentration
1	USAUSA	10.4010.40
2	USAUSA	15.0515.05
3	AndesAndes	20.6420.64
4	AndesAndes	17.3517.35
5	EthiopiaEthiopia	17.4517.45
6	EthiopiaEthiopia	14.0014.00
7	TibetTibet	15.8715.87
8	TibetTibet	19.7319.73

Rows for individuals and columns for variables. Begin by entering your data into a spreadsheet using a spreadsheet program on the computer. Have each row give the data for a single individual or sampling unit. Use each column for a different variable. For example, here are the first 8 rows of a data file containing the hemoglobin concentration for males from different human populations from Example 2.3B. The full data file has 1962 lines or rows, one for each individual.

Include only numbers, and not letters or symbols, in the columns for numeric variables (e.g., HemoglobinConcentration). Entries for categorical variables (e.g., Population) can be numbers or letters. Avoid special characters altogether, such as $,$, %,%, #,#, @,@, and &.&. Leave cells blank that correspond to missing data. It is a good idea to have one of the columns indicate the identity (ID) of each individual so that you can recheck later that you've entered each individual's data correctly. The top line of the spreadsheet should contain the variable names without spaces or punctuation. Include more variables by adding new columns. Include more individuals by adding new rows.

Make the data files human readable. The variable names should be clear and unambiguous so that a new reader (including yourself after memory has faded) can interpret the meaning of each variable without error. "HemoglobinConcentration" takes longer to type than the alternative variable name "HC," but it will be clearer to later readers.

Keep a second file to interpret the data file. Always create another file, which accompanies the data file, containing a clear description of where the data come from, how they were obtained, and an explanation of each variable (including units). This auxiliary file should record the meaning of any codes you use in the data entries. For example, write down that "M" and "F" mean "male" and "female" so that the codes are not confused later with, for example, "mother" and "father."

Save spreadsheet as a plain text file. You want your data saved in a file that can be read by as many programs as possible, now and forever more, such as a ".csv" file (for "comma-separated variables"). With this format, each entry in the same row is separated by a comma, and new rows are separated by a line break. For example, here are the contents of a .csv file with the data from the table on the previous page:

```
Population,HemoglobinConcentration
USA,10.40
USA,15.05
Andes,20.64
Andes,17.35
Ethiopia,17.45
Ethiopia,14.00
Tibet,15.87
Tibet,19.73
```

Every computer package now and in the future can read plain text, including .csv, but not every package will read a file stored in an outdated format from Excel.

Ensure the accuracy and integrity of the data. Always double- or triple-check each data point in a file after you enter it. Read the data aloud to a friend to check against the original source. Graphing the data can help catch outliers caused by data entry errors. Store the data in a safe place, such as in the cloud, and arrange backups.

2.7 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
Two categorical variables	Grouped bar graph
	Mosaic plot
One numerical variable and one categorical variable	Strip chart
	Violin plot
	Multiple histograms
	Cumulative frequency distributions (Chapter 3)

- Data files should have a row for each individual and a column for each variable. Use clear, informative names for variables with no spaces. Save the data using plain text files, which will never become obsolete.

Online resources

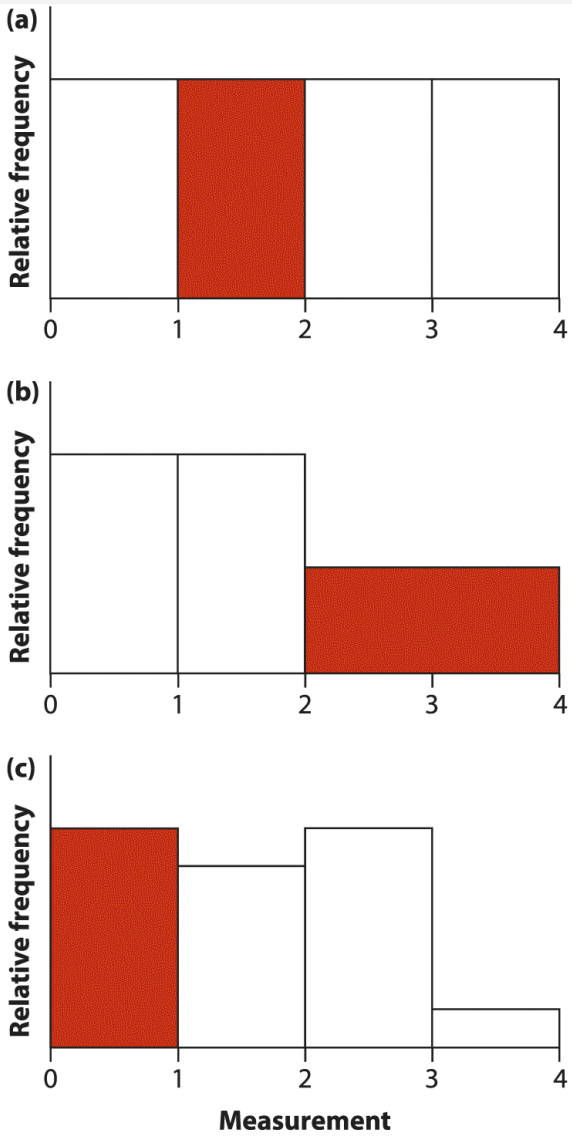
Learning resources associated with this chapter, including data for all examples and most problems, are online at https://whitlockschluter3e.zoology.ubc.ca/chapter02.html.

Chapter 2 Problems

PRACTICE PROBLEMS

Answers to the Practice Problems are provided in the Answers Appendix at the back of the book.

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



Whitlock & Schluter, *The Analysis of Biological Data*, 3e

Description

The horizontal axis is labeled Measurement, ranging from 0 to 4 with increment of 1. The vertical axis is labeled Relative Frequency. The approximate data in the first plot are as follows. 0 to 1, 1 to 2 which is shaded, 2 to 3, 3 to 4 – all are at the same level of frequency. The approximate data in the second plot are as follows. 0 to 1 and 1 to 2 are at the same level. 2 to 3 and 3 to 4, which are shaded, are at lower level. The

approximate data in the third plot are as follows. 0 to 1 which is shaded and 2 to 3 are at the same level. 1 to 2 and 3 to 4 are at various levels.

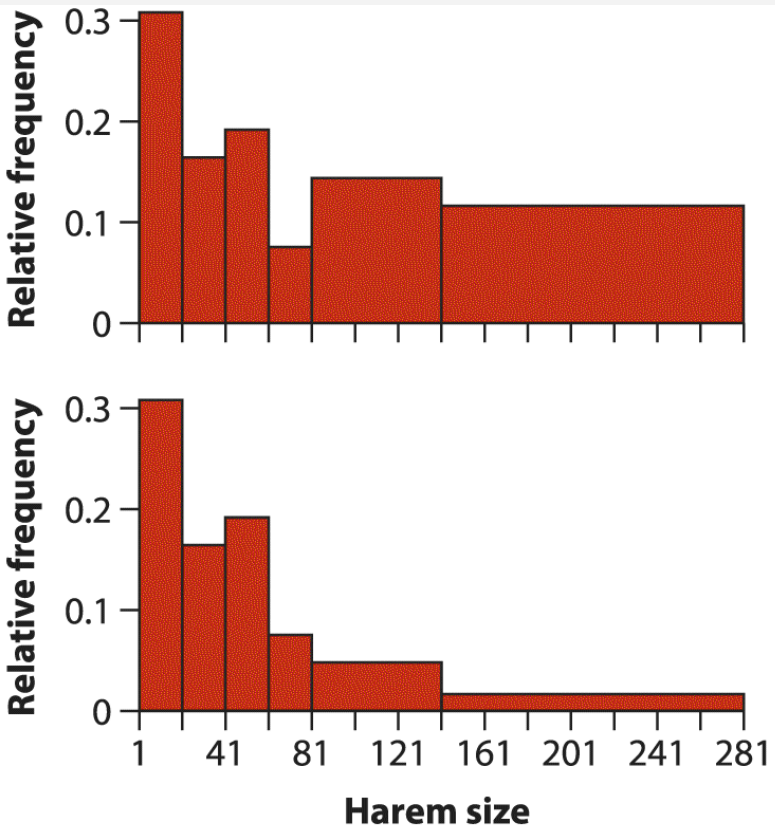
2. Using a graphical method from this chapter, draw three frequency distributions: one that is symmetric, one that is skewed, and one that is bimodal.

   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?



Whitlock & Schluter, *The Analysis of Biological Data*, 3e © 2020 W. H. Freeman and Company

Description

The horizontal axis is labeled Harem size, ranging from 1 to 281 with increments of 40. The vertical axis is labeled Relative Frequency, with points 0, 0 point 1, 0 point 2, and 0 point 3. The approximate data in the first plot are as follows. 0 to 21 harem size, relative frequency 0 point 3; 21 to 41, 0 point 15; 41 to 61, 0 point 2; 61 to 81, 0 point 07; 81 to 141, 0 point 5; 141 to 281, 0 point 25.

The approximate data in the second plot are as follows. 0 to 21, 0 point 3; 21 to 41, 0 point 15; 41 to 61, 0 point 2; 61 to 81, 0 point 07; 81 to 141, 0 point 14; 141 to 281, 0 point 11.

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 2018 of the different species groups in the list of endangered and threatened species under the U.S. Endangered Species Act (U.S. Fish and Wildlife Service 2018). The taxa are listed in alphabetical order in the table. The full data set can be downloaded from whitlockschluter3e.zoology.ubc.ca.

Species group	Number of species
AmphibiansAmphibians	4545
ArachnidsArachnids	1717
BirdsBirds	342342
ClamsClams	124124
CoralsCorals	2424
CrustaceansCrustaceans	2828
FishesFishes	208208
InsectsInsects	9494
MammalsMammals	381381
ReptilesReptiles	146146
SnailsSnails	5555

   a. Rewrite the table, but list the species groups 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 species group. 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 17, the smallest number in the data set? Why?

   e. Create a version of this table that shows the relative frequency of endangered species by species group.

7. Can environmental factors affect 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).

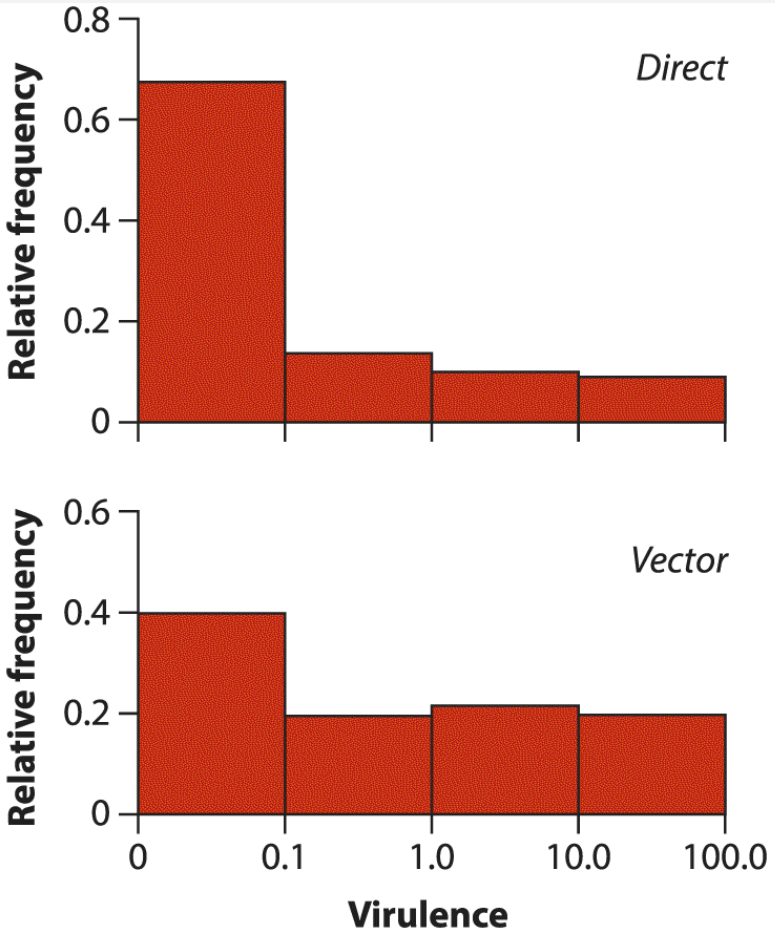
   a. What two variables are compared in this example?

   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 (data 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.

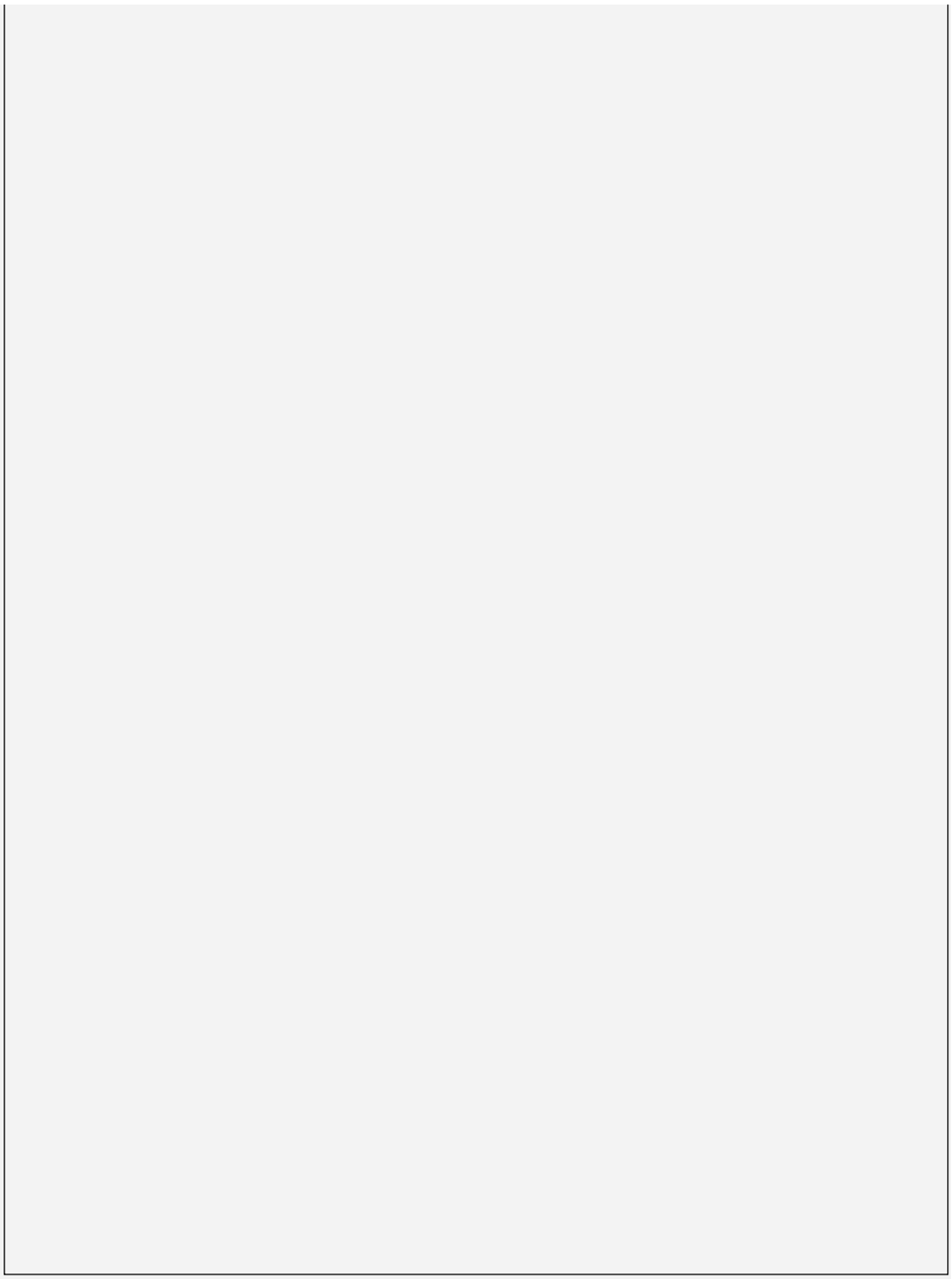


Whitlock & Schluter, *The Analysis of Biological Data*, 3e © 2020 W. H. Freeman and Company

Description

The horizontal axis is labeled Virulence, with points 0, 0 point 1, 1, 10, and 100. The vertical axis is labeled Relative Frequency, with points 0, 0 point 2, 0 point 4, 0 point 6, and 0 point 8. The approximate data in the first plot titled Vector are as follows. 0 to 0 point 1, 0 point 4; 0 point 1 to 1, 0 point 2; 1 to 10, 0 point 21; 10 to 100, 0 point 2.

The approximate data in the second plot titled Direct are as follows. 0 to 0 point 1, 0 point 68; 0 point 1 to 1, 0 point 13; 1 to 10, 0 point 1; 10 to 100, 0 point 09.



a. Identify the type of graph displayed.

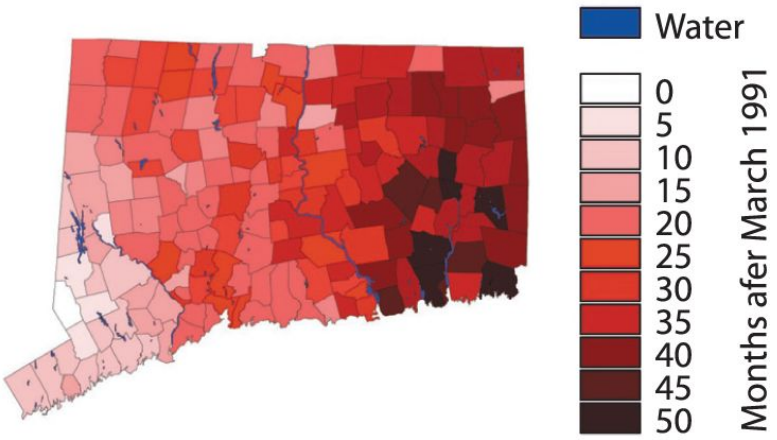
b. What are the two groups being compared in this graph?

c. What variable is being compared between the two groups? Is it numerical or categorical?

d. Explain the units on the vertical (y)(y) axis.

e. What is the main result depicted by this graph?

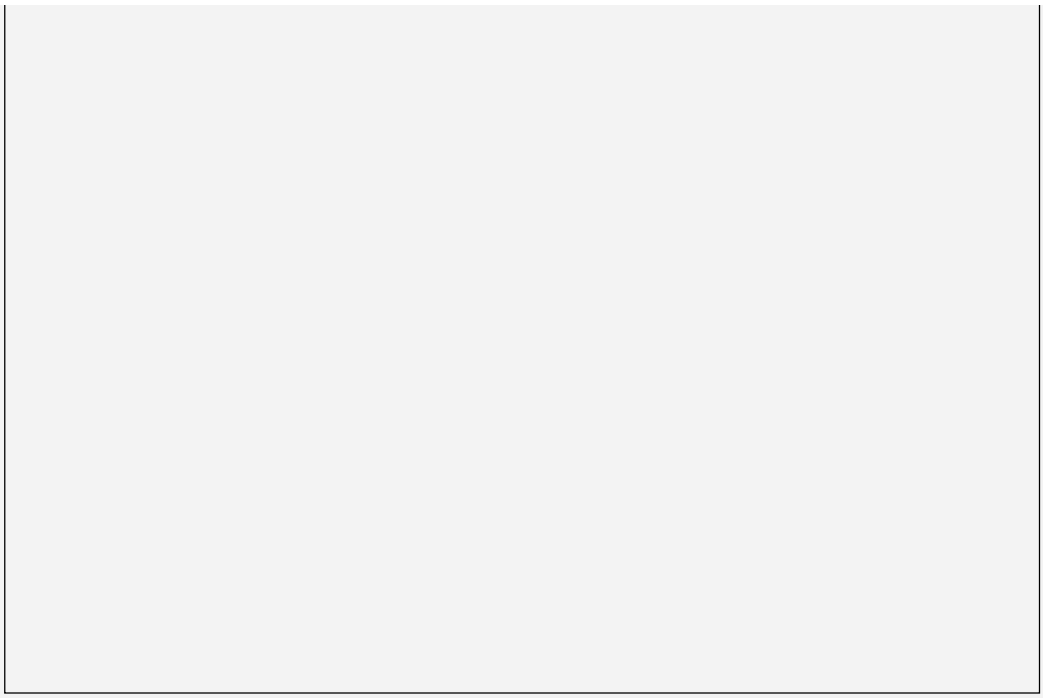
9. Examine the figure below, which indicates the date of first occurrence of rabies in raccoons in the townships of Connecticut, measured by the number of months following March 1, 1991.



Whitlock & Schluter, *The Analysis of Biological Data*, 3e
© 2020 W. H. Freeman and Company
Copyright (2002) National Academy of Sciences, U.S.A.

Description

-



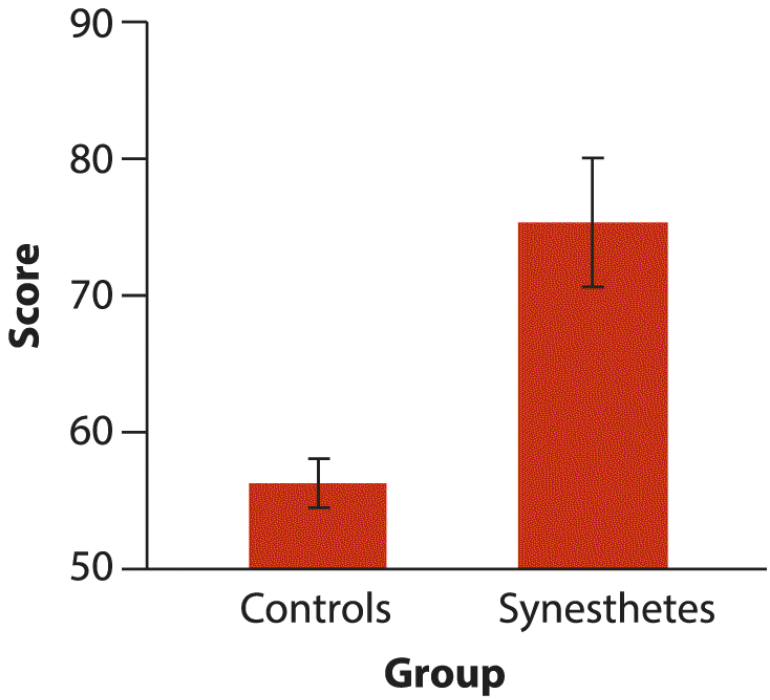
a. Identify the type of graph shown.

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 from a study of the proficiency of two groups of people on a complex visual task involving "rhythmic temporal patterns similar to Morse code" (Saenz & Koch 2008). One group consisted of four auditory synesthetes—healthy adults who experienced sound as well as sight when they observed visual flashes. The other group consisted of 10 adult controls who were not synesthetes. The study was conducted to determine if auditory synesthesia improves performance on the visual tasks. The score of each individual was measured on a scale of 0 to 100. The bars show the average score of the people in each group. (The lines protruding outward from the top edge of each bar are "standard error bars"—we'll learn about them in Chapter 4.) The raw data are available for download at whitlockschluter3e.zoology.ubc.ca.



Whitlock & Schluter, *The Analysis of Biological Data*, 3e

Description

The horizontal axis has two groups: controls and synesthetes. The vertical axis represents score from 50 to 90 with an interval of 10. The approximate data from the graph are as follows: For Controls, interquartile range lies between score 50 and 55. Minimum outlier is 55 and maximum outlier is 58. For Synesthetes, interquartile range lies between score 50 and 75. Minimum outlier is 71 and maximum outlier is 79.

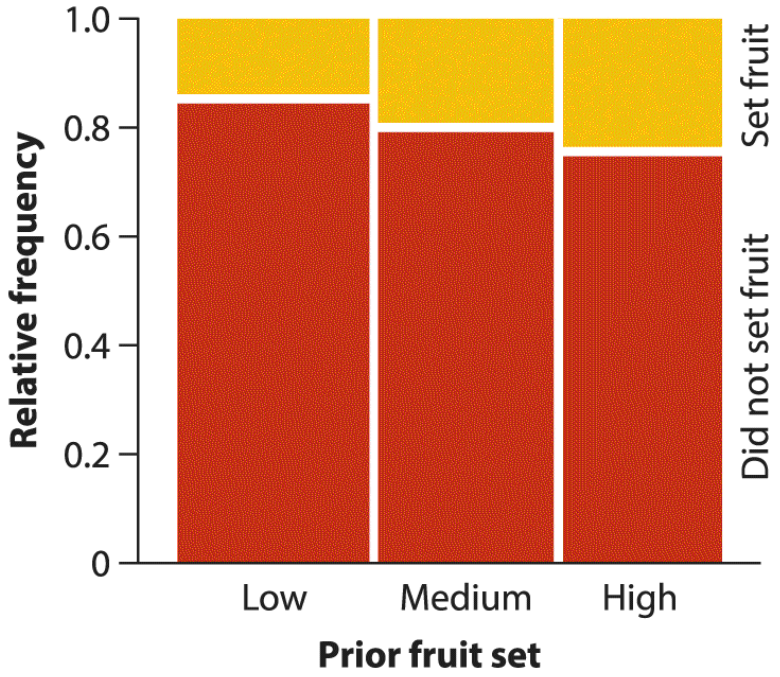
a. Describe the essential findings displayed in the figure.

b. Which two principles of good graph design are violated in this figure?

c. *Computer optional:* Download the data and redraw the plot, following the four principles of good graph design.

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):

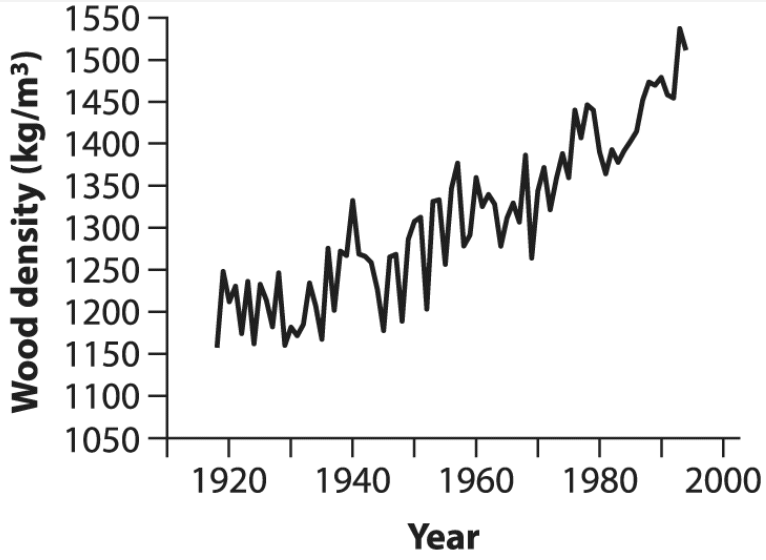


Whitlock & Schluter, *The Analysis of Biological Data*, 3e

Description

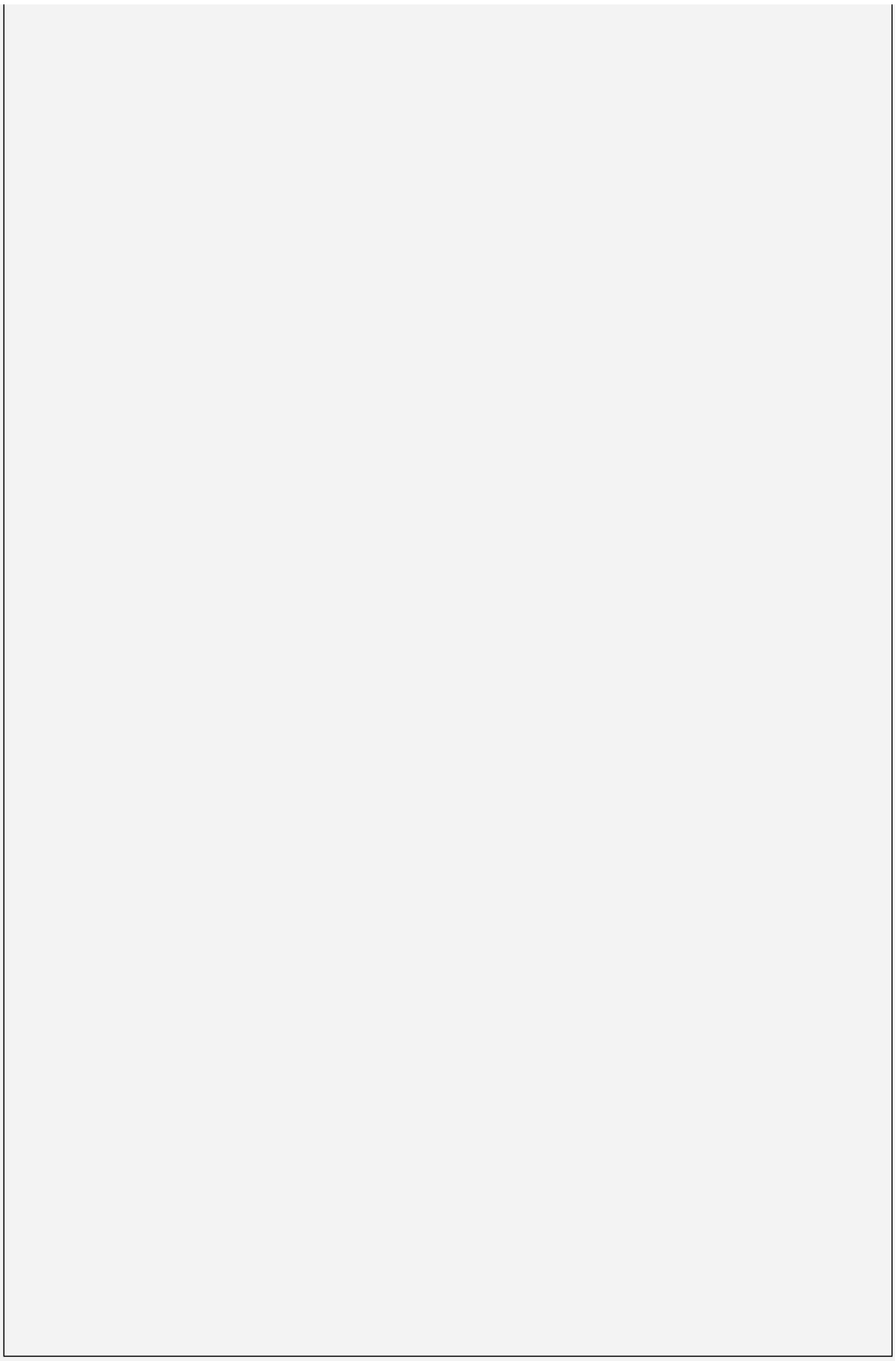
The horizontal axis is labeled Prior fruit set, with data Low, Medium, and High. The vertical axis is labeled Relative Frequency, ranging from 0 to 1 with increments of 0 point 2. The data are as follows. Low, Did not set fruit, 0 to 0 point 85; Low, Set fruit, 0 point 85 to 1; Medium, Did not set fruit, 0 to 0 point 8; Medium, Set fruit, 0 point 8 to 1; High, Did not set fruit, 0 to 0 point 75; High, Set fruit, 0 point 75 to 1.

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

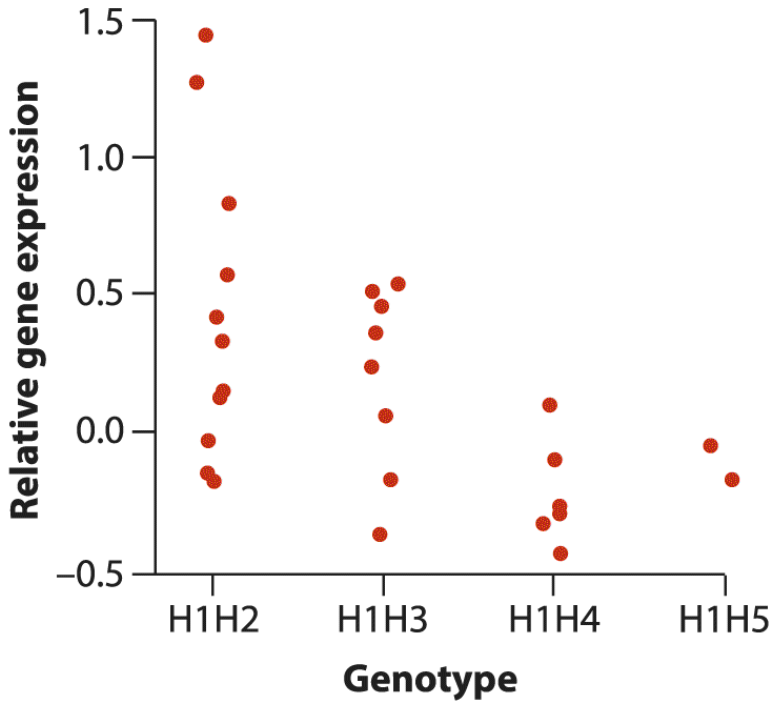


Whitlock & Schluter, *The Analysis of Biological Data*, 3e

Description
The horizontal axis is labeled Year, ranging from 1920 to 2000 with increments of 20. The vertical axis is labeled Wood density in kilogram per meter cubed, ranging from 1050 to 1550 with increments of 50. A line plot starts from 1150 in the year 1918, increases with frequent crests and troughs, and ends at 1500 in the year 1995.



c. Relative expression levels of *Neuropeptide* Y^{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):



Whitlock & Schluter, *The Analysis of Biological Data*, 3e

Description

The horizontal axis is labeled Genotype, with data H1H2, H1H3, H1H4, and H1H5. The vertical axis is labeled Relative gene expression, ranging from negative 0 point 5 to 1 point 5 with increments of 0 point 5. The approximate data are as follows. At H1H2, 11 points are plotted from negative 0 point 2 to 1 point 5. At H1H3, 8 points are plotted from negative 0 point 4 to 0 point 5. At H1H4, 6 points are plotted from negative 0 point 4 to 0 point 2. At H1H5, 2 points are plotted, one at negative 0 point 1 and another at negative 0 point 2.

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. The raw data are available at whitlockschluter3e.zoology.ubc.ca.

	Income level		
	Inadequate	**Adequate**	**Comfortable**
No convictionsNo convictions	4747	128128	9090
ConvictedConvicted	4343	5757	3030

a. What type of table is this?

b. Display these same data in a mosaic plot.

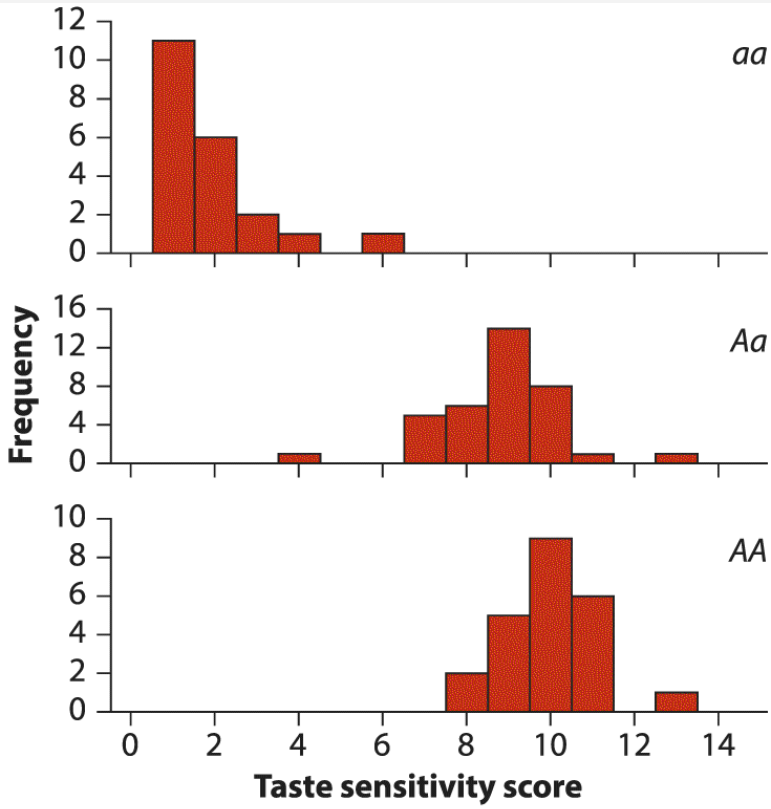
c. What type of variable is "income level"? How should this affect the arrangement of groups in your mosaic plot in part (b)?

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):



Whitlock & Schluter, *The Analysis of Biological Data*, 3e

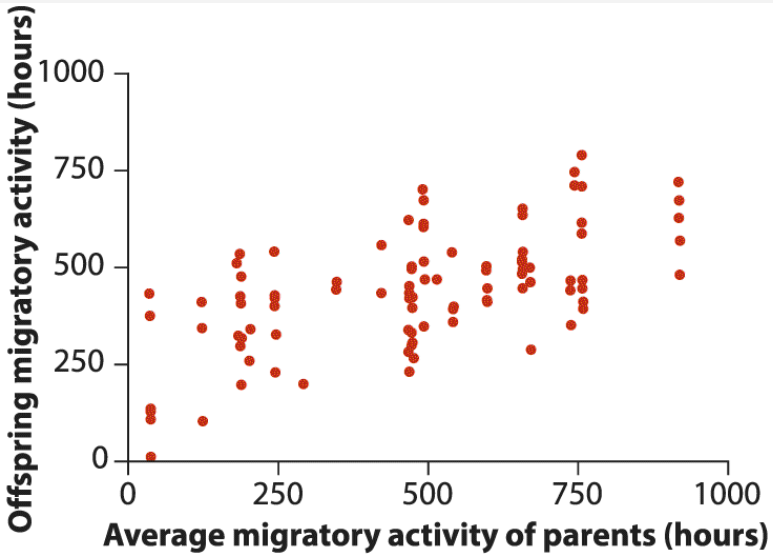
Description

The horizontal axis is labeled Taste sensitivity score, ranging from 0 to 14 with increments of 2. The vertical axis is labeled Frequency. The approximate data in the first plot titled “A A” are as follows: 7 point 5 to 8 point 5, 2; 8 point 5 to 9 point 5, 5; 9 point 5 to 10 point 5, 9; 10 point 5 to 11 point 5, 6; 12 point 5 to 13 point 5, 1.

The approximate data in the second plot titled “A a” are as follows. 3 point 5 to 4 point 5, 1; 6 point 5 to 7 point 5, 5; 7 point 5 to 8 point 5, 6; 8 point 5 to 9 point 5, 14; 9 point 5 to 10 point 5, 8; 10 point 5 to 11 point 5, 1; 12 point 5 to 13 point 5, 1.

The approximate data in the third plot titled “a a” are as follows. 0 point 5 to 1 point 5, 11; 1 point 5 to 2 point 5, 6; 2 point 5 to 3 point 5, 2; 3 point 5 to 4 point 5, 1; 5 point 5 to 6 point 5, 1.

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):

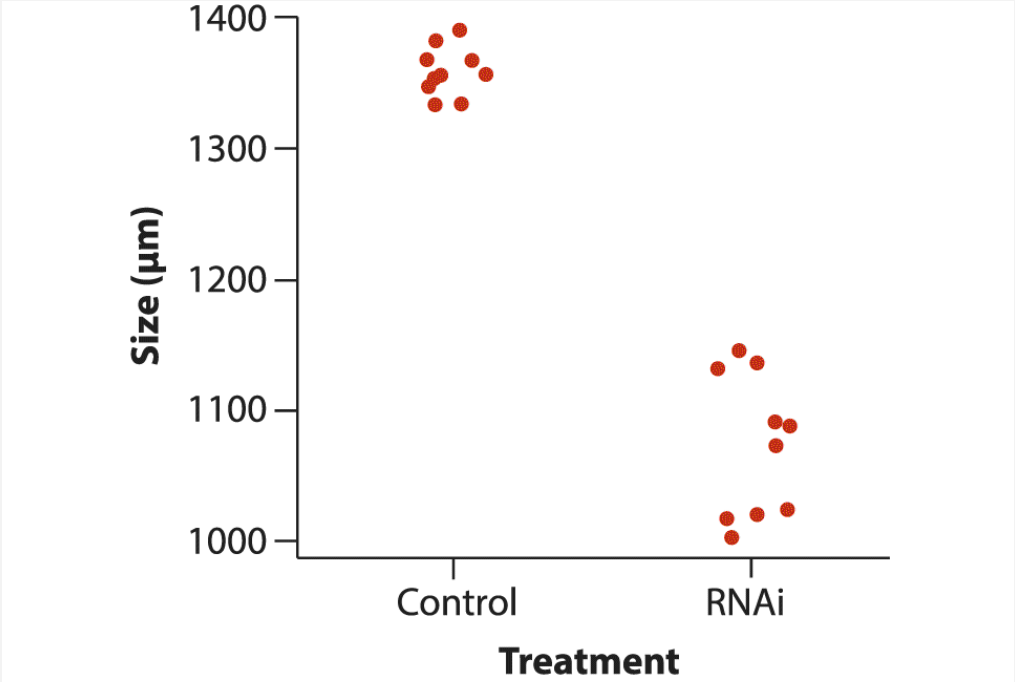


Whitlock & Schluter, *The Analysis of Biological Data*, 3e

Description

The horizontal axis is labeled Average migratory activity of parents in hours, ranging from 0 to 1000 with increments of 250. The vertical axis is labeled Offspring migratory activity in hours, ranging from 0 to 1000 with increments of 250. All data in the graph are approximate. The cluster of points is most dense between (10, 0) and (775, 775).

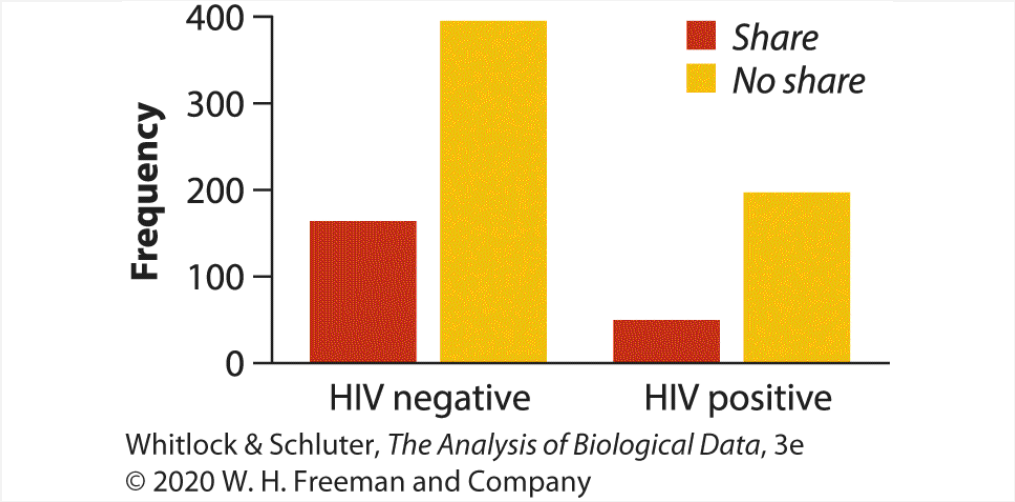
c. Sizes of the second appendage (middle leg) of embryos of water striders in 10 control embryos and 10 embryos dosed with RNAi for the developmental gene *Ultrabithorax* (Khila et al. 2009):



Description

The horizontal axis is labeled Treatment, with data Control and RNAi. The vertical axis is labeled Size in micrometer, ranging from 1000 to 1400 with increments of 100. The approximate data are as follows. The cluster of points is most dense between 1300 and 1400 against Control. The cluster of points is most dense between 1000 and 1150 against RNAi.

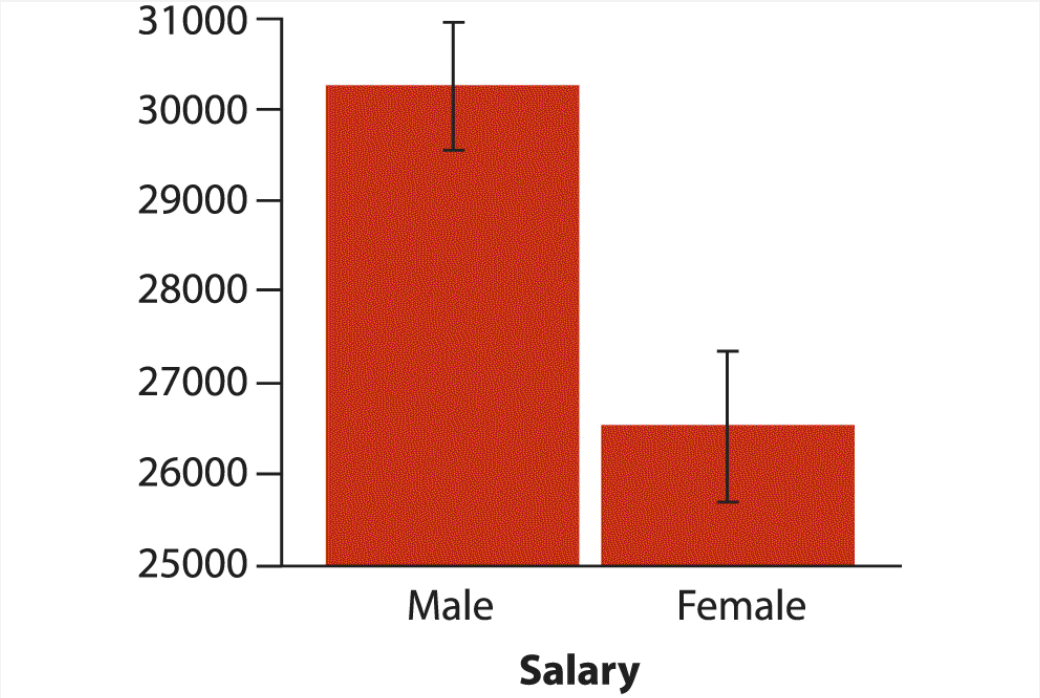
d. The frequency of injection-heroin users who share or do not share needles according to their known HIV infection status (Wood et al. 2001):



Description
The horizontal axis has data HIV negative and HIV positive. The vertical axis is labeled Frequency, ranging from 0 to 400 with increments of 100. The data are as follows. HIV negative, Share, frequency 170; HIV negative, No share, 400; HIV positive, Share, 50; HIV positive, No share, 200.

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.)



Description

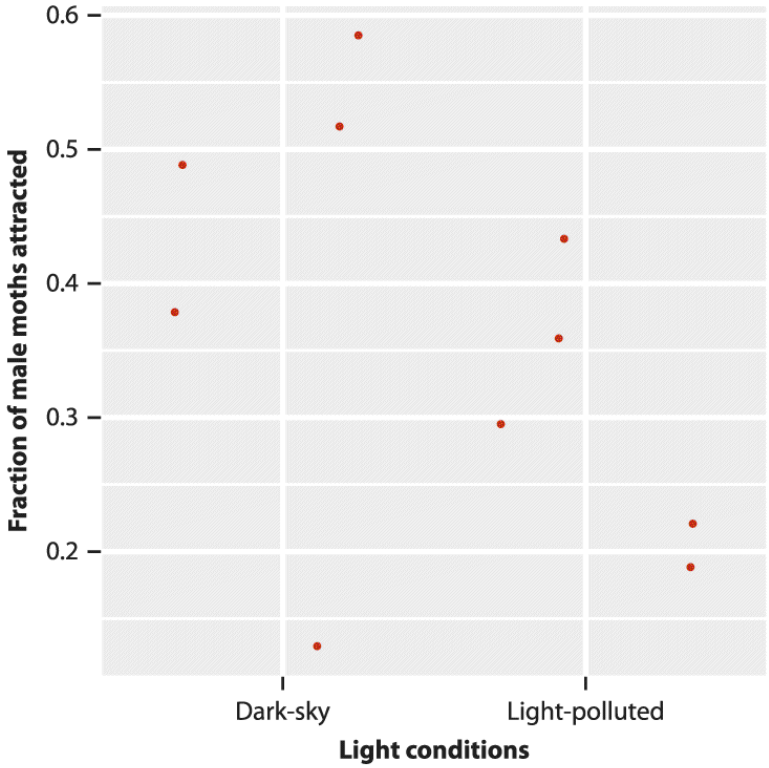
The horizontal axis is labeled Salary, with data Male and Female. The vertical axis has the points ranging from 25000 to 31000 with increments of 1000. The data are as follows. Male, 30250; the error bar ranges from 29500 to 30900; Female, 26500; the error bar ranges from 25600 to 27250.

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. Moths are fatally attracted to human light sources, including fire and street lights. If it is a significant source of mortality, we might predict that moth populations living close to human settlements would gradually evolve to reduce their attraction to its light sources. To test this possibility, Altermatt and Ebert (2016) measured light attraction of ermine moths (*Yponomeuta cagnagella*) from 10 different populations. Five of the populations were located in urban areas with plenty of human lights. The other five populations were located in pristine areas with no light pollution. The data are shown in the graph below. Each dot is the fraction of male moth individuals from a single population that were attracted to light when tested. The plot was computer-generated in R using default settings of the *ggplot* package.



Whitlock & Schluter, *The Analysis of Biological Data*, 3e

Description

The horizontal axis represents the following light conditions: dark-sky and light-polluted. The vertical axis represents fraction of male moths attracted from zero to 0 point 6 with an interval of 0 point 1. The approximate data from the graph are as follows: For dark-sky, majority of data points lie between 0 point 7 5 and 0 point 5 9. For light-polluted, majority of data points lie between 0 point 0 5 and 0 point 4 5.

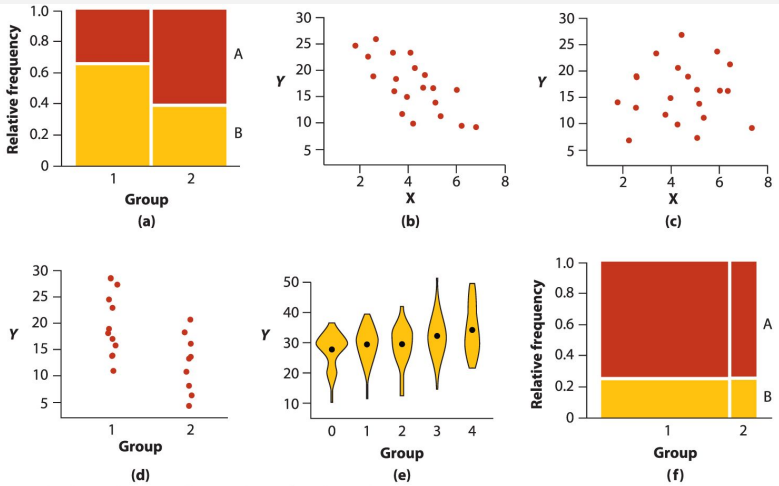
a. What type of graph is this?

b. What is the explanatory variable and what is the response variable in this graph?

c. What two principles of good graph design are violated in this graph?

d. *Computer optional:* Download the data from the book website and redraw the graph on the computer, fixing the problems identified in (c).

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.



Whitlock & Schluter, *The Analysis of Biological Data*, 3e © 2020 W. H. Freeman and Company

FIGURE FOR PROBLEM 16

Description

The first mosaic plot shows the horizontal axis labeled Group, with data 1 and 2. The vertical axis is labeled Relative frequency, ranging from 0 to 1 with increments of 0 point 2. The approximate data are as follows. 1, B, 0 to 0 point 65; 1, A, 0 point 65 to 1; 2, B, 0 to 0 point 4; 2, A, 0 point 4 to 1.

Two scatter plots show the horizontal axis marked from 2 to 8. The vertical axis is marked from 5 to 30. All data in the graphs are approximate. The cluster of points is most dense between (2, 25) and (7, 10) in the first scatter plot. The cluster of points is most dense between (2, 15) and (7, 22) in the second scatter plot.

The dot plot has the horizontal axis labeled Group, with data 1 and 2. The vertical axis is marked from 5 to 30 with increments of 5. 10 Points are plotted from 10 to 30 against 1. 9 Points are plotted from 5 to 20 against 2.

A box plot, having the horizontal axis and the vertical axis marked from 5 to 30 with increments of 5, depicts quartiles for 3 data sets: 1, 2, and 3.

The approximate data are as follows. The plot for 1 shows two vertical lines at the same level, one extending from lower extreme 12 to lower quartile 17 and the other extending from upper quartile 20 to upper extreme 26. A rectangle extends from lower quartile 17 to median 18, while another rectangle of the same height extends from median 18 to upper quartile 20. In line with the vertical line one dot is marked for y equals 29. Similarly, in the plots for 2 and 3, the lines extend from lower extreme 3 to lower quartile 8 and from upper quartile 12 to upper extreme 17. The two rectangles are to the up and below of median 10.

The second mosaic plot shows the horizontal axis labeled Group, with data 1 and 2. The vertical axis is labeled Relative frequency, ranging from 0 to 1 with increments of 0 point 2. The approximate data are as follows. 1, B, 0 to 0 point 25; 1, A, 0 point 25to 1; 2, B, 0 to 0 point 25; 2, A, 0 point 25 to 1.

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. The data are below and also at whitlockschluter3e.zoology.ubc.ca.

TABLE FOR PROBLEM 17

Anemone:	1	2	3	4	5	6	7	8	9	10	
Occasion One	10651065	248248	436436	350350	378378	410410	232232	201201	267267	687687	68
Occasion Two	939939	268268	460460	261261	368368	467467	303303	188188	401401	690690	71

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 the skew is positive or negative.

ASSIGNMENT PROBLEMS

Answers to all Assignment Problems are available for instructors, by contacting DL-WhitlockSchluter3e@macmillan.com.

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). These data are also available at whitlockschluter3e.zoology.ubc.ca.

0.047,0.047, 0.037,0.037, 0.041,0.041, 0.045,0.045, 0.039,0.039, 0.064,0.064, 0.064,0.064, 0.065,

0.065, 0.079,0.079, 0.070,0.070, 0.066,0.066, 0.059,0.059, 0.075,0.075, 0.079,0.079, 0.090,0.090,

0.069,0.069, 0.066,0.066, 0.078,0.078, 0.066,0.066, 0.066,0.066, 0.055,0.055, 0.046,0.046, 0.056,

0.056, 0.067,0.067, 0.075,0.075, 0.048,0.048, 0.077,0.077, 0.081,0.081, 0.066,0.066, 0.172,0.172,

0.080,0.080, 0.078,0.078, 0.048,0.048, 0.096,0.096, 0.0970.097

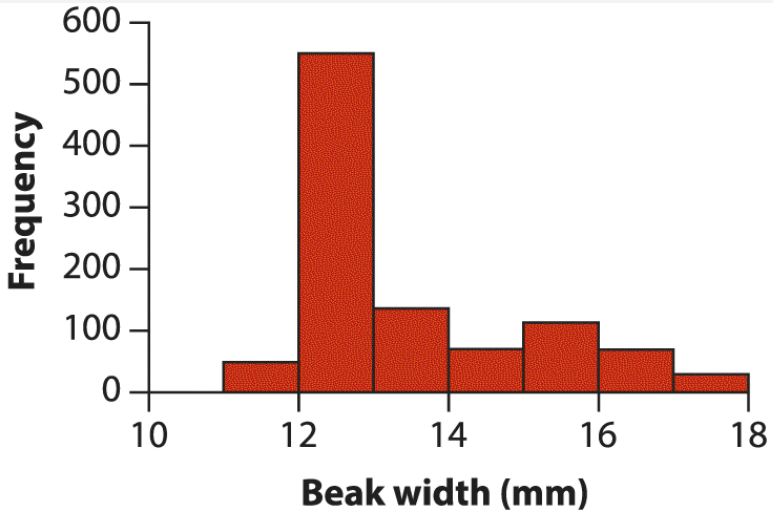
a. Create a graph depicting the frequency distribution of the 35 mass measurements.

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).



Whitlock & Schluter, *The Analysis of Biological Data*, 3e

Description
The horizontal axis is labeled Beak width in millimeter, ranging from 10 to 18 with increments of 2. The vertical axis is labeled Frequency, ranging from 0 to 600 with increments of 100. The approximate data are as follows. 11 to 12, 50; 12 to 13, 550; 13 to 14, 130; 14 to 15, 80; 15 to 16, 100; 16 to 17, 80; 17 to 18, 25.

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.

©photo courtesy of Thomas B. Smith

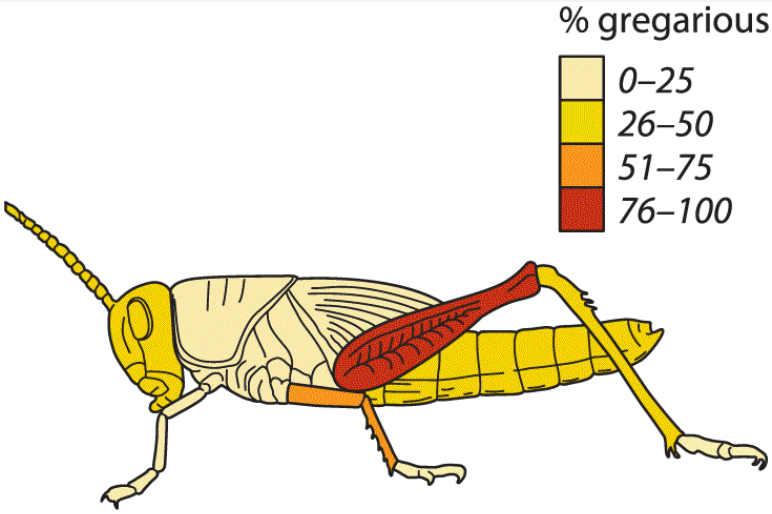
Description

-

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.



Description

-

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. The data are available at whitlockschluter3e.zoology.ubc.ca.

Number of convictions	Frequency
0	265
1	49
2	21
3	19
4	10
5	10
6	2
7	2

8	44
9	22
10	11
11	44
12	33
13	11
14	22
	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?

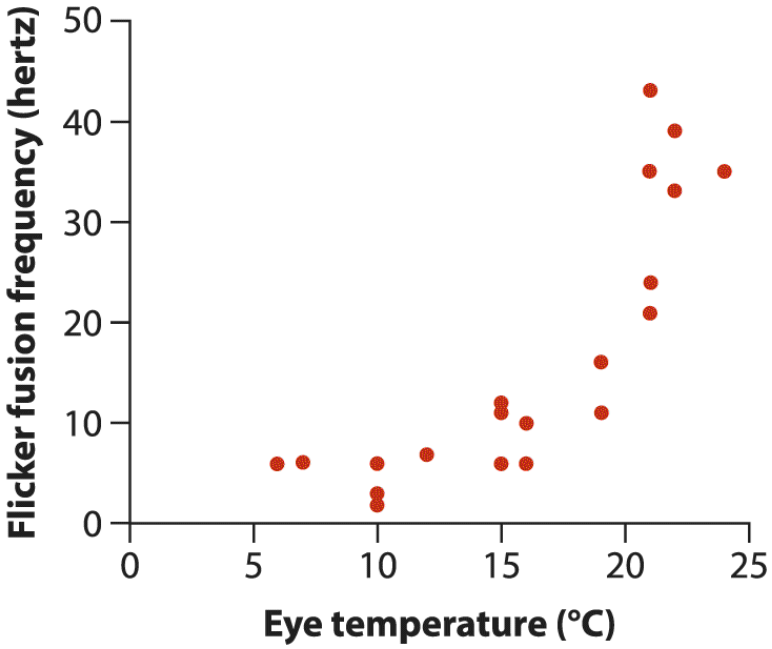
d. What fraction of the boys had no convictions?

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?

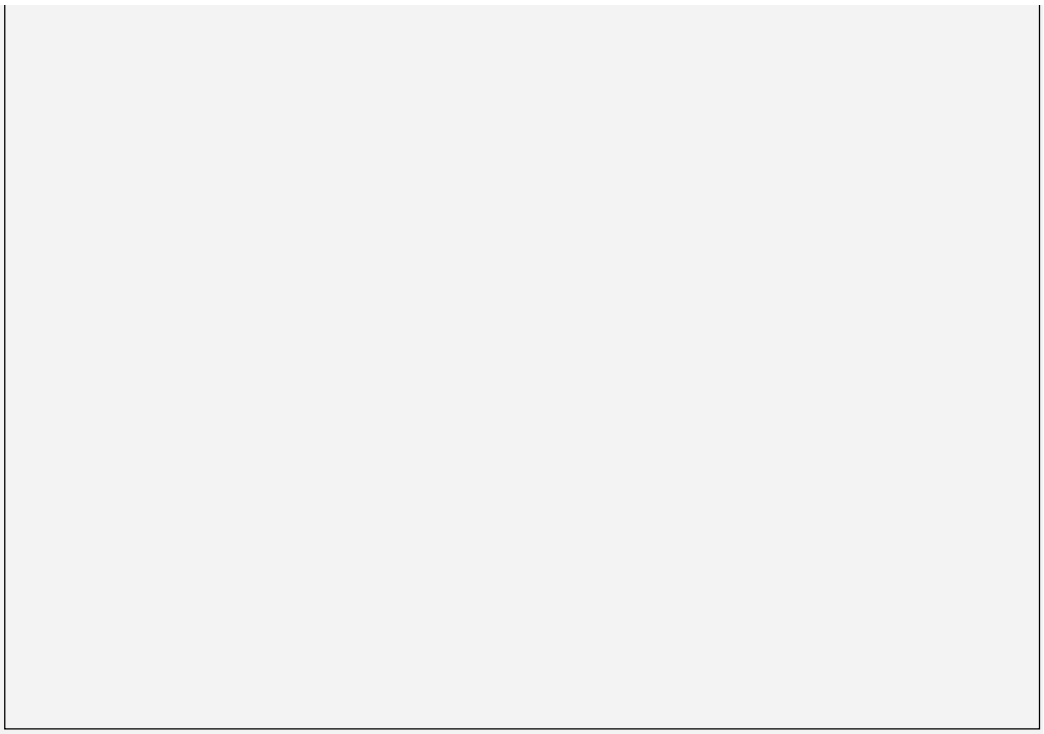
23. Swordfish have a unique “heater organ” that maintains elevated eye and brain temperatures when hunting in deep, cold water. The following graph illustrates the results of a study by Fritsches et al. (2005) that measured how the ability of swordfish retinas to detect rapid motion, measured by the flicker fusion frequency, changes with eye temperature.



Whitlock & Schluter, *The Analysis of Biological Data*, 3e

Description

The horizontal axis is labeled Eye temperature in Celsius, ranging from 0 to 25 with increments of 5. The vertical axis is labeled Flicker fusion frequency in hertz, ranging from 0 to 50 with increments of 10. The approximate data are as follows. (6, 6), (7, 6), (10, 1), (10, 3), (10, 6), (12, 7), (15, 6), (15, 10), (15, 12), (16, 6), (16, 10), (19, 11), (19, 17), (21, 20), (21, 24), (21, 35), (21, 43), (23, 33), (23, 39), (24, 35).



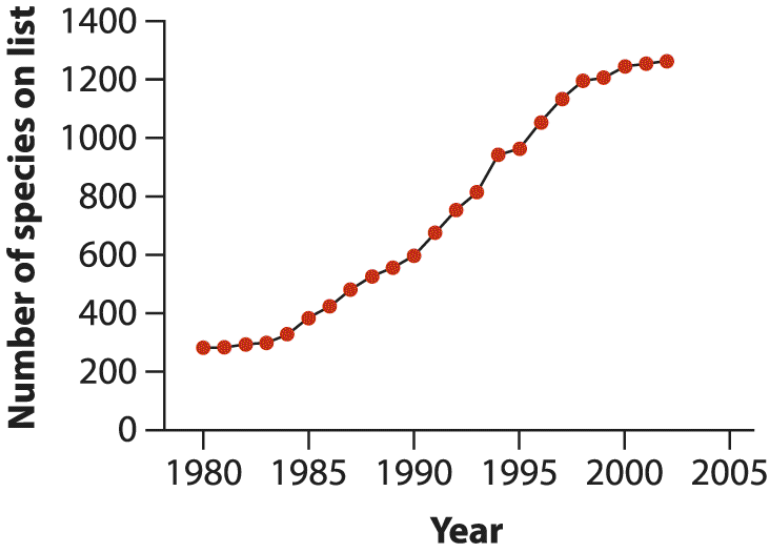
a. What types of variables are displayed?

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):



Whitlock & Schluter, *The Analysis of Biological Data*, 3e

Description

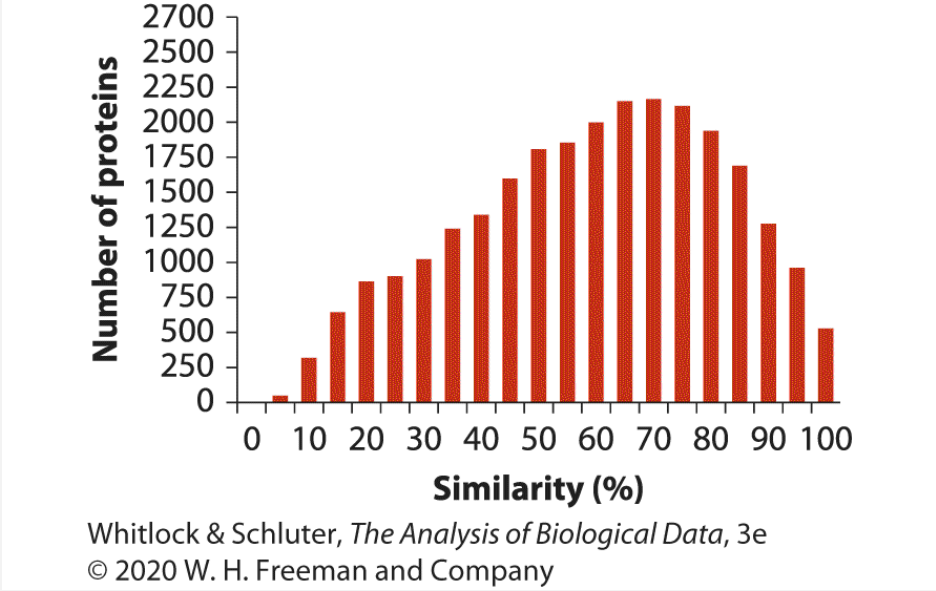
The horizontal axis is labeled Year, ranging from 1980 to 2005 with increments of 5. The vertical axis is labeled Number of species on list, ranging from 0 to 1400 with increments of 200. All data in the graph are approximate. The cluster of points is most dense between (1980, 300) and (2002, 1300). A curve is drawn through the points.

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* (data from Aparicio et al. 2002).



Description
The horizontal axis is labeled Similarity percent, ranging from 0 to 100 with increments of 5. The vertical axis is labeled Number of proteins, ranging from 0 to 2700 with increments of 250. The approximate data

are as follows. 6, 60; 11, 300; 16, 700; 21, 850; 26, 900; 31, 1000; 36, 1250; 41, 1359; 46, 1650; 51, 1750; 56, 1800; 61, 2000; 66, 2200; 71, 2200; 76, 2100; 81, 1900; 86, 1700; 91, 1300; 96, 950; 101, 500.

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. The data are also available at whitlockschluter3e.zoology.ubc.ca.

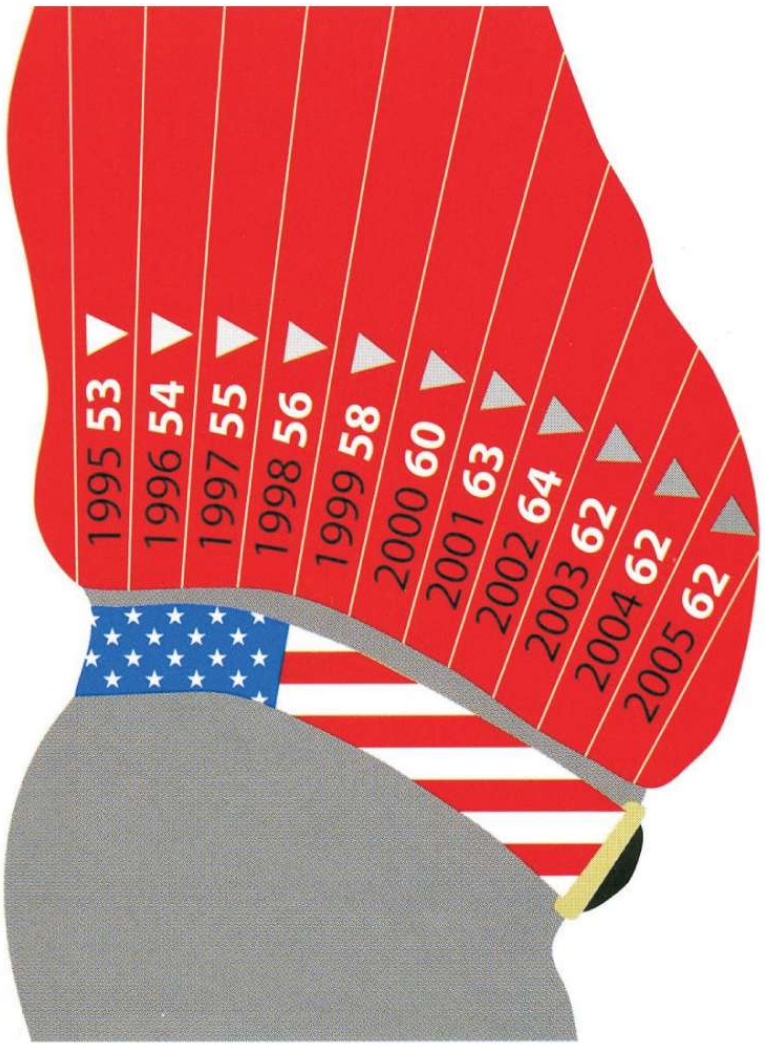
Number of fruits produced previously	Photosynthetic capacity (µmol O2/m2/s) ($\mu mol CO_2/m^2/s$)
1010	13.013.0
1414	11.911.9
55	11.511.5
2424	10.610.6
5050	11.111.1
3737	9.49.4
8989	9.39.3
162162	9.19.1
149149	7.37.3

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.

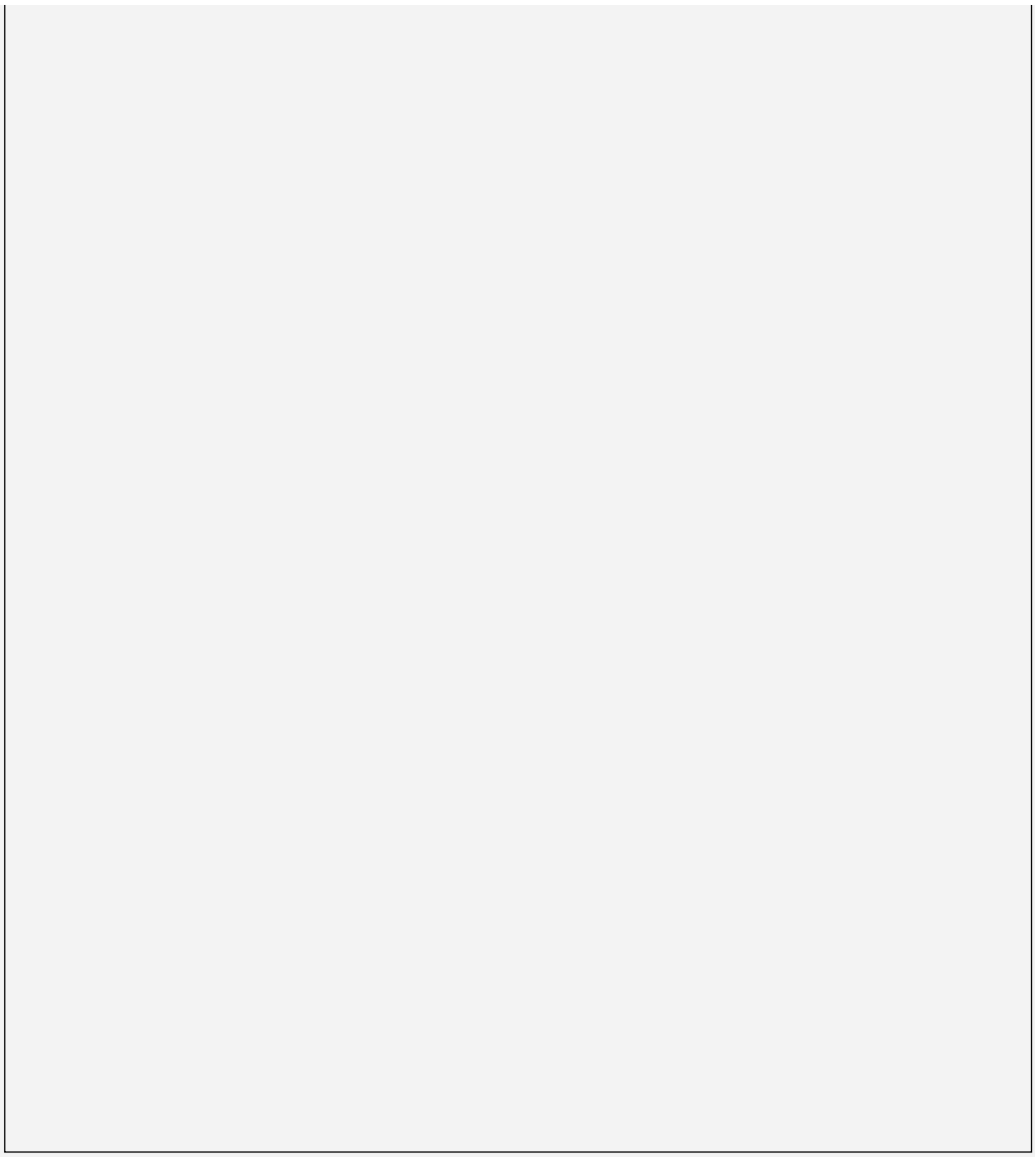
27. Examine the accompanying figure, which displays the percentage of adults over 18 with a "body mass index" greater than 25 in different years. Body mass index is a measure of weight relative to height.



Republished with permission of The Economist, That shrinking feeling, Emma Duncan, 2005; permission conveyed through Copyright Clearance Center.

Description

-



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. The raw data are available online at whitlockschluter3e.zoology.ubc.ca.

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.

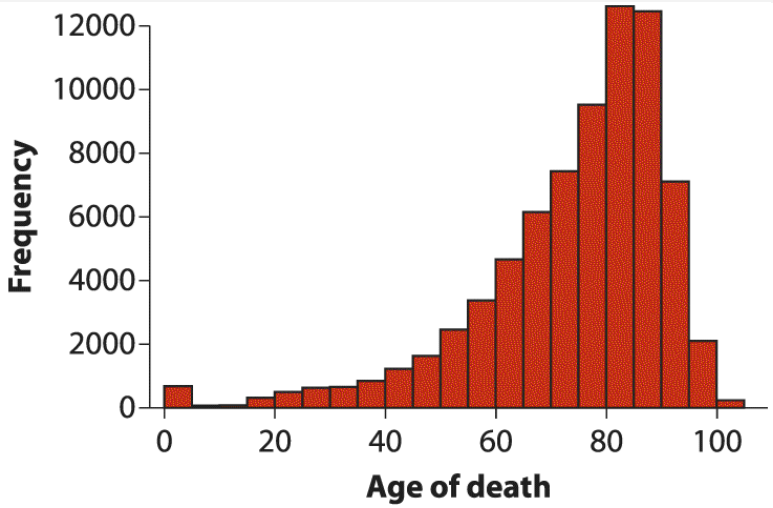
b. Illustrate the same results using a graphical technique instead. Identify the type of graph you used.

29. The graph below shows the distribution of age at death for males from Australia.

a. What kind of graph is this?

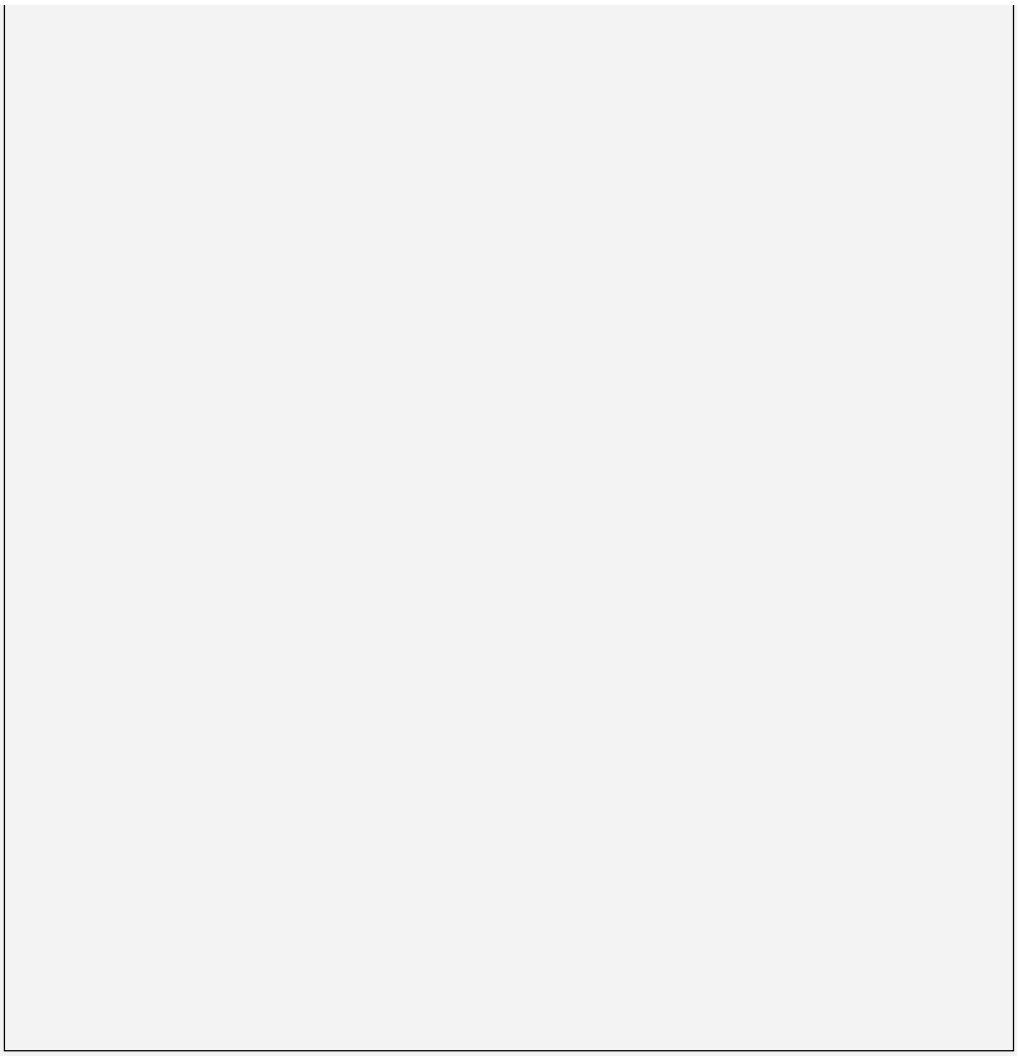
b. Describe the shape of the distribution. Is it symmetric or skewed? If it is skewed, describe the type of skew.

c. Is this distribution bimodal? Where are the mode or modes?

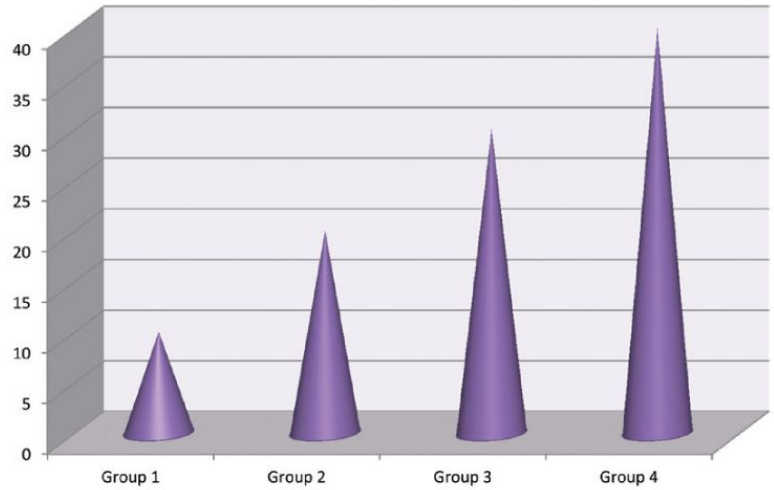


Whitlock & Schluter, *The Analysis of Biological Data*, 3e

Description
The horizontal axis represents age in years from 0 to 100 with an interval of 20. The vertical axis represents frequency from 2000 to 12000 with an interval of 2000. The approximate data from the graph are as follows: 0 to 5 age, 1000 frequency; 20 to 25 age, 600 frequency; 40 to 45 age, 1200 frequency; 60 to 65 age, 3800 frequency; 80 to 85 age, 12000 frequency; 95 to 100 age, 2000 frequency. The top edges of the bars forms an exponential curve till the age 80 to 85 which declines afterwards.



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.



Whitlock & Schluter, *The Analysis of Biological Data*, 3e

Description

-

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.

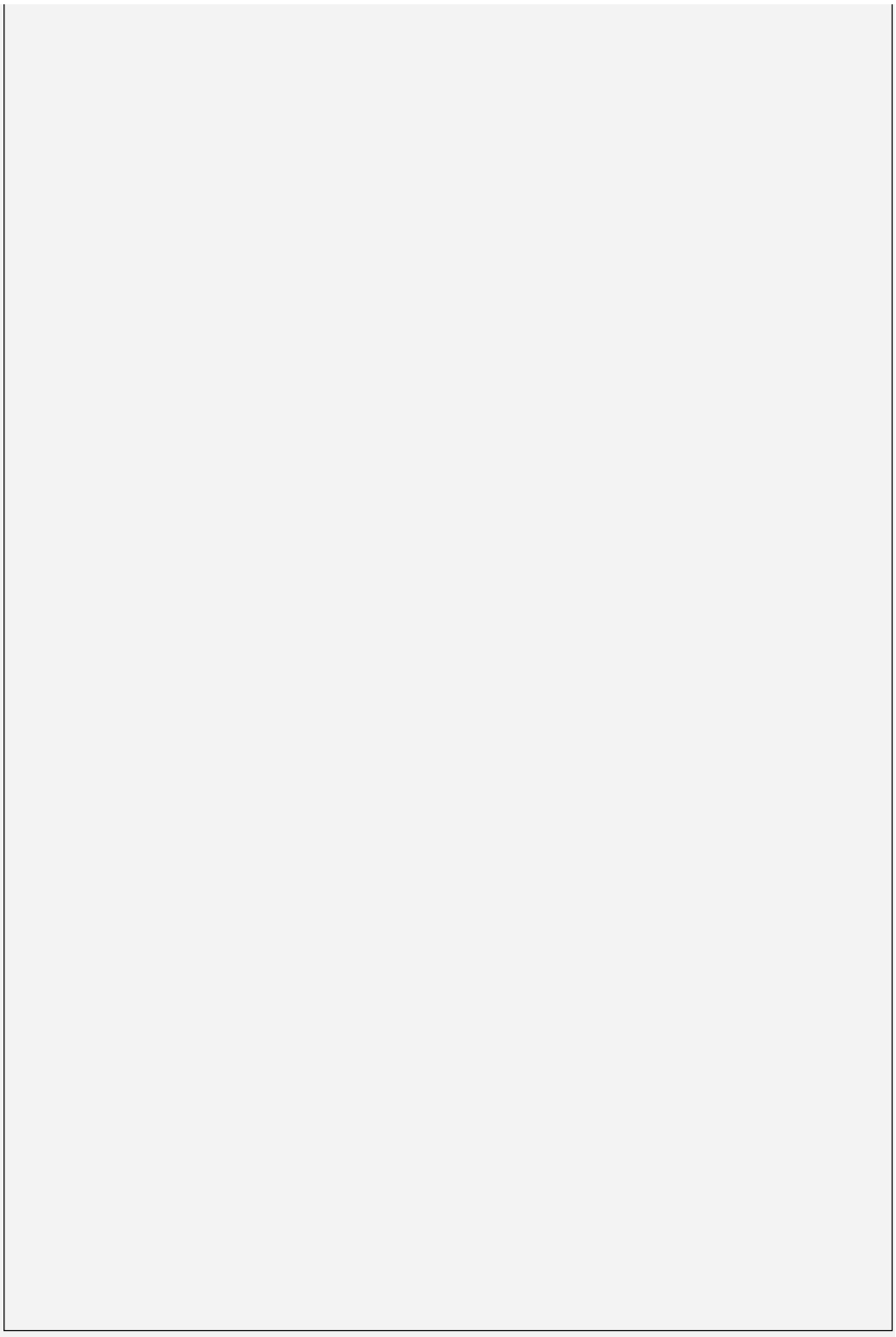
c. The actual frequencies are 10, 20, 30, and 40. Draw a graph that overcomes the problems identified above.

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 enrollment. The following graph shows the frequency distribution of scores (Freakonomics 2011).

Whitlock & Schluter, *The Analysis of Biological Data*, 3e

Description
The horizontal axis is labeled Score on Polish language exam, ranging from 0 to 70 with increments of 10. The vertical axis is labeled Relative Frequency, with points 0, 0 point 01, 0 point 02, and 0 point 03.

The rectangular bars are drawn such that it forms a concave downward curve with a gap from 18 to 20.



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 accidents. 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 accidentsDrivers with accidents	2121	3838
ControlsControls	3838	211211

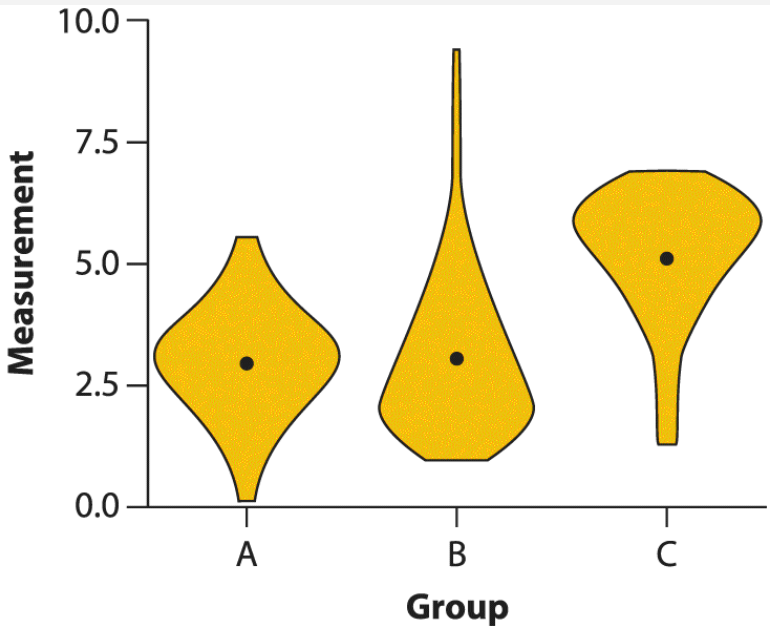
a. What type of table is this?

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? The raw data are available at whitlockschluter3e.zoology.ubc.ca.

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



Whitlock & Schluter, *The Analysis of Biological Data*, 3e

Description

The horizontal axis represents three groups: A, B, and C. The vertical axis represents measurements from 0.0 to 10.0 with an interval of 2.5. The approximate data from graph are as follows: For group A, majority of measurements lie between 1 to 5 with widest bulge at 2 point 7. For group B, majority of measurements lie between 1 and 6 with the widest bulge at 2 point 5. For group C, majority of measurements lie between 2 point 5 and 7 with the widest bulge at 6 point 5.

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 negative skew?

d. Which of the frequency distributions show positive skew?

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: 16.5,16.5, 18.9,18.9, 22.6,22.6, 27.8,27.8, 30.2,30.2, 30.7,30.7, 35.935.9

Females—control: 23.7,23.7, 24.5,24.5, 24.7,24.7, 26.1,26.1, 28.1,28.1, 33.4,33.4, 33.7,33.7, 35.235.2

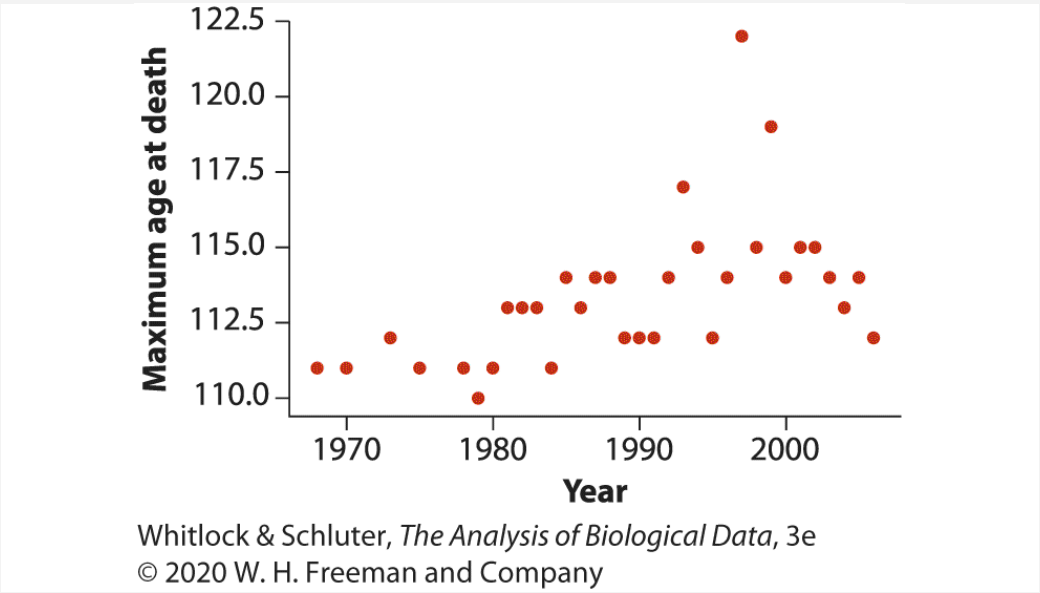
Males—reduced: 23.7,23.7, 28.1,28.1, 29.8,29.8, 31.1,31.1, 36.3,36.3, 37.7,37.7, 39.9,39.9, 39.9,39.9, 40.2,40.2, 40.240.2

Males—control: 24.9,24.9, 25.2,25.2, 29.6,29.6, 33.2,33.2, 34.1,34.1, 35.4,35.4, 38.1,38.1, 38.8,38.8, 40.740.7

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 following graph shows the maximum age of all individuals whose deaths were recorded in a given year, worldwide (Dong et al. 2016).



Whitlock & Schluter, *The Analysis of Biological Data*, 3e

Description

The horizontal axis represents calendar years from 19 70 to 2000 with an interval of 10. The vertical axis represents maximum age at death from 110 point 0 to 122 point 5 with an interval of 2 point 5. The approximate data from graph are as follows: 19 70, 111; 19 73, 112; 19 79, 109; 19 84, 113.5; 19 86, 114; 19 90, 112; 19 93, 117; 19 97, 122.5; 19 98, 119; 2000, 115; and 2005, 112 point 5.

a. What kind of graph is this?

b. Since the data describe a temporal sequence, what other type of plot would be suitable to display them?

c. One individual in this graph is an outlier relative to the others. What is the age of death for the outlier?

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 category 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.