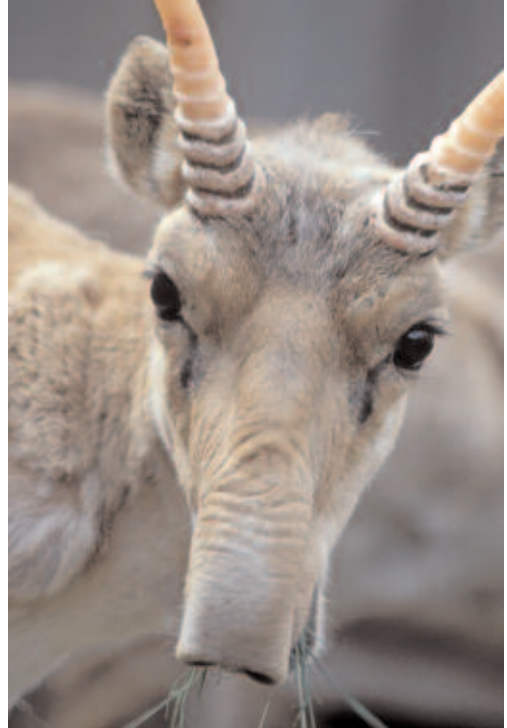


# 3

## Describing data



**D**escriptive statistics, or summary statistics, are quantities that capture important features of frequency distributions. Whereas graphs reveal shapes and patterns in the data, descriptive statistics provide hard numbers. The most important descriptive statistics for numerical data are those measuring the **location** of a frequency distribution and its **spread**. The location tells us something about the average or typical individual—where the observations are centered. The spread tells us how variable the measurements are from individual to individual—how widely scattered the observations are around the center. The **proportion** is the most important descriptive statistic for a categorical variable, measuring the fraction of observations in a given category.

The importance of calculating the location of a distribution seems obvious. How else do we address questions like “Which species is larger?” or “Which treatment group yielded the greatest response?” The importance of describing distribution spread is less obvious but no less crucial, at least in biology. In some fields of science, variability around a central value is instrument noise or measurement error, but in

biology much of the variability signifies real differences among individuals. Different individuals respond differently to treatments, and this variability begs measurement. Measuring variability also gives us perspective. We can ask, “How large are the differences between groups compared with variations within groups?” Biologists also appreciate variation as the stuff of evolution—we wouldn’t be here without variation.

In this chapter, we review the most common statistics to measure the location and spread of a frequency distribution and to calculate a proportion. We introduce the use of mathematical symbols to represent values of a variable, and we show formulas to calculate each summary statistic.

## 3.1 Arithmetic mean and standard deviation

The arithmetic mean is the most common metric to describe the location of a frequency distribution. It is the average of a set of measurements. The standard deviation is the most commonly used measure of distribution spread. Example 3.1 illustrates the basic calculations for means and standard deviations.

### Example 3.1

#### Gliding snakes

When a paradise tree snake (*Chrysopelea paradisi*) flings itself from a treetop, it flattens its body everywhere except for the region around the heart. As it gains downward speed, the snake forms a tight horizontal S shape and then begins to undulate widely from side to side. Lift is thereby created, causing the snake to glide away from the source tree. By orienting the head and anterior part of the body, the snake can change direction during a glide to avoid trees, reach a preferred landing site, and even chase aerial prey.

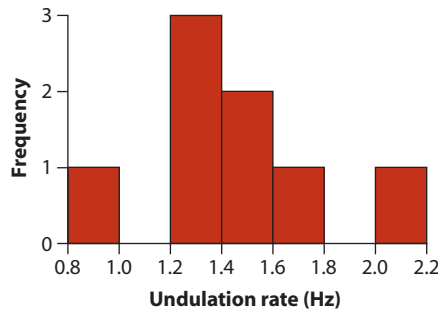


To better understand how lift is generated, Socha (2002) videotaped the glides of eight snakes leaping from a 10-m tower.<sup>1</sup> Among the measurements taken was the rate of side-to-side undulation on each snake. Undulation rates of the eight snakes, measured in Hertz (cycles per second), were as follows:

0.9, 1.4, 1.2, 1.2, 1.3, 2.0, 1.4, 1.6

A histogram of these data is shown in Figure 3.1-1. The frequency distribution has a single peak between 1.2 and 1.4 Hz.

<sup>1</sup> See films of these snakes flying at <http://www.flyingsnake.org/video/video.html>.



**Figure 3.1-1** Undulation rate of gliding paradise tree snakes.  $n = 8$  snakes.

## The sample mean

The **sample mean** is the average of the measurements in the sample, the sum of all the observations divided by the number of observations. To show its calculation, we use the symbol  $Y$  to refer to the variable and  $Y_i$  to represent the measurement of individual  $i$ . For the gliding snake data,  $i$  takes on values between 1 and 8, because there are eight snakes. Thus,  $Y_1 = 0.9$ ,  $Y_2 = 1.4$ ,  $Y_3 = 1.2$ ,  $Y_4 = 1.2$ , and so on.<sup>2</sup>

The sample mean, symbolized as  $\bar{Y}$  (and pronounced “Y-bar”), is calculated as

$$\bar{Y} = \frac{\sum_{i=1}^n Y_i}{n},$$

where  $n$  is the number of observations. The symbol  $\Sigma$  (uppercase Greek letter “sigma”) indicates a sum. The “ $i=1$ ” under the  $\Sigma$  and the “ $n$ ” over it indicate that we are summing over all values of  $i$  between 1 and  $n$ , inclusive

$$\sum_{i=1}^n Y_i = Y_1 + Y_2 + Y_3 \cdots + Y_n.$$

When it is clear that  $i$  refers to individuals 1, 2, 3, . . . ,  $n$ , the formula is often written more succinctly as

$$\bar{Y} = \frac{\Sigma Y_i}{n}.$$

Applying this formula to the snake data yields the mean undulation rate:

$$\bar{Y} = \frac{0.9 + 1.2 + 1.2 + 2.0 + 1.6 + 1.3 + 1.4 + 1.4}{8} = 1.375 \text{ Hz.}$$

<sup>2</sup> We have adopted the simple convention of using upper case letters (e.g.,  $Y$ ) when referring to both variable names and data, and prefer to distinguish the two by context. This is a departure from mathematical convention, which reserves upper case exclusively for random variables.

Based on the histogram in Figure 3.1-1, we see that the value of the sample mean is close to the middle of the distribution. Note that the sample mean has the same units as the observations used to calculate it. In the Quick Formula Summary (Section 3.6), we review how the sample mean is affected when the units of the observations are changed, such as by adding a constant or multiplying by a constant.

The *sample mean* is the sum of all the observations in a sample divided by  $n$ , the number of observations.

## Variance and standard deviation

The **standard deviation** is commonly used measure of the spread of a distribution. It measures how far from the mean the observations typically are. The standard deviation is large if most observations are far from the mean, and it is small if most measurements lie close to the mean.

The standard deviation is calculated from the **variance**, another measure of spread. The standard deviation is simply the square root of the variance. The standard deviation is a more intuitive measure of the spread of a distribution, but the variance has mathematical properties that make it useful sometimes as well. The standard deviation from a sample is usually represented by the symbol  $s$ , and the sample variance is written as  $s^2$ .

To calculate the variance from a sample of data, we must first compute the deviations. A deviation is the difference between a measurement and the mean ( $Y_i - \bar{Y}$ ). Deviations for the measurements of snake undulation rate are listed in Table 3.1-1.

The best measure of the spread of this distribution isn't just the average of the deviations ( $Y_i - \bar{Y}$ ), because this average is always zero (the negative deviations cancel the positive deviations). Instead, we need to average the *squared* deviations (the third column in Table 3.1-1) to find the variance:

**Table 3.1-1** Quantities needed to calculate the standard deviation and variance of snake undulation rate ( $\bar{Y} = 1.375\text{Hz}$ ).

Observations ( $Y_i$ )	Deviations ( $Y_i - \bar{Y}$ )	Squared deviations ( $(Y_i - \bar{Y})^2$ )
0.9	-0.475	0.225625
1.2	-0.175	0.030625
1.2	-0.175	0.030625
1.3	-0.075	0.005625
1.4	0.025	0.000625
1.4	0.025	0.000625
1.6	0.225	0.050625
2.0	0.625	0.390625
Sum	0.000	0.735

$$s^2 = \frac{\sum (Y_i - \bar{Y})^2}{n - 1}.$$

By squaring each number, deviations above and below the mean contribute equally<sup>3</sup> to the variance. The summation in the numerator (top part) of the formula,  $\sum (Y_i - \bar{Y})^2$ , is called the **sum of squares** of  $Y$ . Note that the denominator (bottom part) is  $n - 1$  instead of  $n$ , the total number of observations. Dividing by  $n - 1$  gives a more accurate estimate of the population variance.<sup>4</sup> We provide a shortcut formula for the variance in the Quick Formula Summary (Section 3.6).

For the snake undulation data, the variance (rounded to hundredths) is

$$s^2 = \frac{0.735}{7} = 0.11 \text{ Hz}^2.$$

The variance has units equal to the square of the units of the original data. To obtain the standard deviation, we take the square root of the variance:

$$s = \sqrt{\frac{\sum (Y_i - \bar{Y})^2}{n - 1}}.$$

For the snake undulation data,

$$s = \sqrt{\frac{0.735}{7}} = 0.324037 \text{ Hz}.$$

The standard deviation is always positive and has the same units as the observations from which it was calculated.

The *standard deviation* is a common measure of the spread of a distribution. It indicates just how different measurements typically are from the mean.

The standard deviation has a straightforward connection to the frequency distribution. If the frequency distribution is bell-shaped, as in Figure 2.1-3, then about two-thirds of the observations will lie within one standard deviation of the mean, and about 95% will lie within two standard deviations. In other words, about 67% of the data will fall between  $\bar{Y} - s$  and  $\bar{Y} + s$ , about 95% will fall between  $\bar{Y} - 2s$  and  $\bar{Y} + 2s$ . For an in-depth discussion of standard deviation, see Chapter 10.

This straightforward connection between the standard deviation and the frequency distribution diminishes when the frequency distribution deviates from the bell-shaped

<sup>3</sup> We could have averaged the absolute values of the deviations instead, to yield the mean absolute deviation. Averaging the square of the deviations is more common because the result, the variance, has many more useful mathematical properties.

<sup>4</sup> The reason is that the average squared deviation of the observations around the sample mean is slightly smaller than the average of the squared deviations around the population mean (the quantity being estimated by the sample mean). Dividing by  $n - 1$  instead of  $n$  corrects for this bias.

(normal) distribution. In such cases, the standard deviation is less informative about where the data lie in relation to the mean. This point is explored in greater detail in Section 3.3.

In the Quick Formula Summary (Section 3.6), we review how the sample variance and standard deviation are affected when the units of the observations are changed, such as by adding a constant or multiplying by a constant.

## Rounding means, standard deviations, and other quantities

To avoid rounding errors when carrying out calculations of means, standard deviations, and other descriptive statistics, always retain as many significant digits as your calculator or computer can provide. Intermediate results written down on a page should also retain as many digits as feasible. Final results, however, should be rounded before being presented.

There are no strict rules on the number of significant digits that should be retained when rounding. A common strategy, which we adopt here, is to round descriptive statistics to one decimal point more than the measurements themselves. For example, the undulation rates in snakes were measured to a single decimal place (tenths). We therefore present descriptive statistics with two decimals (hundredths). The mean rate of undulation for the eight snakes, calculated as 1.375 Hz, would be communicated as

$$\bar{Y} = 1.38 \text{ Hz.}$$

Similarly, the standard deviation, calculated as 0.324037 Hz, would be reported as

$$s = 0.32 \text{ Hz.}$$

Note that even though we report the rounded value of the mean as  $\bar{Y} = 1.38$ , we used the more exact value,  $\bar{Y} = 1.375$ , in the calculation of  $s$  to avoid rounding errors.

## Coefficient of variation

For many traits, standard deviation and mean change together when organisms of different sizes are compared. Elephants have greater mass than mice and also more variability in mass. For many purposes, we care more about the relative variation among individuals. A gain of 10 g for an elephant is inconsequential, but it would double the mass of a mouse. On the other hand, an elephant that is 10% larger than the elephant mean may have something in common with a mouse that is 10% larger than the mouse mean. For these reasons, it is sometimes useful to express the standard deviation relative to the mean. The **coefficient of variation** (CV) calculates the standard deviation as a percentage of the mean:

$$\text{CV} = 100\% \frac{s}{\bar{Y}}.$$

A higher CV means that there is more variability, whereas a lower CV means that individuals are more consistently the same. For the snake undulation data, the coefficient of variation is

$$CV = 100\% \frac{0.324}{1.375} = 24\%.$$

The coefficient of variation only makes sense when all of the measurements are greater than or equal to zero.

The *coefficient of variation* is the standard deviation expressed as a percentage of the mean.

The coefficient of variation can also be used to compare the variability of traits that do not have the same units. If we wanted to ask, “What is more variable in elephants, body mass or lifespan?” then the standard deviation is not very informative, because mass is measured in kilograms and lifespan is measured in years. The coefficient of variation would allow us to make this comparison.

## Calculating mean and standard deviation from a frequency table

Sometimes the data include many tied observations and are given in a frequency table. The frequency table in Table 3.1-2, for example, lists the number of criminal

**Table 3.1-2 Number of criminal convictions of a cohort of 395 boys.**

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

convictions of a cohort of 395 boys (Farrington 1994; see Assignment Problem 16 in Chapter 2).

To calculate the mean and standard deviation of the number of convictions, notice first of all that the sample size is *not* 15, the number of rows in Table 3.1-2, but 395, the frequency total:

$$n = 265 + 49 + 21 + 19 + \dots + 2 = 395.$$

Calculating the mean thus requires that the measurement of “0” be represented 265 times, the number “1” be represented 49 times, and so on. The sum of the measurements is thus

$$\begin{aligned} \sum Y_i &= (265 \times 0) + (49 \times 1) + (21 \times 2) + (19 \times 3) \\ &\quad + \dots + (2 \times 14) = 445. \end{aligned}$$

The mean of these data is then

$$\bar{Y} = \frac{445}{395} = 1.126582,$$

which we round to  $\bar{Y} = 1.1$  when presenting the results.

The calculation of standard deviation must also take into account the number of individuals with each value. The sum of the squared deviations is

$$\begin{aligned} \sum (Y_i - \bar{Y})^2 &= 265(0 - \bar{Y})^2 + 49(1 - \bar{Y})^2 + 21(2 - \bar{Y})^2 + \dots \\ &\quad + 2(14 - \bar{Y})^2 = 2377.671. \end{aligned}$$

The standard deviation for these data is therefore

$$s = \sqrt{\frac{2377.671}{395 - 1}} = 2.4566,$$

which we present as  $s = 2.5$ .

These calculations assume that all the data are presented in the table. It would not work, however, for frequency tables in which the data are grouped into intervals, such as Table 2.1-3.

## 3.2 Median and interquartile range

After the sample mean, the *median* is the next most common metric used to describe the location of a frequency distribution. It is often displayed in a box plot alongside the *interquartile range*, another measure of the spread of the distribution. We define and demonstrate these concepts with the help of Example 3.2.

### Example 3.2

#### I'd give my right arm for a female

Male spiders in the genus *Tidarren* are tiny, weighing only about 1% as much as females. They also have disproportionately large pedipalps, copulatory organs that make up about

10% of a male's mass. (See the adjacent photo; the pedipalps are indicated by arrows.) Males load the pedipalps with sperm and then search for females to inseminate. Astonishingly, male *Tidarren* spiders voluntarily amputate one of their two organs, right or left, just before sexual maturity. Why do they do this? Perhaps speed is important to males searching for females, and amputation increases running performance. To test this hypothesis, Ramos et al. (2004) used video to measure the running speed of males on strands of spider silk. The data are presented in Table 3.2-1.



**Table 3.2-1 Running speed (cm/s) of male *Tidarren* spiders before and after voluntary amputation of a pedipalp.**

Spider	Speed before	Speed after	Spider	Speed before	Speed after
1	1.25	2.40	9	2.98	3.70
2	2.94	3.50	10	3.55	4.70
3	2.38	4.49	11	2.84	4.94
4	3.09	3.17	12	1.64	5.06
5	3.41	5.26	13	3.22	3.22
6	3.00	3.22	14	2.87	3.52
7	2.31	2.32	15	2.37	5.45
8	2.93	3.31	16	1.91	3.40

## The median

The **median** is the middle observation in a set of data, the measurement that partitions the ordered measurements into two halves. In other words, the median is the 50th percentile and the 0.5 quantile. To calculate the median, first sort the sample observations from smallest to largest. The sorted measurements of running speed of male spiders before amputation (Table 3.2-1) are 1.25, 1.64, 1.91, 2.31, 2.37, 2.38, 2.84, 2.87, 2.93, 2.94, 2.98, 3.00, 3.09, 3.22, 3.41, 3.55 in cm/s. Let  $Y_{(i)}$  refer to the  $i$ th sorted observation, so  $Y_{(1)}$  is 1.25,  $Y_{(2)}$  is 1.64,  $Y_{(3)}$  is 1.91, and so on. If the number of observations ( $n$ ) is odd, then the median is the middle observation:

$$\text{Median} = Y_{(\lfloor n+1 \rfloor / 2)}$$

If the number of observations is even, as in the spider data, then the median is the average of the middle pair:

$$\text{Median} = [Y_{(n/2)} + Y_{(n/2+1)}] / 2.$$

Thus,  $n/2 = 8$ ,  $Y_{(8)} = 2.87$ , and  $Y_{(9)} = 2.93$  for the spider data (before amputation). The median is the average of these two numbers:

$$\text{Median} = (2.87 + 2.93) / 2 = 2.90 \text{ cm/s.}$$

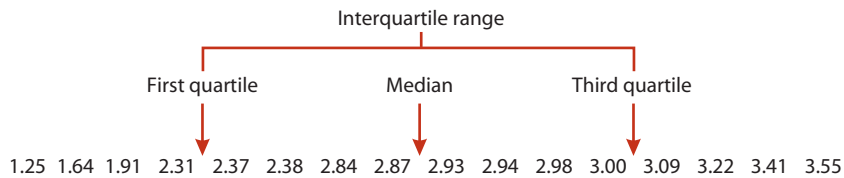
The *median* is the middle measurement of a set of observations.

## The interquartile range

Quartiles are values that partition the data into quarters. The first quartile (the 25th percentile) is the middle value of the measurements lying below the median. The second quartile is the median. The third quartile (the 75th percentile) is the middle value of the measurements larger than the median. The **interquartile range** is the span of the middle half of the data, from the first quartile to the third quartile:

$$\text{Interquartile range} = \text{third quartile} - \text{first quartile}.$$

Figure 3.2-1 shows the meaning of the median, first quartile, third quartile, and interquartile range for the spider data set (before amputation).



**Figure 3.2-1** The first quartile, median, and third quartile break the data set into four equal portions. The median is the middle value, and the first and third quartiles are the middles of the first and second halves of the data. The interquartile range is the span of the middle half of the data.

The first step to calculating the interquartile range is to break the data into halves, leaving out one observation equal to the median if  $n$  is odd.<sup>5</sup>

The smaller half of the before-amputation data is

$$1.25, 1.64, 1.91, 2.31, 2.37, 2.38, 2.84, 2.87.$$

The first quartile is the median of these observations:

$$\text{First quartile} = (2.31 + 2.37)/2 = 2.34.$$

The larger half of the before-amputation measurements is

$$2.93, 2.94, 2.98, 3.00, 3.09, 3.22, 3.41, 3.55,$$

so the third quartile is the median of these numbers:

$$\text{Third quartile} = (3.00 + 3.09)/2 = 3.045.$$

<sup>5</sup> Don't be surprised if your computer program gives slightly different values from ours for the quartiles and the interquartile range. The method given here is simple to calculate, but it does not give the most accurate estimates of the population quantities. Several improved methods are available (Hyndman and Fan 1996).

The interquartile range is then

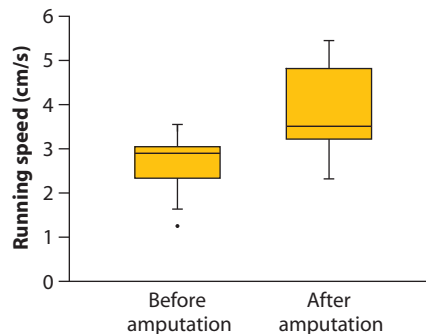
$$\text{Interquartile range} = 3.045 - 2.34 = 0.705 \text{ cm/s.}$$

The *interquartile range* is the difference between the first and third quartiles of the data.

## The box plot

A **box plot** displays the median and interquartile range, along with other quantities of the frequency distribution. Figure 3.2-2 shows a box plot for the spider running-speeds data, both before and after amputation.

Each box in Figure 3.2-2 is bounded at the lower edge by the first quartile and at the upper edge by the third quartile. The horizontal line dividing each box is the median. Two lines, called whiskers, extend outward from the box at each end. The whiskers stop at the smallest and largest “non-extreme” values in the data. “Extreme” values are defined as those lying farther from the box edge than 1.5 times the interquartile range. Extreme values are plotted as isolated dots past the ends of the whiskers.<sup>6</sup> There is one extreme value in the box plots shown in Figure 3.2-2, the smallest measurement for running speed before amputation.



**Figure 3.2-2** Box plot of the running speed of 16 male spiders before and after self-amputation of a pedipalp.

A box plot indicates where the bulk of the measurements lie. A histogram can do the same with greater detail, but a box plot picks out a few of the most important features of the frequency distribution. Because it is compact, a box plot is especially useful when comparing more than one distribution.

<sup>6</sup> Some computer programs extend whiskers all the way to the most extreme values on each end and do not indicate extreme values with isolated dots.

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

### 3.3 How measures of location and spread compare

Which measure of location (the sample mean or the median) is most revealing about the center of a distribution of measurements? And which measure of spread (the standard deviation or the interquartile range) best describes how widely the observations are scattered about the center? To answer these questions, we must compare the behavior of the mean and median, and of the standard deviation and interquartile range, with frequency distributions of different shapes. These alternative measures of location and of spread yield similar information when the frequency distribution is symmetric and unimodal, but the mean and standard deviation become less informative than the median and interquartile range when the data include extreme observations. We compare these measures using Example 3.3.

#### Example 3.3

#### Disarming fish

The marine threespine stickleback is a small coastal fish named for its defensive armor. It has three sharp spines down its back, two pelvic spines underneath, and a series of lateral bony plates down both sides. The armor seems to reduce mortality from predatory fish and diving birds. In contrast, in lakes and streams, where predators are fewer, stickleback populations have reduced armor. (See the photo in the margin for examples of both types. Bony tissue has been stained red to make it more visible.)

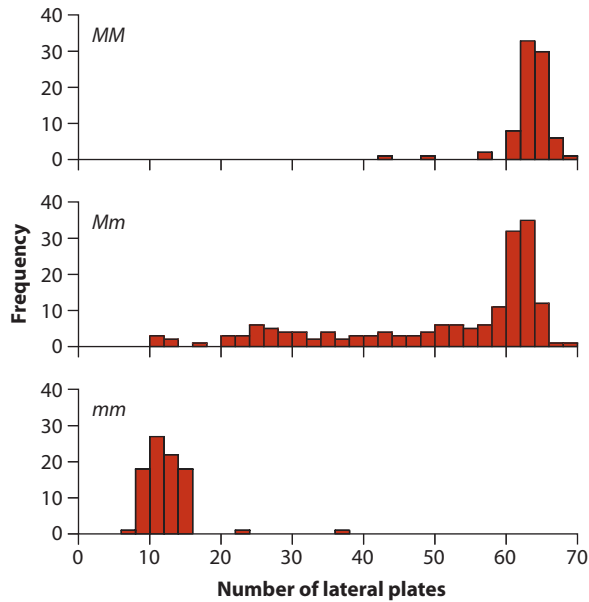


Colosimo et al. (2004) measured the grandchildren of a cross made between a marine and a freshwater stickleback. The study found that much of the difference in number of plates is caused by a single gene, *ectodysplasin*.<sup>7</sup> Fish inheriting two copies of the gene from the marine grandparent, called *MM* fish, had many plates (the top histogram in Figure 3.3-1). Fish inheriting both copies of the gene from the freshwater grandparent (*mm*) had few plates (the bottom histogram in Figure 3.3-1). Fish having one copy from each grandparent (*Mm*) had any of a wide range of plate numbers (the middle histogram in Figure 3.3-1).

<sup>7</sup> Mutations in the same gene in humans cause the loss of hair, teeth, and sweat glands.

## Mean versus median

The mean and median of the three distributions in Figure 3.3-1 are compared in Table 3.3-1. The two measures of location give similar values in the case of the *MM* and *mm* genotypes, whose distributions are fairly symmetric. The mean is smaller than the median in the case of the *Mm* fish, however, whose distribution is strongly asymmetric.



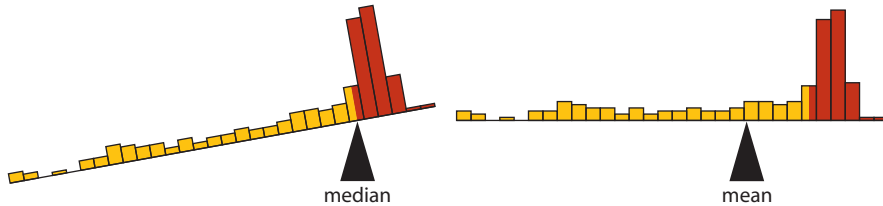
**Figure 3.3-1** Frequency distributions of lateral plate number in three genotypes of stickleback, *MM*, *Mm*, and *mm*, descended from a cross between marine and freshwater grandparents. Plates are counted as the total number down the left and right sides of the fish. The total number of fish: 82 (*MM*), 174 (*Mm*), and 88 (*mm*).

**Table 3.3-1** Descriptive statistics for the number of lateral plates of the three genotypes of threespine sticklebacks<sup>8</sup> discussed in Example 3.3.

Genotype	<i>n</i>	$\bar{Y}$	Median	<i>s</i>	Interquartile range
<i>MM</i>	82	62.8	63	3.4	2
<i>Mm</i>	174	50.4	59	15.1	21
<i>mm</i>	88	11.7	11	3.6	3

<sup>8</sup> When listing descriptive statistics in tables, put the same quantity calculated on different groups into one column. Numbers stacked in a single column are easier to compare than numbers placed side by side in a row. Exchange the columns and rows in Table 3.3-1, and you'll see what we mean.

Why are the median and mean different from one another when the distribution is asymmetric? The answer, shown in Figure 3.3-2, is that the median is the middle measurement of a distribution, whereas the mean is the “center of gravity.”



**Figure 3.3-2** Comparison between the median and the mean using the frequency distribution for the *Mm* genotype (middle panel of Figure 3.3-1). The median is the middle measurement of the distribution (different colors represent the two halves of the distribution). The mean is the center of gravity, the point at which the frequency distribution would be balanced (if observations had weight).

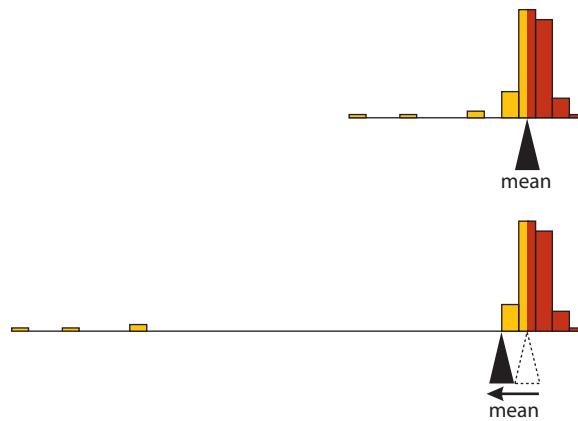
The balancing act illustrated in Figure 3.3-2 suggests that the mean is sensitive to extreme observations. To demonstrate, imagine taking the four smallest observations of the *MM* genotype (top panel in Figure 3.3-1) and moving them far to the left. The median would be completely unaffected, but the mean would shift leftward to a point near the edge of the range of most observations (Figure 3.3-3).

Median and mean measure different aspects of the location of a distribution. The median is the middle value of the data, whereas the mean is its center of gravity.

Thus, the mean is displaced from the location of the “typical” measurement when the frequency distribution is strongly skewed, particularly when there are extreme observations. The mean is still useful as a description of the population as a whole, but it no longer indicates where most of the observations are located. The median is less sensitive to extreme observations, and hence the median is the more informative descriptor of the typical observation in such instances.

## Standard deviation versus interquartile range

Because it is calculated from the square of the deviations, the standard deviation is even more sensitive to extreme observations than is the mean. When the four smallest observations of the *MM* genotype are shifted far to the left, such that the smallest is set to zero (Figure 3.3-3), the standard deviation jumps from 3.4 to 12.0, whereas the interquartile range is not affected. For this reason, the interquartile range is a better indicator of the spread of the main part of a distribution than the standard deviation when the data are strongly skewed to one side or the other, especially when there are extreme observations. On the other hand, the standard deviation is a better indication of the spread among all of the data points.



**Figure 3.3-3** Sensitivity of the mean to extreme observations using the frequency distribution of the *MM* genotypes (see the upper panel in Figure 3.3-1). The two different colors represent the two halves of the distribution. When the four smallest observations of the *MM* genotype are shifted far to the left (lower panel), the mean is displaced downward, to the edge of the range of the bulk of the observations. The median, on the other hand, which is located where the two colors meet, is unaffected by the shift.

## 3.4 Proportions

The proportion is the most important descriptive statistic for a categorical variable.

### Calculating a proportion

The **proportion** of observations in a given category, symbolized  $\hat{p}$ , is calculated as

$$\hat{p} = \frac{\text{Number in category}}{n},$$

where the numerator is the number of observations in the category of interest, and  $n$  is the total number of observations in all categories combined.<sup>9</sup>

For example, of the 344 individual sticklebacks in Example 3.3, 82 had genotype *MM*, 88 were *mm*, and 174 were *Mm* (Table 3.3-1). The proportion of *MM* fish is

$$\hat{p} = \frac{82}{344} = 0.238.$$

The other proportions are calculated similarly, and all three proportions are listed in Table 3.4-1.

<sup>9</sup> The “hat” in  $\hat{p}$  is used to indicate an estimate of the true proportion  $p$ .

**Table 3.4-1** The number of fish of each genotype from a cross between a marine stickleback and a freshwater stickleback (Example 3.3). The sum of the proportions as written does not add precisely to one because of rounding.

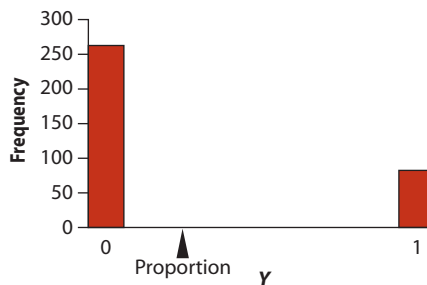
Genotype	Frequency	Proportion
<i>MM</i>	82	0.24
<i>Mm</i>	174	0.51
<i>mm</i>	88	0.26
Total	344	1.00

### The proportion is like a sample mean

The proportion  $\hat{p}$  has properties in common with the arithmetic mean. To see this, let's create a new numerical variable  $Y$  for the stickleback study. Give individual fish  $i$  the value  $Y_i = 1$  if it has the *MM* genotype and give it the value  $Y_i = 0$  otherwise. The sum of all the ones and zeroes,  $\sum Y_i$ , is the frequency of fish having genotype *MM*. The mean of the ones and zeroes is

$$\bar{Y} = \frac{\sum Y_i}{n} = \frac{82}{344} = 0.238,$$

which is just  $\hat{p}$ , the proportion of observations in the first category. If we imagine the  $Y$ -measurements to have weight, then the proportion is their center of gravity (Figure 3.4-1).



**Figure 3.4-1** The distribution of  $Y$ , where  $Y = 1$  if a stickleback is genotype *MM* and 0 otherwise. The mean of  $Y$  is the proportion of *MM* individuals in the sample (0.238).

## 3.5 Summary

- The location of a distribution for a numerical variable can be measured by its mean or by its median. The mean gives the center of gravity of the distribution

and is calculated as the sum of all measurements divided by the number of measurements. The median gives the middle value.

- ▶ The standard deviation measures the spread of a distribution for a numerical variable. It is a measure of the typical distance between observations and the mean. The variance is the square of the standard deviation.
- ▶ The quartiles break the ordered observations into four equal parts. The interquartile range, the difference between the first and third quartiles, is another measure of the spread of a frequency distribution.
- ▶ The mean and median yield similar information when the frequency distribution of the measurements is symmetric and unimodal. The mean and standard deviation become less informative about the location and spread of typical observations than the median and interquartile range when the data include extreme observations.
- ▶ A box plot is a graphical method used to display important features of a frequency distribution. It shows the median, the first and third quartiles, the range of values, and any extreme values.
- ▶ The proportion is the most important descriptive statistic for a categorical variable. It is calculated by dividing the number of observations in the category of interest by  $n$ , the total number of observations in all categories combined.

## 3.6 Quick Formula Summary

### Table of formulas for descriptive statistics

Quantity	Formula
Sample size	$n$
Mean	$\bar{Y} = \frac{\sum Y}{n}$
Variance	$s^2 = \frac{\sum (Y_i - \bar{Y})^2}{n - 1}$
shortcut formula:	$s^2 = \frac{\sum (Y_i^2) - n\bar{Y}^2}{n - 1}$
Standard deviation	$s = \sqrt{\frac{\sum (Y_i - \bar{Y})^2}{n - 1}}$
shortcut formula:	$s = \sqrt{\frac{\sum (Y_i^2) - n\bar{Y}^2}{n - 1}}$
Sum of squares	$\sum (Y_i - \bar{Y})^2 = \sum (Y_i^2) - n\bar{Y}^2$

Coefficient of variation	$CV = 100\% \frac{s}{\bar{Y}}$
Median	$Y_{((n+1)/2)}$ (if $n$ is odd) $(Y_{(n/2)} + Y_{(n/2+1)})/2$ (if $n$ is even), where $Y_{(1)}, Y_{(2)}, \dots, Y_{(n)}$ are the ordered observations
Proportion	$\hat{p} = \frac{\text{Number in category}}{n}$

### Effect of arithmetic operations on descriptive statistics

The table below lists the effect on the descriptive statistics of adding or multiplying all the measurements by a constant. The rules listed in the table are useful when converting measurements from one system of units to another, such as English to metric or degrees Celsius to degrees Fahrenheit.

For example, if the mean temperature is 27°C with a standard deviation of 3°C, then measurements in Celsius can be converted to equivalent measurements in Fahrenheit by the transformation

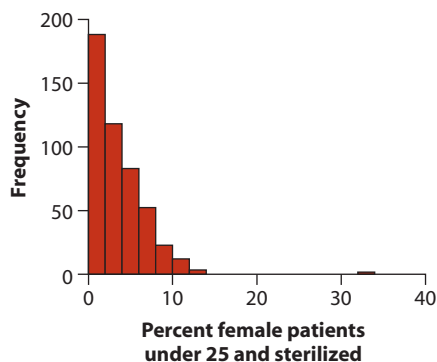
$$^{\circ}\text{F} = 9/5^{\circ}\text{C} + 32.$$

Thus, the new mean temperature is  $9/5(27) + 32 = 80.6^{\circ}\text{F}$ , and the standard deviation is  $9/5(3) = 5.4^{\circ}\text{F}$ .

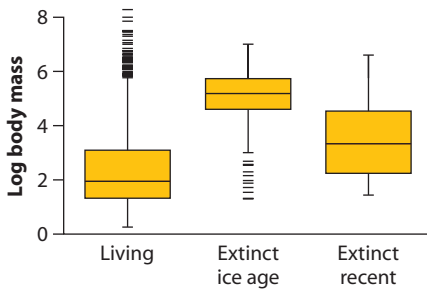
Quantity	Symbol	Effect of adding a constant $c$ to all the measurements, $Y' = Y + c$	Effect of multiplying all the measurements by a constant $c$ , $Y' = cY$
Mean	$\bar{Y}$	$\bar{Y}' = \bar{Y} + c$	$\bar{Y}' = c\bar{Y}$
Standard deviation	$s$	$s' = s$	$s' = cs$
Variance	$s^2$	$s'^2 = s^2$	$s'^2 = c^2s^2$
Median	$M$	$M' = M + c$	$M' = cM$
Interquartile range	$IR$	$IR' = IR$	$IR' = cIR$

## PRACTICE PROBLEMS

1. A review of the performance of hospital gynecologists in two regions of England measured the outcomes of patient admissions under each doctor's care (Harley et al. 2005). One measurement taken was the percentage of patient admissions made up of women under 25 years old who were sterilized. We are interested in describing what constitutes a typical rate of sterilization, so that the behavior of atypical doctors can be better scrutinized. The frequency distribution of this measurement for all doctors is plotted to the right.

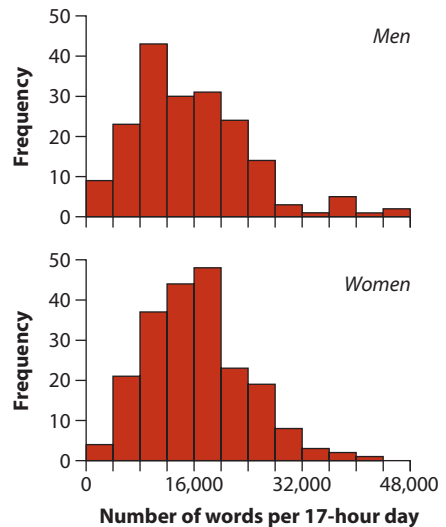


- a. Explain what the vertical axis measures.
- b. What would be the best choice to describe the location of this frequency distribution, the mean or the median, if our goal was to describe the typical individual? Why?
2. The data displayed in the plot below are from a nearly complete record of body masses (in grams, then converted to log base-10) of the world's native mammals (Smith et al. 2003). The data were divided into three groups: those surviving from the last ice age to the present day ( $n = 4061$ ), those who went extinct around the end of the last ice age ( $n = 227$ ), and those driven extinct within the last 300 years (recent;  $n = 44$ ).



- a. What type of graph is this?
- b. What does the horizontal line in the center of each rectangle represent?
- c. What are the horizontal lines at the top and bottom edges of each rectangle supposed to represent?
- d. What are the dashes lying outside of the rectangles?
- e. What are the vertical lines extending above and below each rectangle?
- f. Compare the locations of the three body size distributions. How do they differ?
- g. Compare the shapes of the three frequency distributions. Which are relatively symmetric and which are asymmetric? Explain your reasoning.
- h. Which group's frequency distribution has the lowest spread? Explain your reasoning.
- i. What has been the likely effect of ice-age and recent extinctions on the median body size of mammals?

3. Mehl et al. (2007) wired 396 volunteers with electronically activated recorders that allowed the researchers to calculate the number of words each individual spoke, on average, per 17-hour waking day. They found that the mean number of words spoken was only very slightly higher for the 210 women (16,215 words) than the 186 men (15,669) in the sample. The frequency distribution of the number of words spoken by all individuals of each sex is shown below (modified from Mehl et al. 2007).



- a. What type of graph is shown?
- b. What are the explanatory and response variables in the figure?
- c. What is the mode of the frequency distribution of each sex?
- d. Which sex likely has the higher median number of words spoken per day?
- e. Which sex had the highest variance in number of words spoken per day?
4. The following data are measurements of body mass, in grams, of finches captured in mist nets during a survey of various habitats in Kenya, East Africa (Schluter 1988).

<i>Crimson-rumped waxbill</i>	8, 8, 8, 8, 8, 8, 8, 6,
	7, 7, 7, 8, 8, 8, 7, 7,
<i>Cutthroat finch</i>	16, 16, 16, 12, 16, 15,
	15, 17, 15, 16, 15, 16

*White-browed sparrow weaver* 40, 43, 37, 38, 43, 33, 35, 37, 36, 42, 36, 36, 39, 37, 34, 41



- Calculate the mean body mass of each of these three finch species. Which species is largest, and which is smallest? (*Here and forever after, provide units with your answer.*)
  - Which species has the greatest standard deviation in body mass? Which has the least?
  - Calculate the coefficient of variation (CV) in mass for each finch species. How different are the coefficients between the species? Compare the difference in CVs with the differences in standard deviation calculated in part (b).
  - Given the following measurements of beak length, in mm, of the 16 white-browed sparrow weavers, which measurement is more variable in this species (relative to the mean), body mass or beak length?  
10.6, 10.8, 10.9, 11.3, 10.9, 10.1, 10.7, 10.7, 10.9, 11.4, 10.8, 11.2, 10.7, 10.0, 10.1, 10.7
- Assignment Problem 22 in Chapter 2 presents a cumulative distribution function of the population growth rates of 204 countries recognized by the United Nations.
    - Draw a box plot using the information from the graph in that problem.
    - Label three features of this box plot.
  - The spider data in Example 3.2 consist of *pairs* of measurements made on the same subjects. One measurement is running speed before amputation and the second is running speed after amputation. Calculate a new variable called “change in speed,” defined as the speed after amputation minus the speed before amputation.
    - What are the units of the new variable?
    - Draw a box plot for the change in running speed. Use the method outlined in Section 3.2 to calculate the quartiles.
    - Based on your drawing in part (b), is the frequency distribution of the change in running speed symmetric or asymmetric? Explain how you decided this.
    - What is the quantity measured by the span of the box in part (b)?
    - Calculate the mean change in running speed. Is it the same as the median? Why or why not?
    - Calculate the variance of the change in running speed.
    - What fraction of observations fall within one standard deviation above and below the mean?
  - If you were to add a constant  $k$  to each measurement in Practice Problem 6, what effect would this have on
    - the mean?
    - the standard deviation?
  - If you were to convert all of the observations of change in running speed in Practice Problem 6 from cm/s into mm/s, how would this change
    - the mean?
    - the standard deviation?
    - the median?
    - the interquartile range?
    - the coefficient of variation?
    - the variance?
  - Niderkorn’s (1872; from Pounder 1995) measurements on 114 human corpses provided the first quantitative study on the development of rigor

mortis.<sup>10</sup> The data in the following table give the number of bodies achieving rigor mortis in each hour after death, recorded in one-hour intervals.

Hours	Number of bodies
1	0
2	2
3	14
4	31
5	14
6	20
7	11
8	7
9	4
10	7
11	1
12	1
13	2
Total	114

- Calculate the mean number of hours after death that it took for rigor mortis to set in.
- Calculate the standard deviation in the number of hours until rigor mortis.
- What fraction of observations lie within one standard deviation of the mean (i.e., between  $\bar{Y} - s$  and  $\bar{Y} + s$ )?
- Calculate the median number of hours until rigor mortis sets in. What is the likely explanation for the difference between the median and the mean?

## ASSIGNMENT PROBLEMS

10. The following data are measurements of the number of young cod that “recruited” (grew to the catchable size) to the North Sea population in different years (Beaugrand et al. 2003). The measurements are listed below in  $\log_{10}$  units (i.e., the measurement 5.0 represents  $10^{5.0}$  recruits) and arranged from low to high rather than by year.

5.0, 5.1, 5.2, 5.2, 5.2, 5.2, 5.2, 5.2, 5.3, 5.3, 5.3, 5.3, 5.4, 5.4, 5.4, 5.4, 5.4, 5.5, 5.5, 5.5, 5.5, 5.5, 5.5, 5.5, 5.6, 5.6, 5.6, 5.7, 5.7, 5.7, 5.7, 5.7, 5.8, 5.8, 5.8, 5.9, 5.9, 6.0, 6.0

- What was the mean number of recruits (in  $\log_{10}$  units) over this period?
- What was the standard deviation in the number of recruits (in  $\log_{10}$  units)?

- In what fraction of years did the number of recruits fall within two standard deviations of the mean?

11. The gene for the vasopressin receptor *V1a* is expressed at higher levels in the forebrain of monogamous vole species than promiscuous vole species.<sup>11</sup> Can expression of this gene influence monogamy? To test this, Lim et al. (2004) experimentally enhanced *V1a* expression in the forebrain of 11 males of the meadow vole, a solitary promiscuous species. The percentage of time each male spent huddling with the female provided to him (an index for monogamy) was recorded. The same measurements were taken in 20 control males left untreated.

<sup>10</sup> Rigor mortis is the muscular stiffening that occurs after death. It is caused by linkages forming between actin and myosin in muscle when muscle glycogen is depleted, pH drops, and the level of ATP falls below a critical level.

<sup>11</sup> In monogamous vole species, single males and females form stable mating pairs. In promiscuous voles, no stable pairs form and voles might mate with multiple partners.

Control males: 98, 96, 94, 88, 86, 82, 77, 74, 70, 60, 59, 52, 50, 47, 40, 35, 29, 13, 6, 5

*Vla*-enhanced males: 100, 97, 96, 97, 93, 89, 88, 84, 77, 67, 61

- Display these data in a graph.
- Which group has the higher mean percentage of time spent huddling with females?
- Which group has the higher standard deviation in percentage of time spent huddling with females?

- 12.** Birds of the Caribbean islands of the Lesser Antilles are descended from immigrants originating from larger islands and the nearby mainland. The data presented here are the approximate dates of immigration, in millions of years, of each of 37 bird species now present on the Lesser Antilles (Ricklefs and Bermingham 2001). The dates were calculated from the difference in mitochondrial DNA sequences between each of the species and its closest living relative on larger islands or the mainland.

0.00, 0.00, 0.04, 0.21, 0.29, 0.54, 0.63, 0.88, 0.96, 1.25, 1.67, 1.75, 1.84, 1.96, 2.01, 2.51, 2.72, 3.30, 3.51, 4.05, 4.85, 6.94, 8.73, 10.57, 11.11, 12.45, 14.00, 17.30, 17.92, 18.05, 18.43, 22.48, 22.48, 23.48, 26.32, 26.45, 28.87

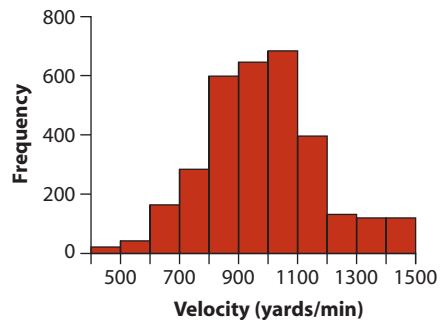
- Plot the data in a histogram and describe the shape of the frequency distribution.
- By viewing the graph alone, approximate the mean and median of the distribution. Which should be greater? Explain your reasoning.
- Calculate the mean and median. Was your intuition in part (b) correct?
- Calculate the first and third quartiles and the interquartile range.
- Draw a box plot for these data.

- 13.** The data in the accompanying table are from an ecological study of the entire rainforest community at El Verde in Puerto Rico (Waide and Reagan 1996). Diet breadth is the number of types of food eaten by an animal species. The number of animal species having each diet breadth is shown in the second column. The total number of species listed is  $n = 127$ .

Diet breadth (number of prey types eaten)	Frequency (number of species)
1	21
2	8
3	9
4	10
5	8
6	3
7	4
8	8
9	4
10	4
11	4
12	2
13	5
14	2
15	1
16	1
17	2
18	1
19	3
20	2
> 20	25
Total	127

- Calculate the median number of prey types consumed by animal species in the community.
- What is the interquartile range in the number of prey types? Use the method outlined in Section 3.2 to calculate the quartiles.
- Can you calculate the mean number of prey types in the diet?

- 14.** Francis Galton (1894) presented the following data on the flight speeds of 3207 “old” homing pigeons traveling at least 90 miles.

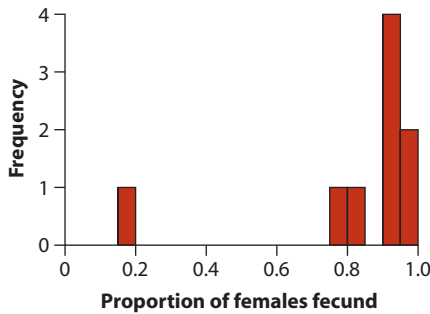


- a. What type of graph is this?
- b. Examine the graph and visually determine the approximate value of the mean (to the nearest 100 yards per minute). Explain how you obtained your estimate.
- c. Examine the graph and visually determine the approximate value of the median (to the nearest 100 yards per minute). Explain how you obtained your estimate.
- d. Examine the graph and visually determine the approximate value of the mode (to the nearest 100 yards per minute). Explain how you obtained your estimate.
- e. Examine the graph and visually determine the approximate value of the standard deviation (to the nearest 50 yards per minute). Explain how you obtained your estimate.

15. On the basis of your answer to Practice Problem 8, describe how each of the following quantities is changed if every measurement in a data set were multiplied by a constant  $k$ :

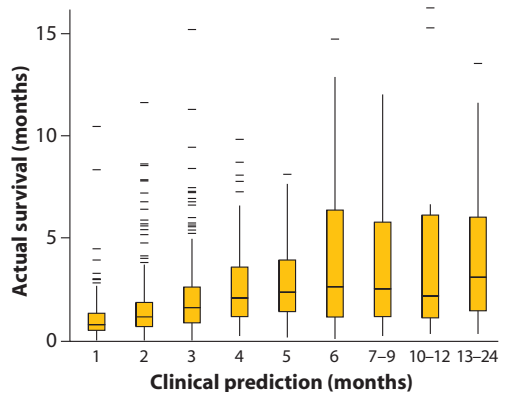
- a. the mean
- b. the standard deviation
- c. the median
- d. the interquartile range
- e. the coefficient of variation
- f. the variance

16. A study of the endangered saiga antelope (pictured at the beginning of the chapter) recorded the fraction of females in the population that were fecund in each year between 1993 and 2001 (Milner-Gulland et al. 2003). A graph of the data is as follows:



- a. Assume that you want to describe the fraction of females that are fecund in a typical year, based on these data. What would be the better choice to describe this fraction, the mean or the median of the measurements? Why?
- b. With the same goal in mind, what would be the better choice to describe the spread of measurements around their center, the standard deviation or the interquartile range? Why?

17. Accurate prediction of the timing of death in patients with a terminal illness is important for their care. The following graph compares the survival times of terminally ill cancer patients with the clinical prediction of their survival times (modified from Glare et al. 2003).



- a. Describe in words what features most of the frequency distributions of actual survival times have in common, based on the box plots for each group.
- b. Describe the differences in shape of actual survival time distributions between those for one to five months predicted survival times and those for six to 24 months.
- c. Describe the trend in median actual survival time with increasing predicted number.
- d. The predicted survival times of terminally ill cancer patients tend to overestimate the medians of actual survival times. Are the *means* of actual survival times likely to be closer to, further from, or no different from

the predicted times than the medians?  
Explain.

18. Measurements of lifetime reproductive success (LRS) of individual wild animals reveals the disparate contributions they make to the next generation. Jensen et al. (2004) estimated LRS of male and female house sparrows in an island population in Norway. They measured LRS of an individual as the total number of “recruits” produced in its lifetime, where a recruit is an offspring that survives to breed one year after birth. Parentage of recruits was determined from blood samples using DNA techniques. (Assessing parentage from observation alone isn’t accurate because pairs aren’t entirely faithful in most bird species.) Their results are tabulated as follows:

Lifetime reproductive success	Frequency	
	Females	Males
0	30	38
1	25	17
2	3	7
3	6	6
4	8	4
5	4	10
6	0	2
7	4	0
8	1	0
> 8	0	0
Total	81	84

- Which sex has the higher mean lifetime reproductive success?
- Can you think of a reason why sexes might have different means in these data?
- Which sex has the higher variance in reproductive success?