

CHAPTER 3

Summary Statistics: Measures of Location and Spread

Data are the essence of scientific investigations, but rarely do we report all the data that we collect. Rather, we summarize our data using **summary statistics**. Biologists and statisticians distinguish between two kinds of summary statistics: measures of **location** and measures of **spread**. Measures of location illustrate where the majority of data are found; these measures include means, medians, and modes. In contrast, measures of spread describe how variable the data are; these measures include the sample standard deviation, variance, and standard errors. In this chapter, we introduce the most common summary statistics and illustrate how they arise directly from the **Law of Large Numbers**, one of the most important theorems of probability.

Henceforth, we will adopt standard statistical notation when describing random variables and statistical quantities or estimators. Random variables will be designated as Y , where each individual observation is indexed with a subscript, Y_i . The subscript i indicates the i th observation. The size of the sample will be denoted by n , and so i can take on any integer value between 1 and n . The arithmetic mean is written as \bar{Y} . Unknown **parameters** (or population statistics) of distributions, such as expected values and variances, will be written with Greek letters (such as μ for the expected value, σ^2 for the expected variance, σ for the expected standard deviation), whereas statistical estimators of those parameters (based on real data) will be written with italic letters (such as \bar{Y} for the arithmetic mean, s^2 for the sample variance, and s for the sample standard deviation).

Throughout this chapter, we use as our example the data illustrated in Figure 2.6, the simulated measurement of tibial spines of 50 linyphiid spiders. These data, sorted in ascending order, are illustrated in Table 3.1.

TABLE 3.1 Ordered measurements of tibial spines of 50 linyphiid spiders (millimeters)

0.155	0.207	0.219	0.228	0.241	0.249	0.263	0.276	0.292	0.307
0.184	0.208	0.219	0.228	0.243	0.250	0.268	0.277	0.292	0.308
0.199	0.212	0.221	0.229	0.247	0.251	0.270	0.280	0.296	0.328
0.202	0.212	0.223	0.235	0.247	0.253	0.274	0.286	0.301	0.329
0.206	0.215	0.226	0.238	0.248	0.258	0.275	0.289	0.306	0.368

This simulated dataset is used throughout this chapter to illustrate measures of summary statistics and probability distributions. Although raw data of this sort form the basis for all of our calculations in statistics, the raw data are rarely published because they are too extensive and too difficult to comprehend. Summary statistics, if they are properly used, concisely communicate and summarize patterns in raw data without enumerating each individual observation.

Measures of Location

The Arithmetic Mean

There are many ways to summarize a set of data. The most familiar is the average, or **arithmetic mean** of the observations. The arithmetic mean is calculated as the sum of the observations (Y_i) divided by the number of observations (n) and is denoted by \bar{Y} :

$$\bar{Y} = \frac{\sum_{i=1}^n Y_i}{n} \quad (3.1)$$

For the data in Table 3.1, $\bar{Y} = 0.253$. Equation 3.1 looks similar to, but is not quite equivalent to, Equation 2.6, which was used in Chapter 2 to calculate the expected value of a discrete random variable:

$$E(X) = \sum_{i=1}^n Y_i p_i$$

where the Y_i 's are the values that the random variable can have, and the p_i 's are their probabilities. For a continuous variable in which each Y_i occurs only once, with $p_i = 1/n$, Equations 3.1 and 2.6 give identical results.

For example, let *Spine length* be the set consisting of the 50 observations in Table 3.1: $\text{Spine length} = \{0.155, 0.184, \dots, 0.329, 0.368\}$. If each element (or event) in *Spine length* is independent of all others, then the probability p_i of any of these 50 independent observations is $1/50$. Using Equation 2.6, we can calculate the expected value of *Spine length* to be

$$E(Y) = \sum_{i=1}^n Y_i p_i$$

where Y_i is the i th element and $p_i = 1/50$. This sum,

$$E(Y) = \sum_{i=1}^n Y_i \times \frac{1}{50}$$

is now equivalent to Equation 3.1, used to calculate the arithmetic mean of n observations of a random variable Y :

$$\bar{Y} = \sum_{i=1}^n p_i Y_i = \sum_{i=1}^n Y_i \times \frac{1}{50} = \frac{1}{50} \sum_{i=1}^n Y_i$$

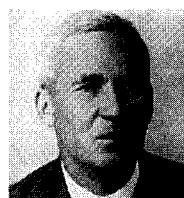
To calculate this expected value of *Spine length*, we used the formula for the expected value of a discrete random variable (Equation 2.6). However, the data given in Table 3.1 represent observations of a continuous, normal random variable. All we know about the expected value of a normal random variable is that it has some underlying true value, which we denote as μ . Does our calculated value of the mean of *Spine length* have any relationship to the unknown value of μ ?

If three conditions are satisfied, the arithmetic mean of the observations in our sample is an **unbiased estimator** of μ . These three conditions are:

1. Observations are made on randomly selected individuals.
2. Observations in the sample are independent of each other.
3. Observations are drawn from a larger population that can be described by a normal random variable.

The fact that \bar{Y} of a sample approximates μ of the population from which the sample was drawn is a special case of the second fundamental theorem of probability, the **Law of Large Numbers**.¹

Here is a description of the Law of Large Numbers. Consider an infinite set of random samples of size n , drawn from a random variable Y . Thus, Y_1 is a sample from Y with 1 datum, $\{y_1\}$. Y_2 is a sample of size 2, $\{y_1, y_2\}$, etc. The Law of Large Numbers establishes that, as the sample size n increases, the arithmetic



Andrei Kolmogorov

¹ The modern (or “strong”) version of the law of large numbers was proven by the Russian mathematician Andrei Kolmogorov (1903–1987), who also studied **Markov processes** such as those used in modern computational Bayesian analysis (see Chapter 5) and fluid mechanics.

mean of Y_i (Equation 3.1) approaches the expected value of Y , $E(Y)$. In mathematical notation, we write

$$\lim_{n \rightarrow \infty} \left(\frac{\sum_{i=1}^n Y_i}{n} = \bar{Y}_n \right) = E(Y) \quad (3.2)$$

In words, we say that as n gets very large, the average of the Y_i 's equals $E(Y)$ (see Figure 3.1).

In our example, the tibial spine lengths of all individuals of linyphiid spiders in a population can be described as a normal random variable with expected value $= \mu$. We cannot measure all of these (infinitely many) spines, but we can measure a subset of them; Table 3.1 gives $n = 50$ of these measurements. If each spine measured is from a single individual spider, each spider chosen for measurement is chosen at random, and there is no bias in our measurements, then the expected value for each observation should be the same (because they come from the same infinitely large population of spiders). The Law of Large Numbers states that the average spine length of our 50 measurements approximates the expected value of the spine length in the entire population. Hence, we can estimate the unknown expected value μ with the average of our observations. As Figure 3.1 shows, the estimate of the true population mean is more reliable as we accumulate more data.

Other Means

The arithmetic average is not the only measure of location of a set of data. In some cases, the arithmetic average will generate unexpected answers. For example, suppose a population of mule deer (*Odocoileus hemionus*) increases in size by 10% in one year and 20% in the next year. What is the average population growth rate each year?² The answer is not 15%!

You can see this discrepancy by working through some numbers. Suppose the initial population size is 1000 deer. After one year, the population size (N_1) will be $(1.10) \times 1000 = 1100$. After the second year, the population size (N_2) will be $(1.20) \times 1100 = 1320$. However, if the average growth rate were 15% per year,

² In this analysis, we use the finite rate of increase, λ , as the parameter for population growth. λ is a multiplier that operates on the population size each year, such that $N_{t+1} = \lambda N_t$. Thus, if the population increases by 10% every year, $\lambda = 1.10$, and if the population decreases by 5% every year, $\lambda = 0.95$. A closely related measure of population growth rate is the instantaneous rate of increase, r , whose units are individuals/(individuals \times time). Mathematically, $\lambda = e^r$ and $r = \ln(\lambda)$. See Gotelli (2001) for more details.

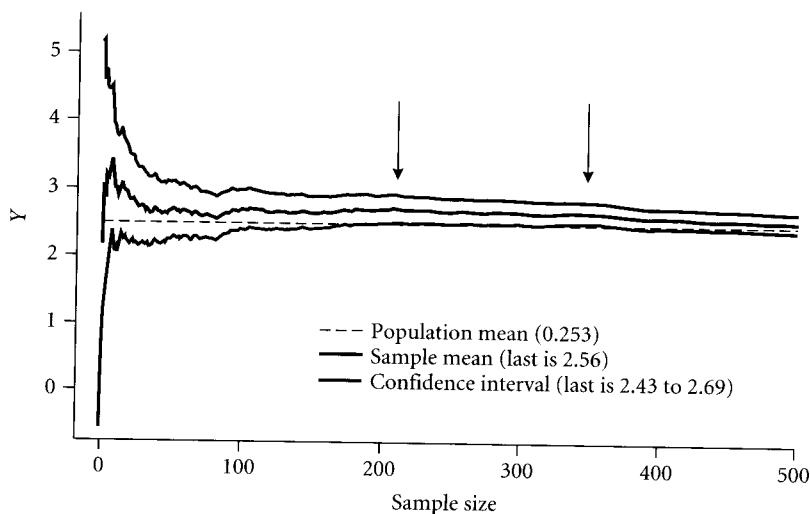


Figure 3.1 Illustration of the Law of Large Numbers and the construction of confidence intervals using the spider tibial spine data of Table 3.1. The population mean (0.253) is indicated by the dotted line. The sample mean for samples of increasing size n is indicated by the central solid line and illustrates the Law of Large Numbers: as the sample size increases, the sample mean approaches the true population mean. The upper and lower solid lines illustrate 95% confidence intervals about the sample mean. The width of the confidence interval decreases with increasing sample size. 95% of confidence intervals constructed in this way should contain the true population mean. Notice, however, that there are samples (between the arrows) for which the confidence interval does not include the true population mean. Curve constructed using algorithms and S-Plus code published by Blume and Royall (2003).

the population size would be $(1.15) \times 1000 = 1150$ after one year and then $(1.15) \times 1150 = 1322.50$ after 2 years. These numbers are close, but not identical; after several more years, the results diverge substantially.

THE GEOMETRIC MEAN In Chapter 2, we introduced the log-normal distribution: if Y is a random variable with a log-normal distribution, then the random variable $Z = \ln(Y)$ is a normal random variable. If we calculate the arithmetic mean of Z ,

$$\bar{Z} = \frac{1}{n} \sum_{i=1}^n Z_i \quad (3.3)$$

what is this value expressed in units of Y ? First, recognize that if $Z = \ln(Y)$, then $Y = e^Z$, where e is the base of the natural logarithm and equals $\sim 2.71828\dots$. Thus, the value of \bar{Z} in units of Y is $e^{\bar{Z}}$. This so-called **back-transformed mean** is called the **geometric mean** and is written as GM_Y .

The simplest way to calculate the geometric mean is to take the antilog of the arithmetic mean:

$$GM_Y = e^{\left[\frac{1}{n} \sum_{i=1}^n \ln(Y_i) \right]} \quad (3.4)$$

A nice feature of logarithms is that the sum of the logarithms of a set of numbers equals the logarithm of their products: $\ln(Y_1) + \ln(Y_2) + \dots = \ln(Y_1 Y_2 \dots Y_n)$. So another way of calculating the geometric mean is to take the n th root of the product of the observations:

$$GM_Y = \sqrt[n]{Y_1 Y_2 \dots Y_n} \quad (3.5)$$

Just as we have a special symbol for adding up a series of numbers:

$$\sum_{i=1}^n Y_i = Y_1 + Y_2 + \dots + Y_n$$

we also have a special symbol for multiplying a series of numbers:

$$\prod_{i=1}^n Y_i = Y_1 \times Y_2 \times \dots \times Y_n$$

Thus, we could also write our formula for the geometric mean as

$$GM_Y = \sqrt[n]{\prod_{i=1}^n Y_i}$$

Let's see if the geometric mean of the population growth rates does a better job of predicting average population growth rate than the arithmetic average does. First, if we express population growth rates as multipliers, the annual growth rates of 10% and 20% become 1.10 and 1.20, and the natural logarithms of these two values are $\ln(1.10) = 0.09531$ and $\ln(1.20) = 0.18232$. The arithmetic average of these two numbers is 0.138815. Back-calculating gives us a geometric mean of $GM_Y = e^{0.138815} = 1.14891$, which is slightly less than the arithmetic mean of 1.20.

Now we can calculate population growth rate over two years using this geometric mean growth rate. In the first year, the population would grow to $(1.14891) \times (1000) = 1148.91$, and in the second year to $(1.14891) \times (1148.91) = 1319.99$. This is the same answer we got with 10% growth in the first year and 20% growth in the second year $[(1.10) \times (1000) \times (1.20)] = 1320$. The values would match perfectly if we had not rounded the calculated growth rate. Notice also that although population size is always an integer variable (0.91 deer can be seen only in a theoretical forest), we treat it as a continuous variable to illustrate these calculations.

Why does GM_Y give us the correct answer? The reason is that population growth is a multiplicative process. Note that

$$\frac{N_2}{N_0} = \left(\frac{N_2}{N_1} \right) \times \left(\frac{N_1}{N_0} \right) \neq \left(\frac{N_2}{N_1} \right) + \left(\frac{N_1}{N_0} \right)$$

However, numbers that are multiplied together on an arithmetic scale can be added together on a logarithmic scale. Thus

$$\ln \left[\left(\frac{N_2}{N_1} \right) \times \left(\frac{N_1}{N_0} \right) \right] = \ln \left(\frac{N_2}{N_1} \right) + \ln \left(\frac{N_1}{N_0} \right)$$

THE HARMONIC MEAN A second kind of average can be calculated in a similar way, using the reciprocal transformation ($1/Y$). The reciprocal of the arithmetic mean of the reciprocals of a set of observations is called the **harmonic mean**:³

$$H_Y = \frac{1}{\frac{1}{n} \sum \frac{1}{Y_i}} \quad (3.6)$$

For the spine data in Table 3.1, $GM_Y = 0.249$ and $H_Y = 0.246$. Both of these means are smaller than the arithmetic mean (0.253); in general, these means are ordered as $\bar{Y} > GM_Y > H_Y$. However, if all the observations are equal ($Y_1 = Y_2 = Y_3 = \dots = Y_n$), all three of these means are identical as well ($\bar{Y} = GM_Y = H_Y$).



³The harmonic mean turns up in conservation biology and population genetics in the calculation of effective population size, which is the equivalent size of a population with completely random mating. If the effective population size is small (< 50), random changes in allele frequency due to genetic drift potentially are important. If population size changes from one year to the next, the harmonic mean

gives the effective population size. For example, suppose a stable population of 100 sea otters passes through a severe bottleneck and is reduced to a population size of 12 for a single year. Thus, the population sizes are 100, 100, 12, 100, 100, 100, 100, 100, 100, and 100. The arithmetic mean of these numbers is 91.2, but the harmonic mean is only 57.6, an effective population size at which genetic drift could be important. Not only is the harmonic mean less than the arithmetic mean, the harmonic mean is especially sensitive to extreme values that are small. Incidentally, sea otters on the Pacific coast of North America did pass through a severe population bottleneck when they were overhunted in the eighteenth and nineteenth centuries. Although sea otter populations have recovered in size, they still exhibit low genetic diversity, a reflection of this past bottleneck (Larson et al. 2002). (Photograph by Warren Worthington, <http://soundwaves.usgs.gov/2002/07/>)

Other Measures of Location: The Median and the Mode

Ecologists and environmental scientists commonly use two other measures of location, the median and the mode, to summarize datasets. The **median** is defined as the value of a set of ordered observations that has an equal number of observations above and below it. In other words, the median divides a dataset into two halves with equal number of observations in each half. For an odd number of observations, the median is simply the central observation. Thus, if we considered only the first 49 observations in our spine-length data, the median would be the 25th observation (0.248). But with an even number of observations, the median is defined as the midpoint between the $(n/2)$ th and $[(n/2)+1]$ th observation. If we consider all 50 observations in Table 3.1, the median would be the average of the 25th and 26th observations, or 0.2485.

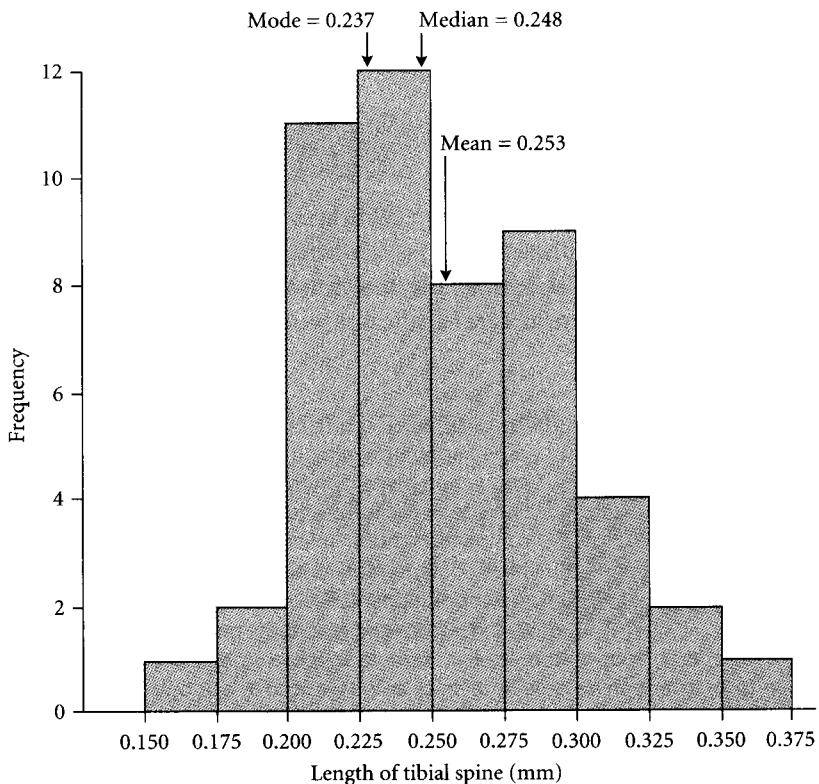


Figure 3.2 Histogram of the tibial spine data from Table 3.1 ($n = 50$) illustrating the arithmetic mean, median, and mode. The mean is the expectation of the data, calculated as the average of the continuous measurements. The median is the midpoint of the ordered set of observations. Half of all observations are larger than the median and half are smaller. The mode is the most frequent observation.

The **mode**, on the other hand, is the value of the observations that occurs most frequently in the sample. The mode can be read easily off of a histogram of the data, as it is the peak of the histogram. Figure 3.2 illustrates the arithmetic mean, the median, and the mode in a histogram of the tibial spine data.

When to Use Each Measure of Location

Why choose one measure of location over another? The arithmetic mean (Equation 3.1) is the most commonly used measure of location, in part because it is familiar. A more important justification is that the Central Limit Theorem (Chapter 2) shows that arithmetic means of large samples of random variables conform to a normal or Gaussian distribution, even if the underlying random variable does not. This property makes it easy to test hypotheses on arithmetic means.

The geometric mean (Equations 3.4 and 3.5) is more appropriate for describing multiplicative processes such as population growth rates or abundance classes of species (before they are logarithmically transformed; see the discussion of the log-normal distribution in Chapter 2 and the discussion of data transformations in Chapter 8). The harmonic mean (Equation 3.6) turns up in calculations used by population geneticists and conservation biologists.

The median or the mode better describe the location of the data when distributions of observations cannot be fit to a standard probability distribution, or when there are extreme observations. This is because the arithmetic, geometric, and harmonic means are very sensitive to extreme (large or small) observations, whereas the median and the mode tend to fall in the middle of the distribution regardless of its spread and shape. In symmetrical distributions such as the normal distribution, the arithmetic mean, median, and mode all are equal. But in asymmetrical distributions, such as that shown in Figure 3.2 (a relatively small random sample from an underlying normal distribution), the mean occurs towards the largest tail of the distribution, the mode occurs in the heaviest part of the distribution, and the median occurs between the two.⁴

⁴People also use different measures of location to support different points of view. For example, the average household income in the United States is considerably higher than the more typical (or median) income. This is because income has a log-normal distribution, so that averages are weighted by the long right-hand tail of the curve, representing the ultrarich. Pay attention to whether the mean, median, or mode of a data set is reported, and be suspicious if it is reported without any measure of spread or variation.

Measures of Spread

It is never sufficient simply to state the mean or other measure of location. Because there is variation in nature, and because there is a limit to the precision with which we can make measurements, we must also quantify and publish the spread, or variability, of our observations.

The Variance and the Standard Deviation

We introduced the concept of variance in Chapter 2. For a random variable Y , the variance $\sigma^2(Y)$ is a measurement of how far the observations of this random variable differ from the expected value. The variance is defined as $E[Y - E(Y)]^2$ where $E(Y)$ is the expected value of Y . As with the mean, the true variance of a population is an unknown quantity. Just as we calculated an estimate \bar{Y} of the population mean μ using our data, we can calculate an estimate s^2 of the population variance σ^2 using our data:

$$s^2 = \frac{1}{n} \sum (Y_i - \bar{Y})^2 \quad (3.7)$$

This value is also referred to as the **mean square**. This term, along with its companion, the **sum of squares**,

$$SS_Y = \sum_{i=1}^n (Y_i - \bar{Y})^2 \quad (3.8)$$

will crop up again when we discuss regression and analysis of variance in Chapters 9 and 10. And just as we defined the standard deviation σ of a random variable as the (positive) square root of its variance, we can estimate it as $s = \sqrt{s^2}$. The square root transformation ensures that the units of standard deviation are the same as the units of the mean.

We noted earlier that the arithmetic mean \bar{Y} provides an unbiased estimate of μ . By unbiased, we mean that if we sampled the population repeatedly (infinitely many times) and computed the arithmetic mean of each sample (regardless of sample size), the grand average of this set of arithmetic means should equal μ . However, our initial estimates of variance and standard deviation are not unbiased estimators of σ^2 and σ , respectively. In particular, Equation 3.7 consistently underestimates the actual variance of the population.

The bias in Equation 3.7 can be illustrated with a simple thought experiment. Suppose you draw a single observation Y_1 from a population and try to estimate μ and $\sigma^2(Y)$. Your estimate of μ is the average of your observations, which in this case is simply Y_1 itself. However, if you estimate $\sigma^2(Y)$ using Equation 3.7, your answer will always equal 0.0 because your lone observation is the same as

your estimate of the mean! The problem is that, with a sample size of 1, we have already used our data to estimate μ , and we effectively have no additional information to estimate $\sigma^2(Y)$.

This leads directly to the concept of **degrees of freedom**. The degrees of freedom represent the number of independent pieces of information that we have in a dataset for estimating statistical parameters. In a dataset of sample size 1, we do not have enough independent observations that can be used to estimate the variance.

The unbiased estimate of the variance, referred to as the **sample variance**, is calculated by dividing the sums of squares by $(n - 1)$ instead of dividing by n . Hence, the unbiased estimate of the variance is

$$s^2 = \frac{1}{n-1} \sum (Y_i - \bar{Y})^2 \quad (3.9)$$

and the unbiased estimate of the standard deviation, referred to as the **sample standard deviation**,⁵ is

$$s = \sqrt{\frac{1}{n-1} \sum (Y_i - \bar{Y})^2} \quad (3.10)$$

Equations 3.9 and 3.10 adjust for the degrees of freedom in the calculation of the sample variance and the standard deviation. These equations also illustrate that you need at least two observations to estimate the variance of a distribution.

For the tibial spine data given in Table 3.1, $s^2 = 0.0017$ and $s = 0.0417$.

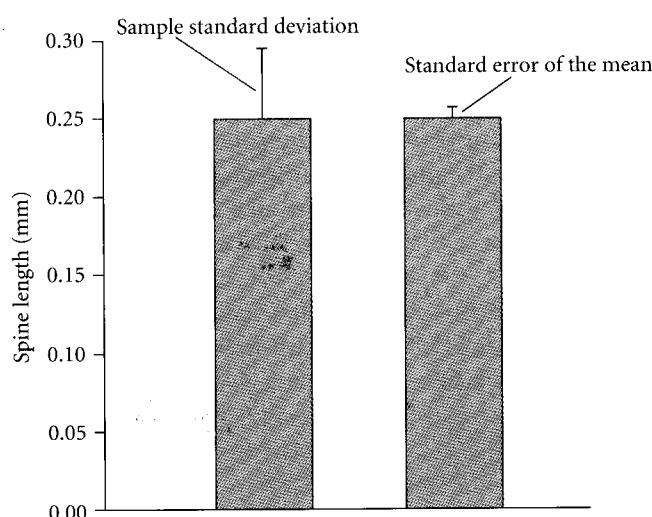
The Standard Error of the Mean

Another measure of spread, used frequently by ecologists and environmental scientists, is the **standard error of the mean**. This measure of spread is abbreviated as $s_{\bar{Y}}$ and is calculated by dividing the sample standard deviation by the square root of the sample size:

$$s_{\bar{Y}} = \frac{s}{\sqrt{n}} \quad (3.11)$$

⁵ This unbiased estimator of the standard deviation is itself unbiased only for relatively large sample sizes ($n > 30$). For smaller sample sizes, Equation 3.10 modestly tends to underestimate the population value of σ (Gurland and Tripathi 1971). Rohlf and Sokal (1995) provide a look-up table of correction factors by which s should be multiplied if $n < 30$. In practice, most biologists do not apply these corrections. As long as the sample sizes of the groups being compared are not greatly different, no serious harm is done by ignoring the correction to s .

Figure 3.3 Bar chart showing the arithmetic mean for the spine data in Table 3.1 ($n = 50$), along with error bars indicating the sample standard deviation (left bar) and standard error of the mean (right bar). Whereas the standard deviation measures the variability of the individual measurements about the mean, the standard error measures the variability of the estimate of the mean itself. The standard error equals the standard deviation divided by \sqrt{n} , so it will always be smaller than the standard deviation, often considerably so. Figure legends and captions should always provide sample sizes and indicate clearly whether the standard deviation or the standard error has been used to construct error bars.



The Law of Large Numbers proves that for an infinitely large number of observations, $\sum Y_i/n$ approximates the population mean μ , where $Y_n = \{Y_i\}$ is a sample of size n from a random variable Y with expected value $E(Y)$. Similarly, the variance of $Y_n = \sigma^2/n$. Because the standard deviation is simply the square root of the variance, the standard deviation of Y_n is

$$\sqrt{\frac{\sigma^2}{n}} = \frac{\sigma}{\sqrt{n}}$$

which is the same as the standard error of the mean. Therefore, the standard error of the mean is an estimate of the standard deviation of the population mean μ .

Unfortunately, some scientists do not understand the distinction between the standard deviation (abbreviated in figure legends as SD) and the standard error of the mean (abbreviated as SE).⁶ Because the standard error of the mean is always smaller than the sample standard deviation, means reported with standard errors appear less variable than those reported with standard deviations (Figure 3.3). However, the decision as to whether to present the sample standard deviation s or the standard error of the mean $s_{\bar{Y}}$ depends on what

⁶ You may have noticed that we referred to the standard error of the mean, and not simply the standard error. The standard error of the mean is equal to the standard deviation of a set of means. Similarly, we could compute the standard deviation of a set of variances or other summary statistics. Although it is uncommon to see other standard errors in ecological and environmental publications, there may be times when you need to report, or at least consider, other standard errors. In Figure 3.1, the standard error of the median = $1.2533 \times s_Y$, and the standard error of the standard deviation = $0.7071 \times s_{\bar{Y}}$. Sokal and Rohlf (1995) provide formulas for standard errors of other common statistics.

inference you want the reader to draw. If your conclusions based on a single sample are representative of the entire population, then report the standard error of the mean. On the other hand, if the conclusions are limited to the sample at hand, it is more honest to report the sample standard deviation. Broad observational surveys covering large spatial scales with large number of samples more likely are representative of the entire population of interest (hence, report $s_{\bar{Y}}$), whereas small, controlled experiments with few replicates more likely are based on a unique (and possibly unrepresentative) group of individuals (hence, report s).

We advocate the reporting of the sample standard deviation, s , which more accurately reflects the underlying variability of the actual data and makes fewer claims to generality. However, as long as you provide the sample size in your text, figure, or figure legend, readers can compute the standard error of the mean from the sample standard deviation and vice versa.

Skewness, Kurtosis, and Central Moments

The standard deviation and the variance are special cases of what statisticians (and physicists) call **central moments**. A central moment (*CM*) is the average of the deviations of all observations in a dataset from the mean of the observations, raised to a power r :

$$CM = \frac{1}{n} \sum_{i=1}^n (Y_i - \bar{Y})^r \quad (3.12)$$

In Equation 3.12, n is the number of observations, Y_i is the value of each individual observation, \bar{Y} is the arithmetic mean of the n observations, and r is a positive integer. The first central moment ($r = 1$) is the sum of the differences of each observation from the sample average (arithmetic mean), which always equals 0. The second central moment ($r = 2$) is the variance (Equation 3.5).

The third central moment ($r = 3$) divided by the standard deviation cubed (s^3) is called the **skewness** (denoted as g_1):

$$g_1 = \frac{1}{ns^3} \sum_{i=1}^n (Y_i - \bar{Y})^3 \quad (3.13)$$

Skewness describes how the sample differs in shape from a symmetrical distribution. A normal distribution has $g_1 = 0$. A distribution for which $g_1 > 0$ is **right-skewed**: there is a long tail of observations greater than (i.e., to the right of) the mean. In contrast, $g_1 < 0$, is **left-skewed**: there is a long tail of observations less than (i.e., to the left of) the mean (Figure 3.4).

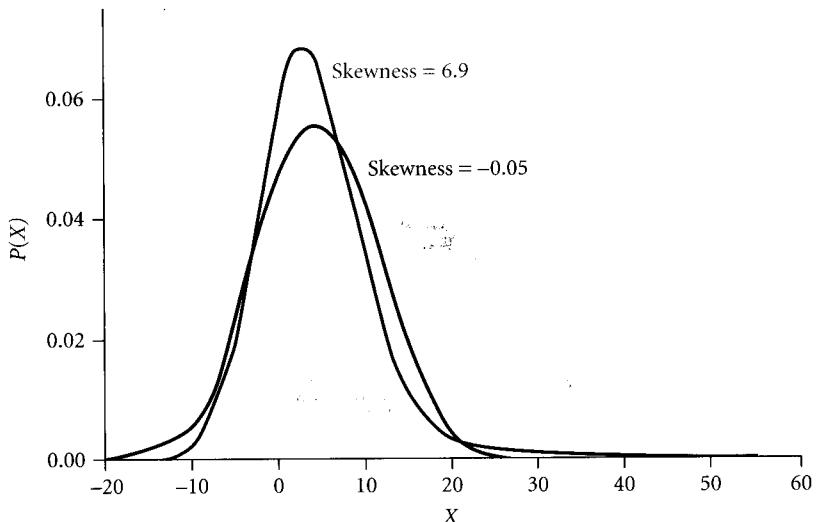


Figure 3.4 Continuous distributions illustrating skewness (g_1). Skewness measures the extent to which a distribution is asymmetric, with either a long right- or left-hand probability tail. The green curve is the log-normal distribution illustrated in Figure 2.8; it has positive skewness, with many more observations to the right of the mean than to the left (a long right tail), and a skewness measure of 6.9. The black curve represents a sample of 1000 observations from a normal random variable with identical mean and standard deviation as the log-normal distribution. Because these data were drawn from a symmetric normal distribution, they have approximately the same number of observations on either side of the mean, and the measured skewness is approximately 0.

The **kurtosis** is based on the fourth central moment ($r = 4$):

$$g_2 = \left[\frac{1}{ns^4} \sum_{i=1}^n (Y_i - \bar{Y})^4 \right] - 3 \quad (3.14)$$

Kurtosis measures the extent to which a probability density is distributed in the tails versus the center of the distribution. Clumped or **platykurtic** distributions have $g_2 < 0$; compared to a normal distribution, there is more probability mass in the center of the distribution, and less probability in the tails. In contrast, **leptokurtic** distributions have $g_2 > 0$. Leptokurtic distributions have less probability mass in the center, and relatively fat probability tails (Figure 3.5).

Although skewness and kurtosis were often reported in the ecological literature prior to the mid-1980s, it is uncommon to see these values reported now. Their statistical properties are not good: they are very sensitive to outliers, and

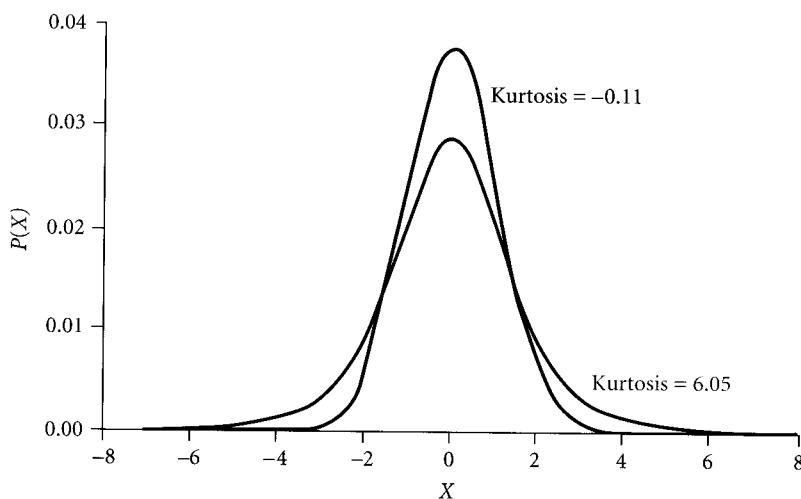


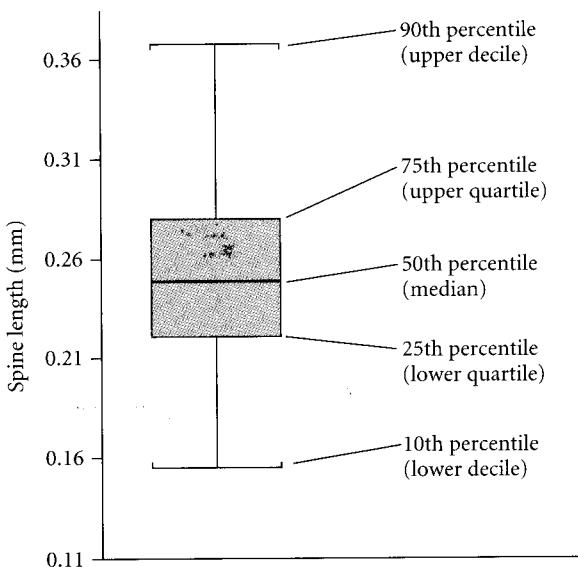
Figure 3.5 Distributions illustrating kurtosis (g_2). Kurtosis measures the extent to which the distribution is fat-tailed or thin-tailed compared to a standard normal distribution. Fat-tailed distributions are leptokurtic, and contain relatively more area in the tails of the distribution and less in the center. Leptokurtic distributions have positive values for g_2 . Thin-tailed distributions are platykurtic, and contain relatively less area in the tails of the distribution and more in the center. Platykurtic distributions have negative values for g_2 . The black curve represents a sample of 1000 observations from a normal random variable with mean = 0 and standard deviation = 1 ($X \sim N(0,1)$); its kurtosis is nearly 0. The green curve is a sample of 1000 observations from a t distribution with 3 degrees of freedom. The t distribution is leptokurtic and has a positive kurtosis ($g_2 = 6.05$ in this example).

to differences in the mean of the distribution. Weiner and Solbrig (1984) discuss the problem of using measures of skewness in ecological studies.

Quantiles

Another way to illustrate the spread of a distribution is to report its **quantiles**. We are all familiar with one kind of quantile, the **percentile**, because of its use in standardized testing. When a test score is reported as being in the 90th percentile, 90% of the scores are lower than the one being reported, and 10% are above it. Earlier in this chapter we saw another example of a percentile—the median, which is the value located at the 50th percentile of the data. In presentations of statistical data, we most commonly report upper and lower **quartiles**—the values for the 25th and 75th percentiles, and upper and lower **deciles**—the values for the 10th and 90th percentiles. These values for the spine data are illustrated concisely in a **box plot** (Figure 3.6). Unlike the variance and standard

Figure 3.6 Box plot illustrating quantiles of data from Table 3.1 ($n = 50$). The line indicates the 50th percentile (median), and the box encompasses 50% of the data, from the 25th to the 75th percentile. The vertical lines extend from the 10th to the 90th percentile.



deviation, the values of the quantiles do not depend on the values of the arithmetic mean or median. When distributions are asymmetric or contain **outliers** (extreme data points that are not characteristic of the distribution they were sampled from; see Chapter 8), box plots of quantiles can portray the distribution of the data more accurately than conventional plots of means and standard deviations.

Using Measures of Spread

By themselves, measures of spread are not especially informative. Their primary utility is for comparing data from different populations or from different treatments within experiments. For example, analysis of variance (Chapter 10) uses the values of the sample variances to test hypotheses that experimental treatments differ from one another. The familiar *t*-test uses sample standard deviations to test hypotheses that the means of two populations differ from each other. It is not straightforward to compare variability itself across populations or treatment groups because the variance and standard deviation depend on the sample mean. However, by discounting the standard deviation by the mean, we can calculate an independent measure of variability, called the **coefficient of variation**, or *CV*.

The *CV* is simply the sample standard deviation divided by the mean, s/\bar{Y} , and is conventionally multiplied by 100 to give a percentage. The *CV* for our spine data = 16.5%. If another population of spiders had a *CV* of tibial spine

length = 25%, we would say that our first population is somewhat less variable than the second population.

A related index is the **coefficient of dispersion**, which is calculated as the sample variance divided by the mean (s^2/\bar{Y}). The coefficient of dispersion can be used with discrete data to assess whether individuals are clumped or hyperdispersed in space, or whether they are spaced at random as predicted by a Poisson distribution. Biological forces that violate independence will cause observed distributions to differ from those predicted by the Poisson.

For example, some marine invertebrate larvae exhibit an aggregated settling response: once a juvenile occupies a patch, that patch becomes very attractive as a settlement surface for subsequent larvae (Crisp 1979). Compared to a Poisson distribution, these aggregated or clumped distributions will tend to have too many samples with high numbers of occurrence, *and* too many samples with 0 occurrences. In contrast, many ant colonies exhibit strong territoriality and kill or drive off other ants that try to establish colonies within the territory (Levings and Traniello 1981). This segregative behavior also will push the distribution away from the Poisson. In this case the colonies will be hyperdispersed: there will be too few samples in the 0 frequency class *and* too few samples with high numbers of occurrence.

Because the variance and mean of a Poisson random variable both equal λ , the coefficient of dispersion (*CD*) for a Poisson random variable = $\lambda/\lambda = 1$. On the other hand, if the data are clumped or aggregated, *CD* > 1.0, and if the data are hyperdispersed or segregated, *CD* < 1.0. However, the analysis of spatial pattern with Poisson distributions can become complicated because the results depend not only on the degree of clumping or segregation of the organisms, but also on the size, number, and placement of the sampling units. Hurlbert (1990) discusses some of the issues involved in fitting spatial data to a Poisson distribution.

Some Philosophical Issues Surrounding Summary Statistics

The fundamental summary statistics—the sample mean, standard deviation, and variance—are estimates of the actual population-level **parameters**, μ , σ , and σ^2 that we obtain directly from our data. Because we can never sample the entire population, we are forced to estimate these unknown parameters by \bar{Y} , s , and s^2 . In doing so, we make a fundamental assumption: that there is a true fixed value for each of these parameters. The Law of Large Numbers proves that if we sampled our population infinitely many times, the average of the infinitely many \bar{Y} 's that we calculated from our infinitely many samples would equal μ .

The Law of Large Numbers forms the foundation for what has come to be known as **parametric**, **frequentist**, or **asymptotic statistics**. Parametric statis-

tics are so called because the assumption is that the measured variable can be described by a random variable or probability distribution of known form with defined, fixed parameters. Frequentist or asymptotic statistics are so called because we assume that if the experiment were repeated infinitely many times, the most frequent estimates of the parameters would converge on (reach an asymptote at) their true values.

But what if this fundamental assumption—that the underlying parameters have true, fixed values—is false? For example, if our samples were taken over a long period of time, there might be changes in spine length of spider tibias because of phenotypic plasticity in growth, or even evolutionary change due to natural selection. Or, perhaps our samples were taken over a short period of time, but each spider came from a distinct microhabitat, for which there was a unique expectation and variance of spider tibia length. In such a case, is there any real meaning to an estimate of a single value for the average length of a tibial spine in the spider population? Bayesian statistics begin with the fundamental assumption that population-level parameters such as μ , σ , and σ^2 are themselves random variables. A Bayesian analysis produces estimates not only of the values of the parameters but also of the inherent variability in these parameters.

The distinction between the frequentist and Bayesian approaches is far from trivial, and has resulted in many years of acrimonious debate, first among statisticians, and more recently among ecologists. As we will see in Chapter 5, Bayesian estimates of parameters as random variables often require complex computer calculations. In contrast, frequentist estimates of parameters as fixed values use simple formulas that we have outlined in this chapter. Because of the computational complexity of the Bayesian estimates, it was initially unclear whether the results of frequentist and Bayesian analyses would be quantitatively different. However, with fast computers, we are now able to carry out complex Bayesian analyses. Under certain conditions, the results of the two types of analyses are quantitatively similar. The decision of which type of analysis to use, therefore, should be based more on a philosophical standpoint than on a quantitative outcome (Ellison 2004). However, the interpretation of statistical results may be quite different from the Bayesian and frequentist perspectives. An example of such a difference is the construction and interpretation of confidence intervals for parameter estimates.

Confidence Intervals

Scientists often use the sample standard deviation to construct a **confidence interval** around the mean (Figure 3.1). For a normally distributed random vari-

able, approximately 67% of the observations occur within ± 1 standard deviation of the mean, and approximately 96% of the observations occur within ± 2 standard deviations of the mean.⁷ We use this observation to create a 95% confidence interval, which for large samples is the interval bounded by $(\bar{Y} - 1.96s_{\bar{Y}}, \bar{Y} + 1.96s_{\bar{Y}})$. What does this interval represent? It means that the probability that the true population mean μ falls within the confidence interval = 0.95:

$$P(\bar{Y} - 1.96s_{\bar{Y}} \leq \mu \leq \bar{Y} + 1.96s_{\bar{Y}}) = 0.95 \quad (3.15)$$

Because our sample mean and sample standard error of the mean are derived from a single sample, this confidence interval will change if we sample the population again (although if our sampling is random and unbiased, it should not change by very much). Thus, this expression is asserting that the probability that the true population mean μ falls within a single calculated confidence interval = 0.95. By extension, if we were to repeatedly sample the population (keeping the sample size constant), 5% of the time we would expect that the true population mean μ would lie outside of this confidence interval.

Interpreting a confidence interval is tricky. A common misinterpretation of the confidence interval is “There is a 95% chance that the true population mean μ occurs within this interval.” Wrong. The confidence interval either does or does not contain μ ; unlike Schrödinger’s quantum cat (see Footnote 9, Chapter 1), μ cannot be both in and out of the confidence interval simultaneously. What you can say is that, 95% of the time, an interval calculated in this way will contain the fixed value of μ . Thus, if you carried out your sampling experiment 100 times, and created 100 such confidence intervals, approximately 95 of them would contain μ and 5 would not (see Figure 3.1 for an example of when a 95% confidence interval does not include the true population mean μ). Blume and Royall (2003) provide further examples and a more detailed pedagogical description.

⁷ Use the “two standard deviation rule” when you read the scientific literature, and get into the habit of quickly estimating rough confidence intervals for sample data. For example, suppose you read in a paper that average nitrogen content of a sample of plant tissues was $3.4\% \pm 0.2$, where 0.2 is the sample standard deviation. Two standard deviations = 0.4, which is then added to and subtracted from the mean. Therefore, approximately 95% of the observations were between 3.0% and 3.8%. You can use this same trick when you examine bar graphs in which the standard deviation is plotted as an error bar. This is an excellent way to use summary statistics to spot check reported statistical differences among groups.

This rather convoluted explanation is not satisfying, and it is not exactly what you would like to assert when you construct a confidence interval. Intuitively, you would like to be saying how confident you are that the mean is inside of your interval (i.e., you're 95% sure that the mean is in the interval). A frequentist statistician, however, can't assert that. If there is a fixed population mean μ , then it's either inside the interval or not, and the probability statement (Equation 3.15) asserts how probable it is that this particular confidence interval includes μ . On the other hand, a Bayesian statistician turns this around. Because the confidence interval is fixed (by your sample data), a Bayesian can calculate the probability that the population mean (itself a random variable) occurs within the confidence interval. Bayesian statisticians refer to these intervals as **credibility intervals**, in order to distinguish them from frequentist confidence intervals. See Chapter 5 and Ellison (1996) for further details.

Generalized Confidence Intervals

We can, of course, construct any percentile confidence interval, such as a 90% confidence interval or a 50% confidence interval. The general formula for an $n\%$ confidence interval is

$$P(\bar{Y} - t_{\alpha[n-1]} s_{\bar{Y}} \leq \mu \leq \bar{Y} + t_{\alpha[n-1]} s_{\bar{Y}}) = (1 - \alpha) \quad (3.16)$$

where $t_{\alpha[n-1]}$ is the critical value of a ***t*-distribution** with probability $P = \alpha$, and sample size n . This probability expresses the percentage of the area under the two tails of the curve of a *t*-distribution (Figure 3.7). For a standard normal curve (a *t*-distribution with $n = \infty$), 95% of the area under the curve lies within ± 1.96 standard deviations of the mean. Thus 5% of the area ($P = 0.05$) under the curve remains in the two tails beyond the points ± 1.96 .

So what is this *t*-distribution? Recall from Chapter 2 that an arithmetic transformation of a normal random variable is itself a normal random variable. Consider the set of sample means $\{\bar{Y}_k\}$ resulting from a set of replicate groups of measurements of a normal random variable with unknown mean μ . The deviation of the sample means from the population mean $\bar{Y}_k - \mu$ is also a normal random variable. If this latter random variable $(\bar{Y}_k - \mu)$ is divided by the unknown population standard deviation σ , the result is a standard normal random variable (mean = 0, standard deviation = 1). However, we don't know the population standard deviation σ , and must instead divide the deviations of each

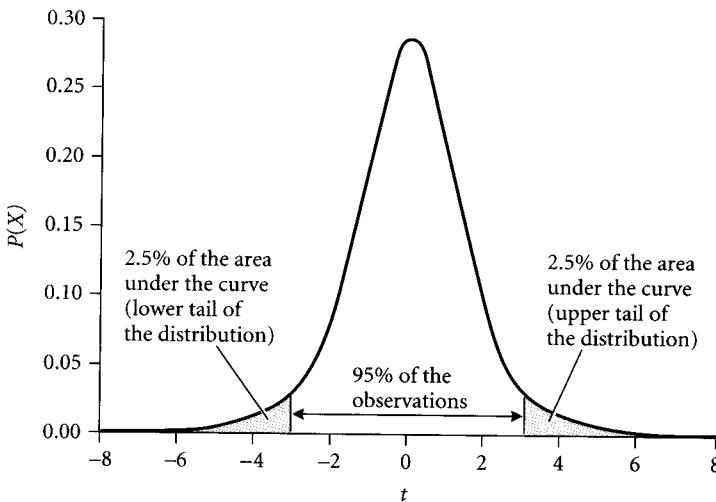


Figure 3.7 t -distribution illustrating that 95% of the observations, or probability mass lies within ± 1.96 standard deviations of the mean (mean = 0) percentiles. The two tails of the distribution each contain 2.5% of the observations or probability mass of the distribution. Their sum is 5% of the observations, and the probability $P = 0.05$ that an observation falls within these tails. This distribution is identical to the t -distribution illustrated in Figure 3.5.

mean by the estimate of the standard error of the mean of each sample ($s_{\bar{Y}_k}$). The resulting t -distribution is similar, but not identical, to a standard normal distribution. This t -distribution is leptokurtic, with longer and heavier tails than a standard normal distribution.⁸

⁸ This result was first demonstrated by the statistician W. S. Gossett, who published it using the pseudonym "Student." Gossett at the time was employed at the Guinness Brewery, which did not allow its employees to publish trade secrets; hence the need for a pseudonym. This modified standard normal distribution, which Gossett called a t -distribution, is also known as the Student's distribution, or the Student's t -distribution. As the number of samples increases, the t -distribution approaches the standard normal distribution in shape. The construction of the t -distribution requires the specification of the sample size n and is normally written as $t_{[n]}$. For $n = \infty$, $t_{[\infty]} \sim N(0,1)$.

Because a t -distribution for small n is leptokurtic (see Figure 3.5), the width of a confidence interval constructed from it will shrink as sample size increases. For example, for $n = 10$, 95% of the area under the curve falls between ± 2.228 . For $n = 100$, 95% of the area under the curve falls between ± 1.990 . The resulting confidence interval is 12% wider for $n = 10$ than for $n = 100$.

Summary

Summary statistics describe expected values and variability of a random sample of data. Measures of location include the median, mode, and several means. If the samples are random or independent, the arithmetic mean is an unbiased estimator of the expected value of a normal distribution. The geometric mean and harmonic mean are also used in special circumstances. Measures of spread include the variance, standard deviation, standard error, and quantiles. These measures describe the variation of the observations around the expected value. Measures of spread can also be used to construct confidence or credibility intervals. The interpretations of these intervals differ between frequentist and Bayesian statisticians.