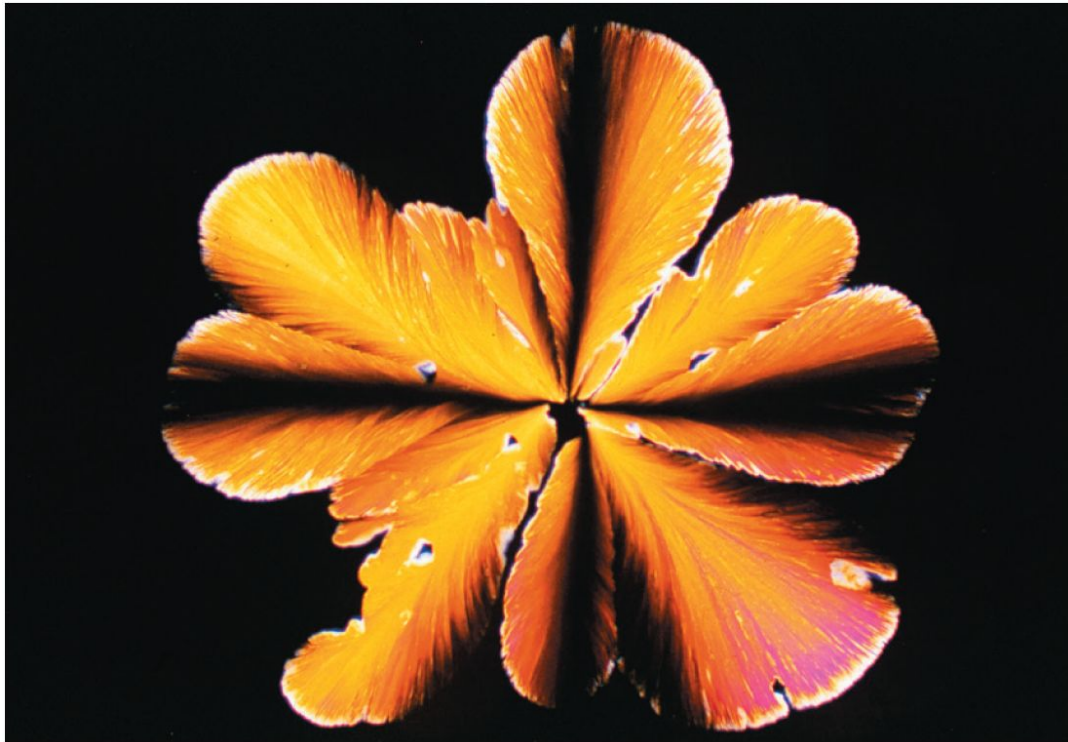


## Chapter 4 Estimating with uncertainty

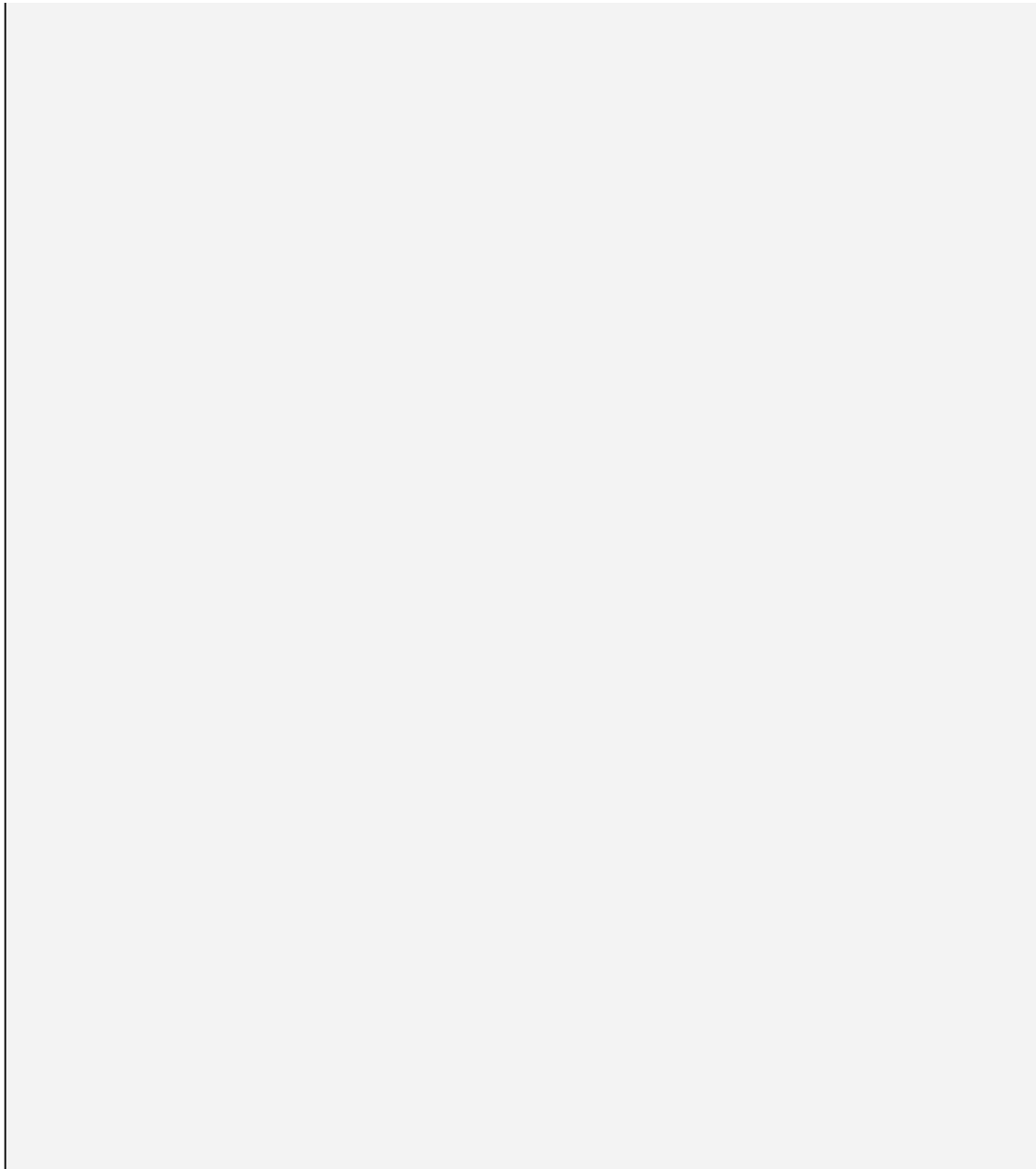


© Molecular Expressions

DNA crystal

### Description

-



When biologists carry out a study, their goals are usually more ambitious than mere description of the resulting data. Rather, data are gathered so that something may be discovered about the larger population from which the sample came. The descriptive statistics measured on a sample are used to estimate parameters of the population. Such estimation is possible when the sample is a random sample. Recall from [Chapter 1](#) that in a random

sample, all individuals in the population have an equal chance of being selected and individuals are sampled independently.

For example, the sample mean  $\bar{Y}$  is used to estimate the true mean of the population, symbolized using the Greek letter  $\mu$  (mu; pronounced “mew”). The sample standard deviation  $s$  is used to estimate the population standard deviation, symbolized by  $\sigma$  (lowercase Greek letter sigma). Likewise, the sample proportion  $\hat{p}$  (pronounced “p-hat”) is an estimate<sup>1</sup> of the population proportion  $p$ .

For an estimate of a population parameter to be useful, we also need to quantify its precision. We need a measure of how far the estimate is likely to be from the target parameter being estimated. If precision is high, then our uncertainty is low. We can be reasonably confident that our estimate is close to the truth. If instead precision is low, then our uncertainty is high, and we’ll need more data to reduce it.

In [Chapter 4](#), we explain the basics of estimation and the uncertainties involved when making generalizations about a population from a random sample. We introduce the standard error, a key measure of the precision of an estimate, and we demonstrate it in the case of the sample mean. We add a brief, conceptual introduction to the confidence interval, another important means of describing the precision of an estimate.

## 4.1 The sampling distribution of an estimate

**Estimation** is the process of inferring a population parameter from sample data. The value of an estimate calculated from data is almost never exactly the same as the value of the population parameter being estimated, because sampling is influenced by chance. The crucial question is “In the face of chance, how much can we trust an estimate?” In other words, what is its *precision*? To answer this question, we need to know something about how the sampling process might affect the estimates we get. We use the **sampling distribution** of the estimate, which is the probability distribution of all the values for an estimate that we *might* have obtained when we sampled the population. We illustrate the concept of a sampling distribution using samples from a known population, the genes of the human genome.

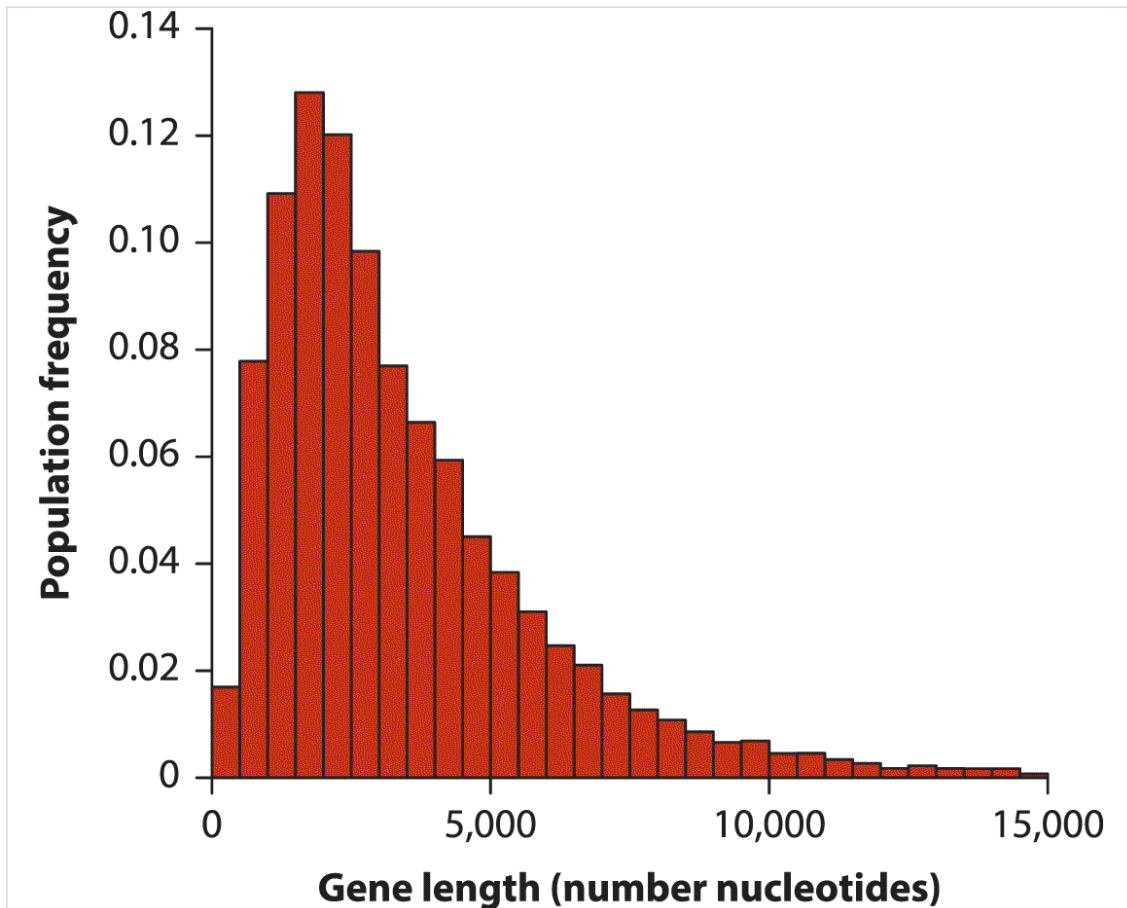
---

### EXAMPLE 4.1: The length of human genes

The international Human Genome Project was the largest coordinated research effort in the history of biology. It yielded the DNA sequence of all 23 human chromosomes, each consisting of millions of nucleotides chained end to end.<sup>2</sup> These encode the genes whose products—RNA and proteins—shape the growth and development of each individual.

We obtained the lengths of all 20,290 known and predicted genes of the published genome sequence ([Hubbard et al. 2005](#)).<sup>3</sup> The length of a gene refers to the total number of nucleotides comprising the coding regions. The frequency distribution of gene lengths in the population of genes is

shown in [Figure 4.1-1](#). The figure includes only genes up to 15,000 nucleotides long; in addition, there are 26 longer genes.<sup>4</sup>



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

#### FIGURE 4.1-1

Distribution of gene lengths in the known human genome. The graph is truncated at 15,000 nucleotides; 26 larger genes are too rare to be visible in this histogram.

#### Description

The horizontal axis represents gene length in number of nucleotides from 0 to 15000 with an interval of 5000. The vertical axis represents population frequency from 0 to 0 point 1 4 with an interval of 0 point 0 2. The approximate data from graph are as follows: The top edges of the bars show a stepwise increase first till 2000. Then, it shows a curve in decreasing manner reaching almost 0 at 15,000. For 0 to 500 nucleotides, population frequency

is 0 point 1 9. For 500 to 1000, it is 0 point 0 7 9. For 1000 to 1500, it is 0 point 1 0 9. For 1500 to 2000, it is 0 point 1 3. The bars then decrease gradually and reaches almost 0 at 15,000.

---

The histogram in [Figure 4.1-1](#) is like those we have seen before, except that it shows the distribution of lengths in the *population* of genes, not simply those in a *sample* of genes. Because it is the population distribution, the relative frequency of genes of a given length interval in [Figure 4.1-1](#) represents the *probability* of obtaining a gene of that length

when sampling a single gene at random. The probability distribution of gene lengths is positively skewed, having a long tail extending to the right.

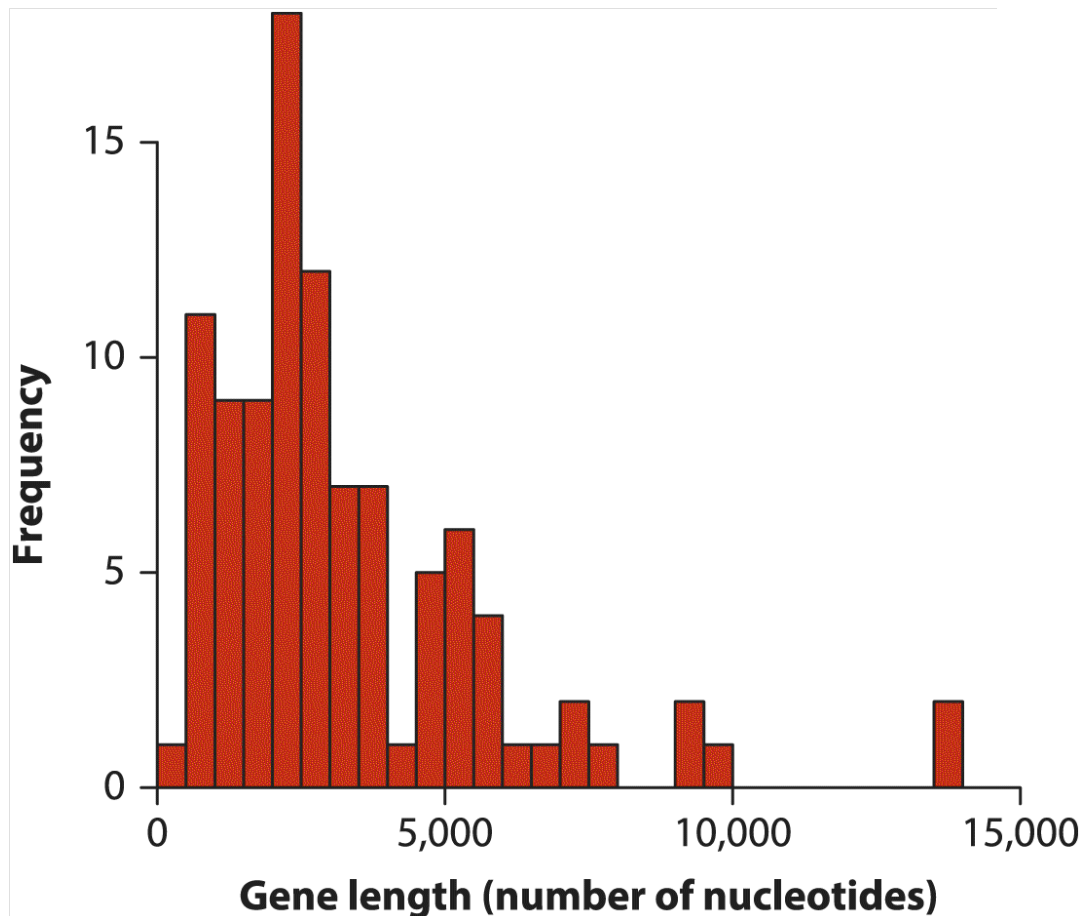
The population mean and standard deviation of gene length in the human genome are listed in [Table 4.1-1](#). These quantities are referred to as *parameters* because they are quantities that describe the population.

TABLE 4.1-1 Population mean and standard deviation of gene length in the known human genome.		
Name	Parameter	Value (nucleotides)
Mean	$\mu$	3511.5
Standard deviation	$\sigma$	2833.2

In real life, we would not usually know the parameter values of the study population, but in this case we do. We'll take advantage of this to illustrate the process of sampling.

## Estimating mean gene length with a random sample

To begin, we collected a single random sample of  $n=100$  genes from the known human genome.<sup>5</sup> A histogram of the lengths of the resulting sample of genes is shown in [Figure 4.1-2](#).



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

#### FIGURE 4.1-2

Frequency distribution of gene lengths in a unique random sample of  $n=100$  genes from the human genome.

#### Description

The horizontal axis represents gene length in number of nucleotides from 0 to 15,000 with an interval of 5,000. The vertical axis represents frequency from 0 to 15 with an interval of 5. The approximate data from graph are as follows: For gene length 0 to 500, frequency is 1. For gene length 500 to 1000, frequency is 11. For gene length 1000 to 1500, frequency is 9. For gene length 1500 to 2000, frequency is 9. For gene length 2000 to 2500, frequency is 19. For gene length 2500 to 3000, frequency is 12. For gene length 3000 to 3500, frequency is 7. For gene length 3500 to 4000, frequency is 7. For gene length 4000 to 4500, frequency is 1. For gene length 4500 to 5000, frequency



is 5. For gene length 5000 to 5500, frequency is 6. For gene length 5500 to 6000, frequency is 4. For gene length 6000 to 6500, frequency is 1. For gene length 6500 to 7000, frequency is 1. For gene length 7000 to 7500, frequency is 2. For gene length 7500 to 8000, frequency is 1. For gene length 8000 to 9000, frequency is 0. For gene length 9000 to 9500, frequency is 2. For gene length 9500 to 10000, frequency is 1. For gene length 10000 to 13500, frequency is 0. For gene length 13500 to 14000, frequency is 2. For gene length 14000 to 15000, frequency is 0.

The frequency distribution of the random sample ([Figure 4.1-2](#)) is not an exact replica of the population distribution ([Figure 4.1-1](#)), because of chance. The two distributions nevertheless share important features, including approximate location, spread, and shape. For example, the sample frequency distribution is skewed to the right like the true population distribution.

The sample mean and standard deviation of gene length from the sample of 100 genes are listed in [Table 4.1-2](#). How close are these estimates to the

population mean and standard deviation listed in [Table 4.1-1](#)? The sample mean is  $\bar{Y}=3328.1$ ,  $\bar{Y} = 3328.1$ , which is 183 nucleotides shorter than the true value, the population mean of  $\mu=3511.5$ ,  $\mu = 3511.5$ . The sample standard deviation  $s=2521.6$ ,  $s = 2521.6$  is also different from the standard deviation of gene length in the population  $\sigma=2833.2$ ,  $\sigma = 2833.2$ . We shouldn't be surprised that the sample estimates differ from the parameter (population) values. Such differences are virtually inevitable because of chance in the random sampling process.

**TABLE 4.1-2 Mean and standard deviation of gene length  $\bar{Y}$  in our unique random sample of  $n=100$  genes from the human genome.**

Name	Statistic	Sample value (number of nucleotides)
Mean	$\bar{Y}$	3328.1
Standard deviation	$s$	2521.6

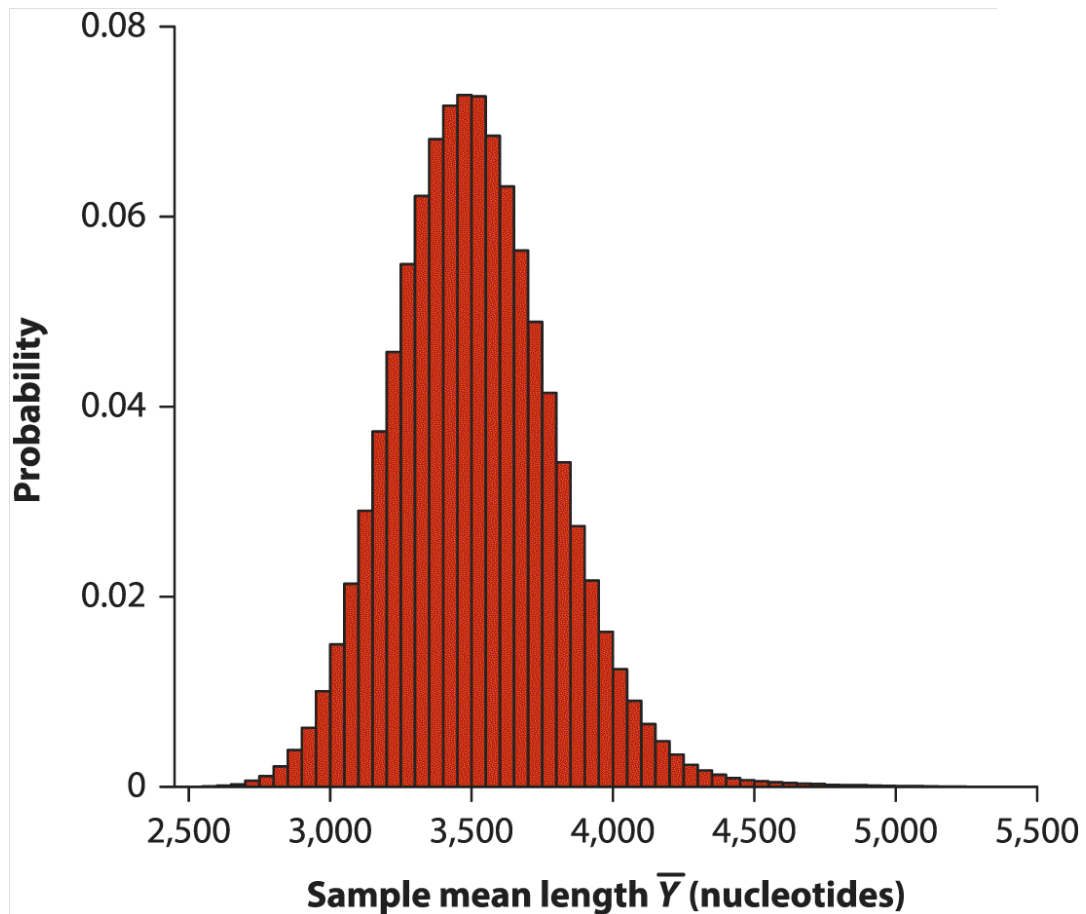
## The sampling distribution of $\bar{Y}$

We obtained  $\bar{Y}=3328.1$  nucleotides in our single sample, but by chance we might have obtained a different value. For example, when we took a second random sample of 100 genes, we found  $\bar{Y}=3668.8$ . Each new sample will usually generate a different estimate of the same parameter. If we were able to repeat this sampling an infinite number of times, we could create the probability distribution of our estimate. The probability distribution of values we might obtain for an estimate makes up the estimate's [sampling distribution](#).

The *sampling distribution* is the probability distribution of all values for an estimate that we might obtain when we sample a population.

The sampling distribution represents the “population” of values for an estimate. It is not a real population, like the squirrels in Muir Woods or all the retirees basking in the Florida sunshine. Rather, the sampling distribution is an imaginary population of values for an estimate. Taking a random sample of  $n$  observations from a population and calculating  $\bar{Y}$  is equivalent to randomly sampling a *single* value of  $\bar{Y}$  from its sampling distribution.

To visualize the sampling distribution for mean gene length, we used the computer to take a vast number of random samples of  $n=100$  genes from the human genome. We calculated the sample mean  $\bar{Y}$  each time. The resulting histogram in [Figure 4.1-3](#) shows the values of  $\bar{Y}$  that might be obtained when randomly sampling 100 genes, together with their probabilities.



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

### FIGURE 4.1-3

The sampling distribution of mean gene length  $\bar{Y}$  when  $n=100$ .  
 $n = 100$ . Note the change in scale from [Figure 4.1-2](#).

### Description

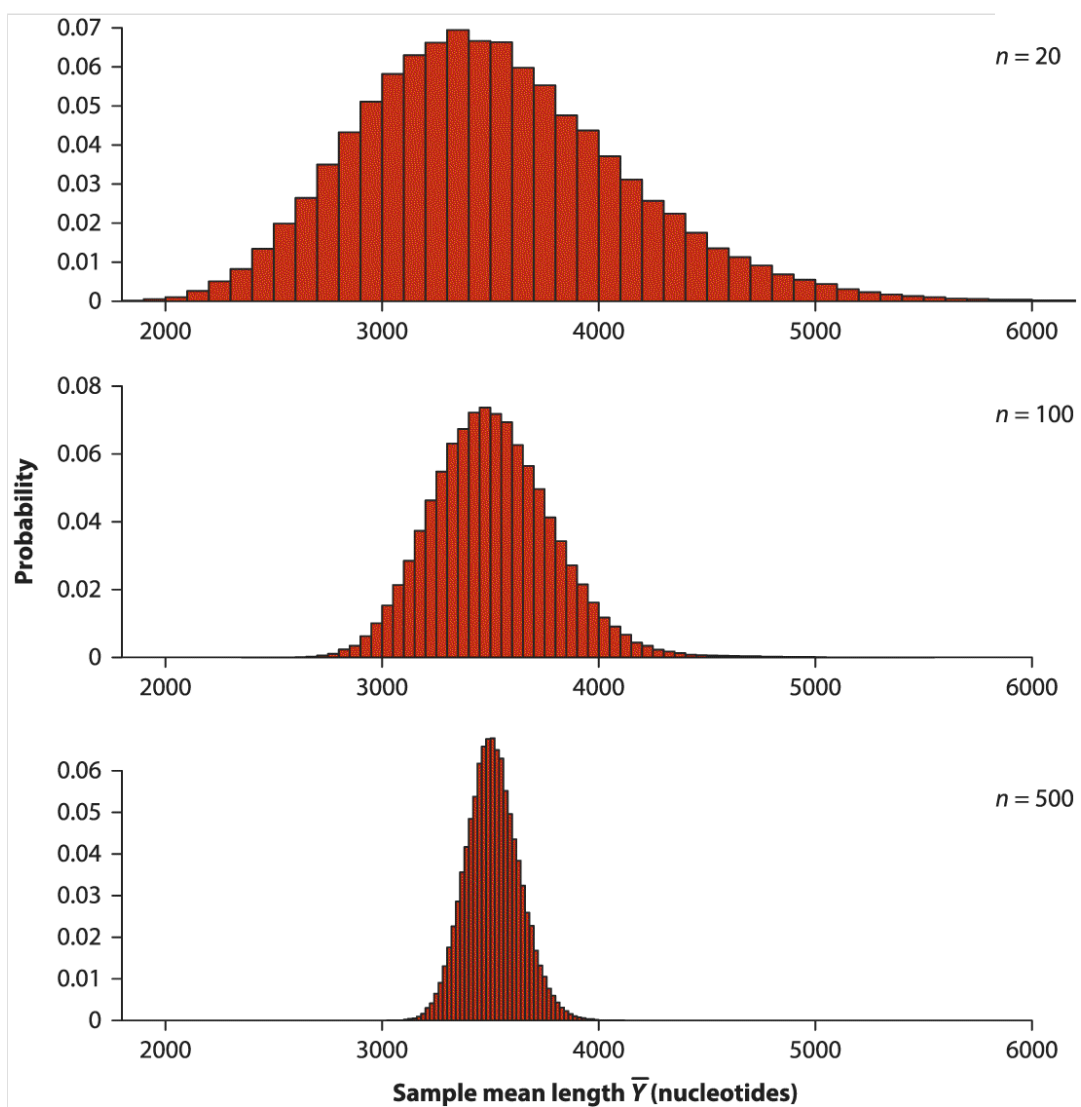
The horizontal axis of the graph is sample mean length  $\bar{Y}$ -bar (in nucleotides) ranging from 2500 to 5500 with an interval of 500. The vertical axis of the graph is probability ranging from 0 to 0 point 0 8 with an interval of 0 point 0 2. The approximate data from the graph are as follows:

Most of the rectangular bars are drawn between 2800 and 4500 in such a way that the edges of the bars can be connected by a concave downward curve. The curve starts at almost 0 probability for 2,700 to 2,750 sample mean

length, shows peak at 0.72 probability for 3,450 to 3,500 sample mean length, and ends at almost 0 probability for 4,800 to 4,850 sample mean length.

[Figure 4.1-3](#) makes plain that although the population mean  $\mu$  is a constant (3511.5), its estimate  $\bar{Y}$  is a variable. Each new sample yields a different  $\bar{Y}$  value from the one before. We don't ever see the sampling distribution of  $\bar{Y}$  because ordinarily we have only one sample, and therefore only one  $\bar{Y}$ . Notice that the sampling distribution for  $\bar{Y}$  is centered exactly on the true mean,  $\mu$ . This means that  $\bar{Y}$  is an unbiased estimate of  $\mu$ . [6](#)

The spread of the sampling distribution of an estimate depends on the sample size. The sampling distribution of  $\bar{Y}$  based on  $n=100$  is narrower than that based on  $n=20$ , and that based on  $n=500$  is narrower still (Figure 4.1-4). The larger the sample size, the narrower the sampling distribution. And the narrower the sampling distribution, the more precise the estimate. Thus, larger samples are desirable whenever possible because they yield more precise estimates. The same is true for the sampling distributions of estimates of other population quantities, not just  $\bar{Y}$ .



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

#### FIGURE 4.1-4

Comparison of the sampling distributions of mean gene length  $\bar{Y}$  when  $n=20$ ,  $n=100$ , and  $n=500$ .

##### Description

The horizontal axis for each graph shows sample mean length  $\bar{Y}$  (in nucleotides) ranging from 2000 to 6000 with an interval of 1000. The vertical axis of each graph represents probability, but with different ranges.

In the first graph, the vertical axis ranges from 0 to 0.7 with an interval of 0.1. Most of the rectangular bars are present between 2,000 and 5,700 in such a way that the edges of the bars can be connected by a concave downward curve. The approximate data from the graph are as follows: The curve starts at almost 0 probability for 2,000 to 2,100 sample mean length, shows peak at 0.65 probability for 3,300 to 3,400 sample mean length, and ends at almost 0 probability from 5,200 to 5,700 sample mean length.

In the second graph, the vertical axis ranges from 0 to 0.8 with an interval of 0.2 and most of the rectangular bars are present between 2700 and 4400 in such a way that the edges of the bars can be connected by a concave downward curve. The approximate data from the graph are as follows: The curve starts at almost 0 probability for 2,700 to 2,800 sample mean length, shows peak at 0.7 probability for 3,500 to 3,570 sample mean length, and ends at almost 0 probability for 4,250 to 4,400 sample mean length.

In the third graph, the vertical axis ranges from 0 to 0.6 with an interval of 0.1 and most of the rectangular bars are present between 3200 and 3800 in such a way that the edges of the bars can be connected by a concave downward curve. The approximate data from the graph are as follows: The curve starts at almost 0 probability for 3,200 to 3,250 sample mean length, shows peak at 0.7 probability for 3,500 to 3,540 sample

mean length, and ends at almost 0 probability for 3,700 to 3,750 sample mean length.

Increasing sample size reduces the spread of the sampling distribution of an estimate, increasing precision.



## 4.2 Measuring the uncertainty of an estimate

In this section, we show how the sampling distribution is used to measure the uncertainty of an estimate.

### Standard error

The standard deviation of the sampling distribution of an estimate is called the standard error. Because it reflects the differences between an estimate and the target parameter, the standard error reflects the precision of an estimate. Estimates with smaller standard errors are more precise than those with larger standard errors. The smaller the standard error, the less uncertainty there is about the target parameter in the population.

The *standard error* of an estimate is the standard deviation of the estimate's sampling distribution.

### The standard error of $\bar{Y}$

The standard error of the sample mean is particularly simple to calculate, so we show it here. We can represent the standard error of the mean with the symbol  $\sigma_{\bar{Y}}$ . It has a remarkably straightforward relationship with  $\sigma$ , the population standard deviation of the variable  $Y$ :

$$\sigma_{\bar{Y}} = \frac{\sigma}{\sqrt{n}}.$$

The standard error decreases with increasing sample size. [Table 4.2-1](#) lists the standard error of the sample mean based on random samples of  $n=20$ , 100, and 500 from the known human genome.

**TABLE 4.2-1 Standard error of the sampling distributions of mean gene length  $\bar{Y}$  according to sample size. These measure the spread of the three sampling distributions in [Figure 4.1-4](#).**

Sample size, $n$	Standard error, $\sigma_{\bar{Y}}$ (nucleotides)
20	633.5
100	281.6
500	125.0

Note that as the sample size increases in [Table 4.2-1](#), the standard error of  $\bar{Y}$  gets smaller. Compare this to [Figure 4.1-4](#), where the estimates based on smaller sample sizes are more likely to be further from the true value of the population mean, on average. The smaller the standard error, the more likely it is that the estimate is close to the true value of the parameter.

# The standard error of $\bar{Y}$ from data

The trouble with the formula for the standard error of the mean  $(\sigma_{\bar{Y}})$  is that we almost never know the value of the population standard deviation  $(\sigma)$ , and so we cannot calculate  $\sigma_{\bar{Y}}$ . The next best thing is to approximate the standard error of the mean by using the sample standard deviation ( $s$ ) as an estimate of  $\sigma$ . To

show that it is approximate, we will use the symbol  $SE_{\bar{Y}}$ . The approximate standard error of the mean is

$$SE_{\bar{Y}} = \frac{s}{\sqrt{n}}.$$

According to this simple relationship, all we need is one random sample to approximate the spread of the entire sampling distribution for  $\bar{Y}$ . The quantity  $SE_{\bar{Y}}$  is usually called the “[standard error of the mean](#).”

The *standard error of the mean* is estimated from data as the sample standard deviation,  $s$ , divided by the square root of the sample size,  $n$ .

Calculating  $SE_{\bar{Y}}$  is so routine in biology that a sample mean should never be reported without it. For example, if we were submitting the results of our unique random sample of 100 genes in [Figure 4.1-2](#) for publication, we would calculate  $SE_{\bar{Y}}$  from the results in [Table 4.1-2](#) as follows:

$$SE_{\bar{Y}} = \frac{2521.6}{\sqrt{100}} = 252.2.$$

We would then report the sample mean in the text of the paper as  $3328.1 \pm 252.2(SE)$ .

Every estimate, not just the mean, has a sampling distribution with a standard error, including the proportion, median, correlation, difference between means, and so on. In the rest of this book, we will give formulas to calculate standard errors of many kinds of

estimates. The standard error is the usual way to indicate uncertainty of an estimate.

## 4.3 Confidence intervals

The [confidence interval](#) is another common way to quantify uncertainty about the value of a parameter. It is a range of numbers, calculated from the data, that is likely to contain within its span the unknown value of the target parameter. In this section, we introduce the concept without showing exact calculations. Confidence intervals can be calculated for means, proportions, correlations, differences between means, and other population parameters, as later chapters will demonstrate.

A *confidence interval* is a range of values surrounding the sample estimate that is likely to contain the population parameter.

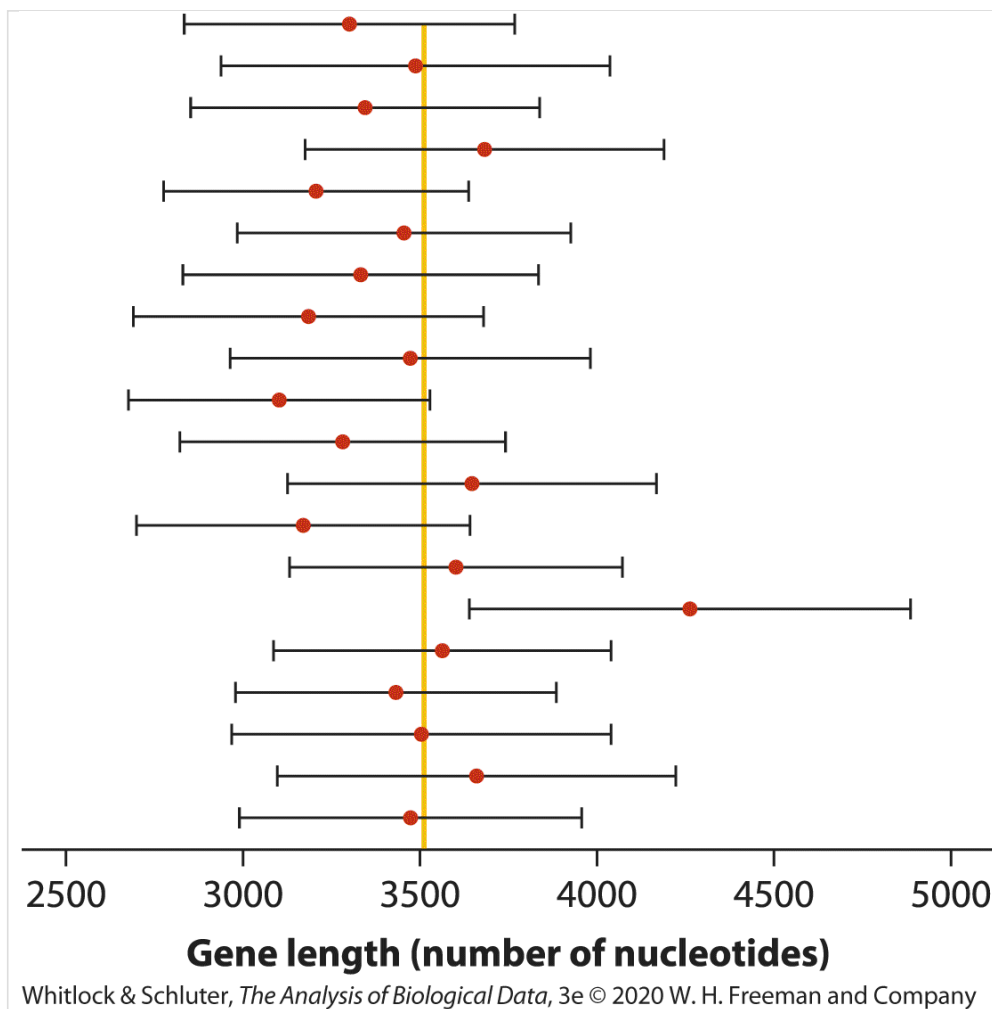
For example, we'll start by describing the [95% confidence interval for the mean](#). This confidence interval is a range likely to contain the value of the true population mean  $\mu$ . It is calculated from the data and extends above and below the sample estimate  $\bar{Y}$ . You will encounter confidence intervals frequently in the biological literature. We'll show you in [Chapter 11](#) how to calculate an exact confidence interval for the mean, but for now we give you the result and its interpretation. The 95% confidence interval for the mean calculated from the unique sample of 100 genes ([Table 4.1-2](#)) is

$$2827.8 < \mu < 3828.4$$

For this example, 2827.8 is the *lower limit* of the confidence interval, whereas 3828.4 is the *upper limit*. This calculation allows us to say, “We are 95% confident that the true mean lies between 2827.8 and 3828.4 nucleotides.” We do *not* say that “there is a 95% probability that the population mean falls between 2827.8 and 3828.4 nucleotides,” which is a common misinterpretation of the confidence interval (2827.8 and 3828.4 are both constants, and the true mean either is or is not between them, so there's no probability involved).

To better understand the correct interpretation of “95% confidence,” imagine that 20 researchers independently take unique random samples of  $n=100$  genes

from the human genome. Each researcher calculates an estimate  $\bar{Y}$  and then a 95% confidence interval for the parameter (the population mean,  $\mu$ ). Each researcher ends up with a different estimate and a different 95% confidence interval, because by chance their samples are not the same (Figure 4.3-1). On average, however, 19 out of 20 (95%) of the researchers' intervals will contain the value of the population parameter. On average, therefore, 1 out of 20 intervals (5%) will *not* contain the parameter value. None of the researchers will know for sure whether his or her own confidence interval contains the value of the unknown parameter, but each can be “95% confident” that it does.<sup>7</sup>



**FIGURE 4.3-1**

The 95% confidence intervals for the mean calculated from 20 separate random samples of  $n=100$  genes from the known human genome. Dots indicate the sample means  $\bar{Y}$ ; the vertical line (gold) represents the population mean,

$\mu=3328.1$ ,  $\mu = 3328.1$ . In this example, 19 of 20 intervals included the population mean, whereas one interval did not.

### Description

A set of 20 line segments are drawn over a horizontal number line that represents gene length (in number of nucleotides) ranging from 2500 to 5000 with an interval of 500. Each line segment represents confidence intervals for a sample with its corresponding mean value  $\bar{Y}$  plotted as a dot, at the center of each line segment. A vertical line representing the population mean at 3500 passes perpendicularly through all the line segments. Majority of the dots are present between 3100 and 3400. 19 out of 20 line segments are intersected by the population mean line, except one line segment that has a mean value of 4,250. All the values are approximate.

All numbers falling between the lower and upper bounds of a confidence interval can be regarded as the most plausible values for the parameter, given the data sampled. Values falling outside the confidence interval are less plausible. For example, on the basis of our random sample of 100 genes, we can say that a mean gene length of 2500 nucleotides in the whole genome is not among the most plausible, because it falls outside the 95% confidence interval,  $2827.8 < \mu < 3828.4$ . However, a mean gene length of 3000 nucleotides falls within the 95% confidence interval, and so remains among the most plausible.

In general, the width of the 95% confidence interval is a good measure of our uncertainty about the true value of the parameter. If the confidence interval is broad, then uncertainty is high and the data are not very informative about the location of the population parameter. If the confidence interval is narrow, on the other hand, then we can be confident that the parameter is close to the estimated value.

The 95% confidence interval provides a most-plausible range for a parameter. Values lying within the interval are most plausible, whereas those outside are less plausible, based on the data.

## The 2SE rule of thumb

A good “quick-and-dirty” approximation to the 95% confidence interval for the population mean is obtained by adding and subtracting two standard errors from the sample mean (the so-called 2SE rule of thumb). This calculation assumes that the sample is a random sample.

A rough approximation to the 95% confidence interval for a mean can be calculated as the sample mean plus and minus two standard errors.



For our unique random sample of 100 genes ([Figure 4.1-2](#)), for example, the sample mean of gene length was  $\bar{Y} = 3328.1$  nucleotides, and its standard error was  $SE_{\bar{Y}} = 252.2$  nucleotides. Two standard errors below the mean provides the lower limit of the approximate confidence interval:

$$\bar{Y} - 2SE_{\bar{Y}} = 3328.1 - (2 \times 252.2) = 2823.7, \bar{Y} - 2SE_{\bar{Y}} = 3328.1 - (2 \times 252.2) = 2823.7,$$

and two standard errors above the mean provides the upper limit:

$$\bar{Y} + 2SE_{\bar{Y}} = 3328.1 + (2 \times 252.2) = 3832.5. \bar{Y} + 2SE_{\bar{Y}} = 3328.1 + (2 \times 252.2) = 3832.5.$$

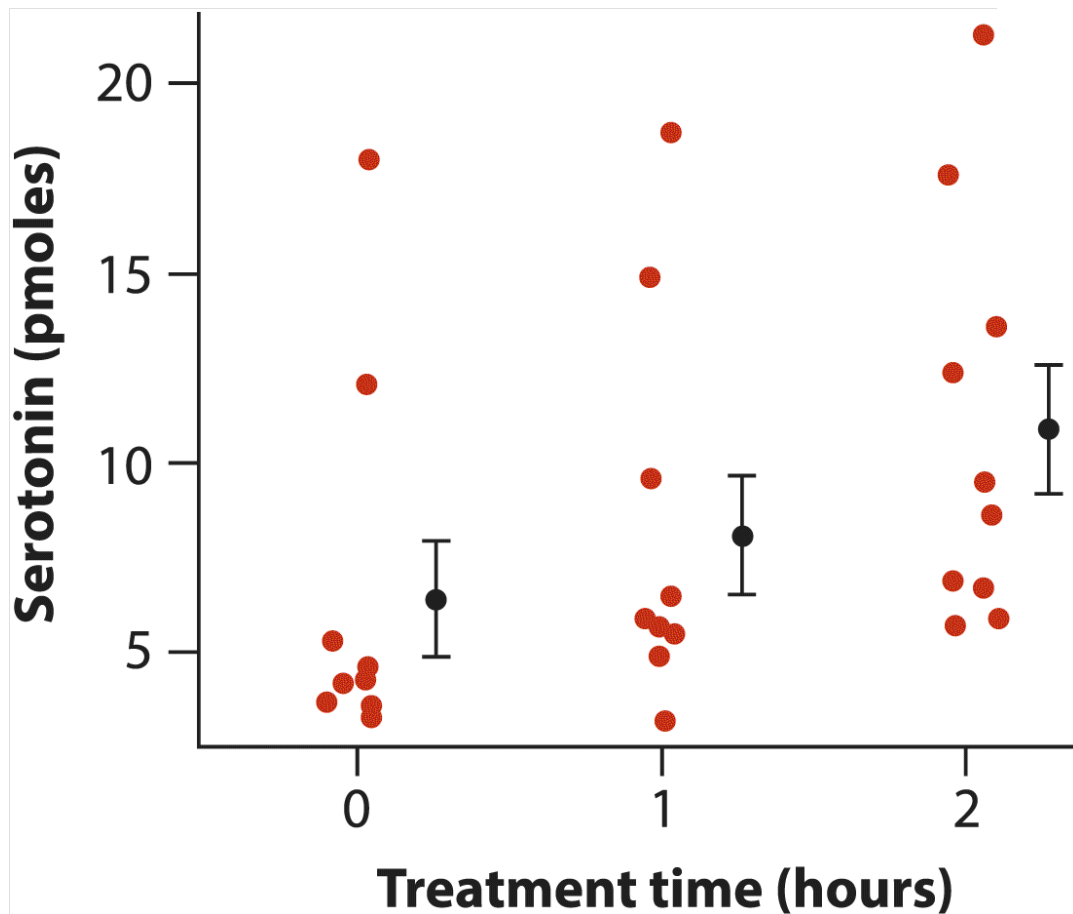
According to the 2SE rule of thumb, then, the 95% confidence interval for the mean gene length in the population can be approximated as

$$2823.7 < \mu < 3832.5. 2823.7 < \mu < 3832.5.$$

This is not too far off from the more exact confidence interval (i.e., between 2827.8 and 3828.4 nucleotides) that we calculated previously. Although approximate, the 2SE rule is simple and works reasonably well.

## 4.4 Error bars

Standard errors or confidence intervals for the mean (and other parameters) are often illustrated graphically with “[error bars](#).” Error bars are lines on a graph that extend outward from the sample estimate to illustrate the precision of estimates, reflecting uncertainty about the value of the parameter being estimated. For example, [Figure 4.4-1](#) reproduces the strip chart of locust serotonin data shown previously in [Chapter 2 \(Figure 2.1-2\)](#), but adds error bars to illustrate the standard error of the sample mean serotonin level in each of the three experimental treatments. The lines projecting outward from the sample mean indicate one standard error above the mean and one standard error below the mean. Remember that standard error bars, unlike the whiskers on a box plot, are not intended to span a specified fraction of the data. Error bars indicate uncertainty about the population parameter, not variability in the data (even though variability in the data contributes to uncertainty).



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

#### FIGURE 4.4-1

Strip chart of locust serotonin data (from [Figure 2.1-2](#)) with error bars added to illustrate the standard error (SE) of the mean for each treatment. Each filled black dot indicates the sample mean. Lines projecting outward indicate one SE above the mean and one SE below the mean.

#### Description

The horizontal axis is labeled “Treatment time in hours”, marked 0, 1, and 2. The vertical axis is labeled “Serotonin in pmoles”, ranging from 5 to 20 with increments of 5. When treatment hours is 0, the serotonin lies between 1 and 18. The error bar indicates: minimum serotonin as 5; median as 6; maximum serotonin as 7 point 5.

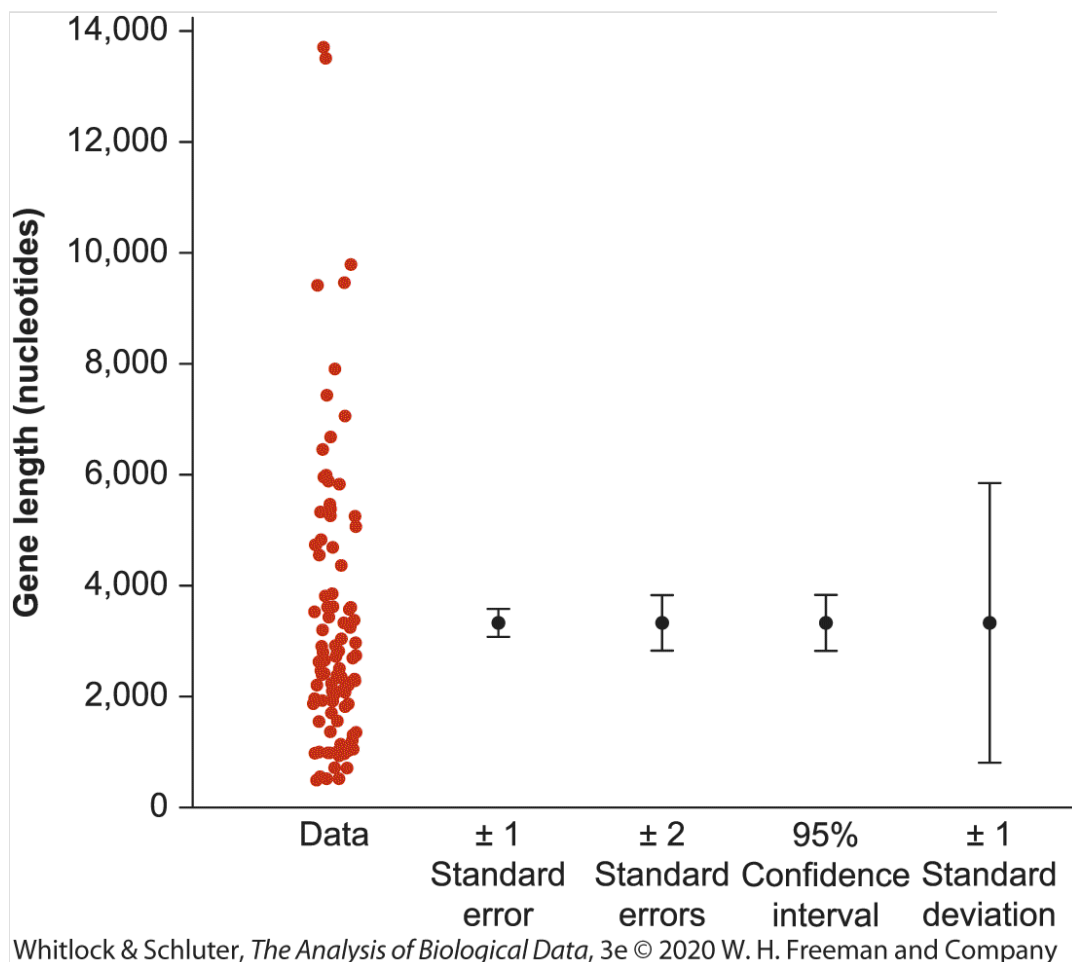
When treatment hours is 1, the serotonin lies between 1 and 19. The error bar indicates: minimum serotonin as 6; median as 8; maximum serotonin as 10.

When treatment hours is 2, the serotonin lies between 6 and 22. The error bar indicates: minimum serotonin as 9; median as 11; maximum serotonin as 12 point 5.

*Error bars* are lines on a graph extending outward from the sample estimate to illustrate uncertainty about the value of the parameter being estimated.

Error bars are used for multiple purposes, so they don't always show the same measure of precision. Often they are used to illustrate standard errors, but sometimes error bars show confidence intervals instead. They may even indicate *two* standard errors rather than one. For example,

[Figure 4.4.2](#) draws an error bar for each of these three measures of precision for the mean of the same random sample. Notice that because they are measuring different quantities, the three error bars do not have the same span. Therefore, it is crucial to read carefully the caption of any figure that has error bars to determine which measure of uncertainty is being shown. (And when you draw graphs yourself, it is important to give this information clearly in the figure legend.) Most commonly, error bars are used either for 95% confidence intervals or for standard errors. Because these two quantities differ approximately by a factor of two, you can see how knowing the meaning of the error bars is important.



**FIGURE 4.4-2**

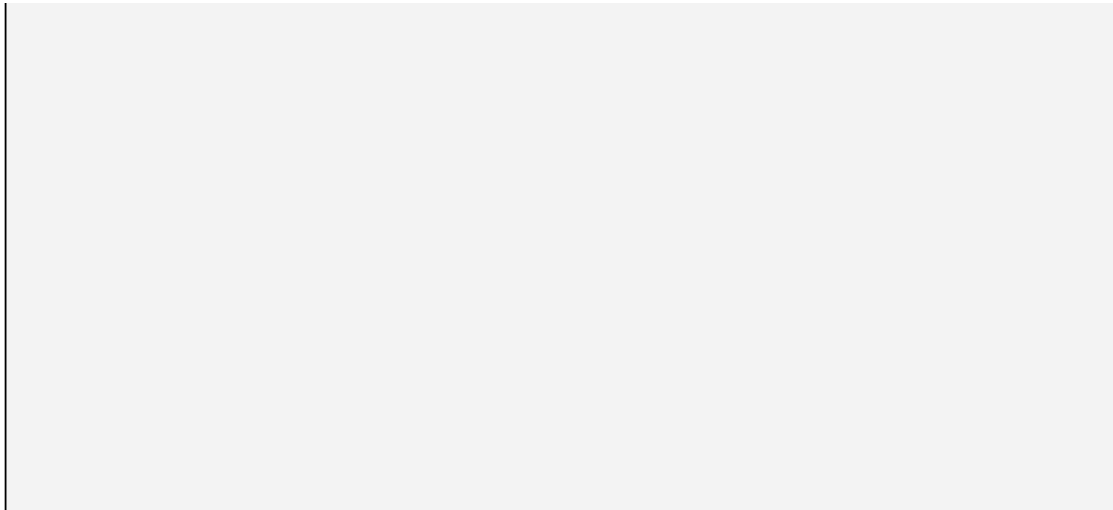
Comparison of alternative error bars calculated from gene lengths in the unique random sample of  $n=100$   **$n = 100$**  genes ([Example 4.1](#)). The data are plotted as a strip chart on the left. The filled black circles indicate the sample mean of gene length, 3328.1 nucleotides. The leftmost error bar visualizes one standard error of the mean (SE). The line extending above the black dot indicates one SE above the mean; the line extending below indicates one SE below the mean. The adjacent error bar indicates two standard errors above and below the mean. The third error bar indicates the 95% confidence interval for the mean. The rightmost error bar indicates one standard deviation above and below the sample mean.

### Description

The horizontal axis marks Data, plus or minus 1 standard error, plus or minus 2 standard errors, 95 percent confidence interval, and plus or minus 1 standard deviation along the horizontal axis. The vertical axis is labeled “Gene length in nucleotides”, ranging from 0 to 6000 with increments of 1000.

The approximate data are as follows. The clusters of points are most dense between 1 and 5000 against Data.

The error bar against plus or minus 1 standard error indicates: minimum gene length as 2250; maximum gene length as 2500; median as 2375. The error bar against plus or minus 2 standard errors indicates: minimum gene length as 2100; maximum gene length as 2600; median as 2350. The error bar against 95 percent confidence interval indicates: minimum gene length as 2100; maximum length as 2600; median as 2350. The error bar against plus or minus 1 standard deviation indicates: minimum gene length as 1000; maximum length as 3800; median as 2400.



Finally, error bars are sometimes used to indicate the standard deviation of the data, but we recommend against this practice to minimize confusion. Error bars are a poor method for illustrating variability in the data, and they are redundant if you show the data. We added an error bar for the standard deviation to [Figure 4.4-2](#) only to show how different—and potentially misleading—it can be. Use error bars only to illustrate the precision of estimates, not variability in the data.

## 4.5 Summary

- Estimation is the process of inferring a population parameter from sample data.
- All estimates have a sampling distribution, which is the probability distribution of all the possible values of the estimate that might be obtained under random sampling with a given sample size.
- The standard error of an estimate is the standard deviation of its sampling distribution. The standard error measures precision. The smaller the standard error, the more precise is the estimate.
- The usual formulas for standard errors and confidence intervals assume that sampling is random.
- The standard error of an estimate declines with increasing sample size.
- The confidence interval is a range of values calculated from sample data that is likely to contain within its span the value of the target parameter. On average, 95% confidence intervals calculated from independent random samples will include the value of the parameter 19 times out of 20.
- The 2SE rule of thumb (i.e., the sample mean plus or minus two standard errors) provides a rough approximation to the 95% confidence interval for a mean.
- Add error bars to graphs to illustrate standard errors or confidence intervals. Make sure to clarify which is being illustrated in the figure legend.



## 4.6 Quick formula summary

### Standard error of the mean

**What is it for?** Measuring the precision of the sample estimate  $\bar{Y}$  of the population mean  $\mu$ .

**What does it assume?** The sample is a random sample.

**Estimate:**  $SE_{\bar{Y}}$

**Parameter:**  $\sigma_{\bar{Y}}$

**Formula:**  $SE_{\bar{Y}} = \frac{s}{\sqrt{n}}$

where  $s$  is the sample standard deviation and  $n$  is the sample size.

$SE_{\bar{Y}}$  estimates the quantity  $\sigma_{\bar{Y}} = \frac{\sigma}{\sqrt{n}}$ , where  $\sigma_{\bar{Y}}$  is the standard error of the sample mean, and  $\sigma$  is the standard deviation of  $Y$  in the population.

#### Online resources

Learning resources associated with this chapter are online at <https://whitlockschluter3e.zoology.ubc.ca/chapter04.html>.

## Chapter 4 Problems

### PRACTICE PROBLEMS

*Answers to the Practice Problems are provided in the [Answers Appendix](#) at the back of the book.*

1. **Calculation practice: Standard error of the mean and approximate confidence intervals for the mean.** We will use the same data for systolic blood pressure collected for Calculation [Practice Problem 1 in Chapter 3](#). Here again are the data points:

112, 112, 128, 128, 108, 108, 129, 129, 125, 125, 153, 153, 155, 155, 132, 132, 137, 137

The mean is 131.0 mm Hg, and the variance is 254.5.

- a. What is  $s^s$ , the standard deviation of these data?
  - b. What is  $n^n$ , the sample size?
  - c. Calculate the standard error of the mean.
  - d. Using the 2SE rule of thumb, calculate an approximate 95% confidence interval for the mean. Provide the lower and upper limits.
2. Examine the times to rigor mortis of the 114 human corpses tabulated in [Practice Problem 9 of Chapter 3](#).
    - a. What is the standard error of the mean time to rigor mortis?
    - b. The standard error calculated in part (a) measures the spread of what frequency distribution?
    - c. What assumption does your calculation in part (a) require?

3. Examine the frequency distribution of gene lengths in the human genome displayed in [Figure 4.1-1](#). Is the population median gene length in the human genome likely to be larger, smaller, or equal to the population mean? Explain.
4. As a general rule, is the spread of the sampling distribution for the sample mean mainly determined by the magnitude of the mean or by the sample size?
5. Seven of the 100 human genes that we sampled randomly from the human genome (in [Example 4.1](#)) were found to occur on the X chromosome. The sample fraction of genes on the X was thus  $\hat{p} = 7/100 = 0.07$ . For each of the following statements, specify whether it is true or false:
  - a.  $\hat{p} = 0.07$  is the fraction of all human genes on the X chromosome.
  - b.  $\hat{p} = 0.07$  estimates  $p$ , the fraction of all human genes on the X chromosome.
  - c.  $\hat{p}$  has a sampling distribution representing the frequency distribution of values of  $\hat{p}$  that we might obtain when we randomly sample 100 genes from the human genome.
  - d. The fraction of all human genes on the X chromosome has a sampling distribution.
  - e. The standard deviation of the sampling distribution of  $\hat{p}$  is the standard error of  $\hat{p}$ .
6. In a poll of 1641 people carried out in Canada in November 2005, 73% of people surveyed agreed with the statement that “you don’t really

expect that politicians will keep their election promises once they are in power” ([CBC News 2005](#)).

- a. What is the parameter being estimated?
  - b. What is the value of the sample estimate?
  - c. What is the sample size?
  - d. The poll also reported that “the results are considered accurate within 2.5 percentage points, 19 times out of 20.” Explain what this statement likely refers to.
7. The following data are flash durations, in milliseconds, of a sample of 35 male fireflies of the species *Photinus ignitus* ([Cratsley and Lewis 2003](#); see [Assignment Problem 19 in Chapter 2](#)):
- 79,79, 80,80, 82,82, 83,83, 86,86, 85,85, 86,86, 86,86, 88,88, 87,87, 89,  
89, 89,89, 90,90, 92,92, 94,94, 92,92, 94,94, 96,96, 95,95, 95,95, 95,95,  
96,96, 98,98, 98,98, 98,98, 101,101, 103,103, 106,106, 108,108, 109,109,  
112,112, 113,113, 118,118, 116,116, 119,119
- a. Estimate the sample mean flash duration. What does this quantity estimate?
  - b. Is the estimate in part (a) likely to equal the population parameter? Why or why not?
  - c. Calculate a standard error for your sample estimate.
  - d. What does the quantity in part (c) measure?
  - e. Using an approximate method, calculate a rough 95% confidence interval for the population mean.
  - f. Provide an interpretation for the interval you calculated in part (e).

8. Imagine that the results of a study calculated a sample mean of zero, with a narrow 95% confidence interval for the population mean.<sup>8</sup> The most appropriate conclusion is that (choose one):
- The population mean is likely to be zero or close to zero.
  - The population mean is probably zero, but there is some chance that it is either slightly less than zero or slightly greater than zero.
  - We can be reasonably certain the mean differs from zero.
9. One of the great discoveries of biology is that organisms have a class of genes called “regulatory genes,” whose only job is to regulate the activity of other genes. How many genes does the typical regulatory gene regulate? A study of interaction networks in yeast (*S. cerevisiae*) came up with the following data for 109 regulatory genes ([Guelzim et al. 2002](#)).

Number of genes regulated	Frequency
11	2020
22	1010
33	77
44	77
55	88
66	88
77	55
88	22
99	44
1010	44
1111	33
1212	44

13	13	55
14	14	11
15	15	22
16	16	11
17	17	33
18	18	22
19	19	22
20	20	33
22	22	33
25	25	11
26	26	11
28	28	11
29	29	11
37	37	11
Total	Total	109

- What type of graph should be used to display these data?
  - What is the estimated mean number of genes regulated by a regulatory gene in the yeast genome?
  - What is the standard error of the mean?
  - Explain what this standard error measures.
  - What assumption are you making in part (c)?
10. Refer to the previous problem ([Practice Problem 9](#)).
- Using an approximate method, provide a rough 95% confidence interval for the population mean.

- b. Provide an interpretation of the interval you calculated in part (a).
11. [Goldman et al. \(1988\)](#) analyzed data on 405 patients with white blood cell cancer (chronic myelogenous leukemia) who received bone marrow transplants. They estimated the probability of relapse within 4 years of treatment to be 0.19, with a 95% confidence interval of 0.12 to 0.28. Which of the following statements are true?
- The population proportion is 0.19.
  - The population proportion is likely to be between 0.12 and 0.28.
  - There is a 95% chance that the population proportion is between 0.12 and 0.28.
  - A population proportion of 0.30 is more plausible than a value of 0.25.
12. An absentminded (and not too clever) scientist friend of yours has just analyzed his data, and he has two numbers—25.4 and 2.54—written on a scrap of paper. He says: “I remember that one of these is the standard deviation of my data and the other is the standard error of the mean, but I can’t remember which is which. Can you help?”
- Which number is the standard deviation and which is the standard error of the mean?
  - What was your friend’s sample size?
13. The following is a list of sample means for human adult height, in centimeters. Each was calculated in the same way from samples taken from the same hypothetical population.
- 160.5, 160.5, 162.5, 162.5, 161.7, 161.7, 160.2, 160.2, 163.7, 163.7, 159.8, 159.8, 160.6, 160.6, 161.1, 161.1

The true mean of the population is 158.7 cm. Use the jargon of estimation to describe the likely type of problem in the sampling

process.

14. When planning to obtain a sample from a population of interest, what can you do to make the standard error of the mean smaller?

#### ASSIGNMENT PROBLEMS

*Answers to all Assignment Problems are available for instructors, by contacting [DL-WhitlockSchluter3e@macmillan.com](mailto:DL-WhitlockSchluter3e@macmillan.com).*

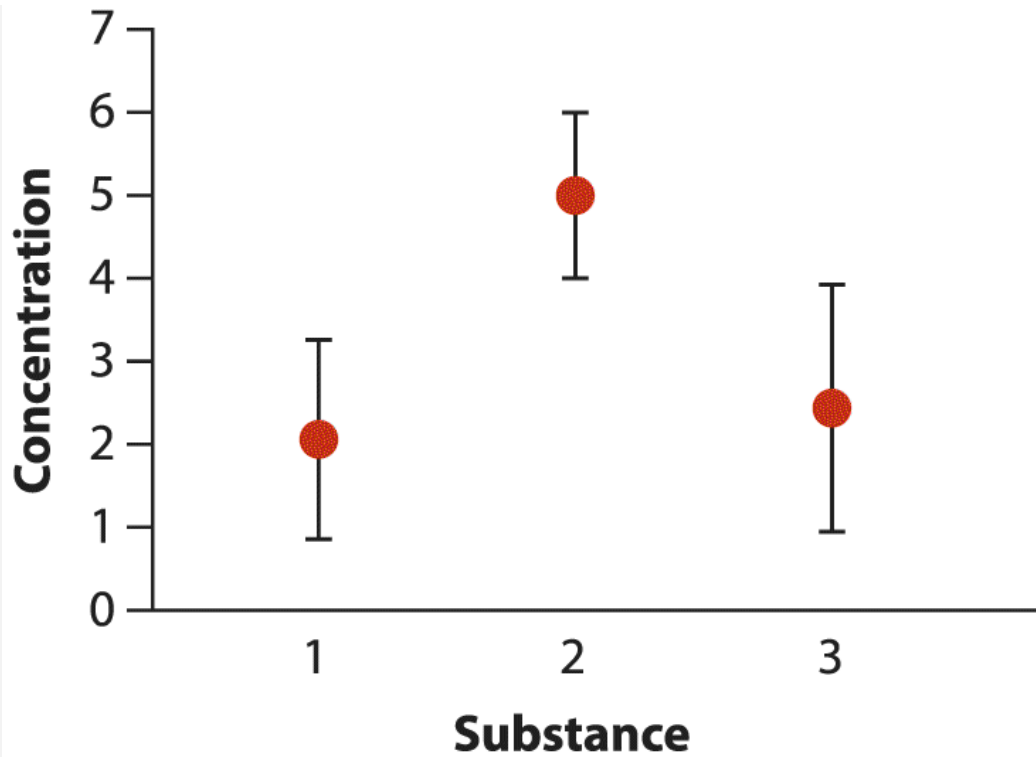
15. A massive survey of sexual attitudes and behavior in Britain between 1999 and 2001 contacted 16,998 households and interviewed 11,161 respondents aged 16–44 years (one per responding household). The frequency distributions of ages of men and women respondents were the same. The following results were reported on the number of heterosexual partners that individuals had over the previous five-year period ([Johnson et al. 2001](#)).

	Sample size, $n$	Mean	Standard deviation
Men	4620	3.8	6.7
Women	6228	2.4	4.6

- a. What is the standard error of the mean in men? What is it in women? Assume that the sampling was random.
- b. Which is a better descriptor of the variation among men in the number of sexual partners, the standard deviation or the standard error? Why?
- c. Which is a better descriptor of uncertainty in the estimated mean number of partners in women, the standard deviation or the standard error? Why?



- d. A mysterious result of the study is the discrepancy between the mean number of partners of heterosexual men and women. If each sex obtains its partners from the other sex, then the true mean number of heterosexual partners should be identical. Considering aspects of the study design, suggest an explanation for the discrepancy.
  - e. Using the 2SE rule of thumb, calculate an approximate 95% confidence interval for the mean number of heterosexual partners for men.
16. Our unique random sample of 100 human genes from the human genome ([Example 4.1](#)) was found to have a median length of 2640.5 nucleotides. Specify whether each of the following statements is true or false.
- a. The median gene length of all human genes is 2640.5 nucleotides.
  - b. The median gene length of all human genes is estimated to be 2640.5 nucleotides.
  - c. The sample median has a sampling distribution with a standard error.
  - d. A random sample of 1000 genes would likely yield an estimate of the median closer to the population median than a random sample of 100 genes.
17. The following figure is from the website of a U.S. national environmental laboratory.<sup>9</sup> It displays sample mean concentrations, with 95% confidence intervals, of three radioactive substances. The text accompanying the figure explained that “*the first plotted mean is  $2.0 \pm 1.1$ ,  $2.0 \pm 1.1$ , so there is a 95% chance that the actual result is between 0.9 and 3.1, a 2.5% chance it is less than 0.9, and a 2.5% chance it is greater than 3.1.*” Is this a correct interpretation of a confidence interval? Explain.



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

#### Description

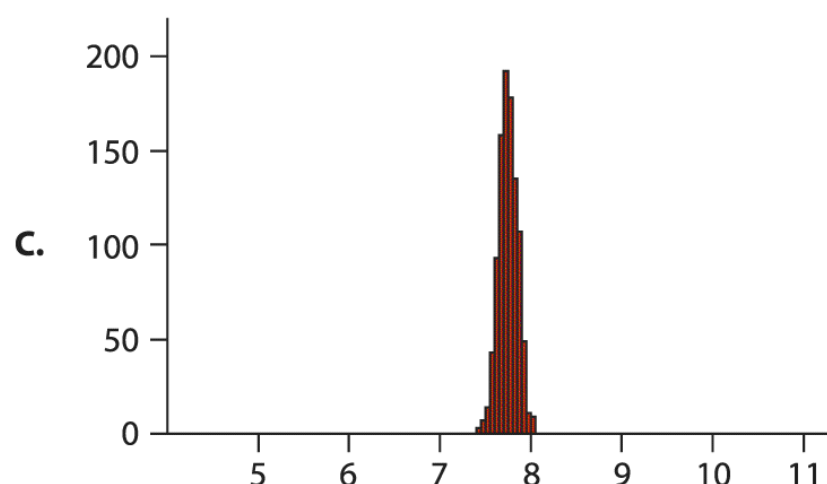
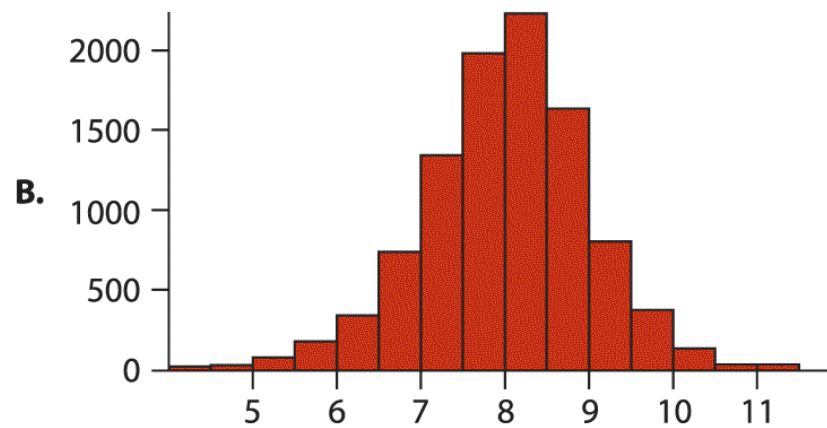
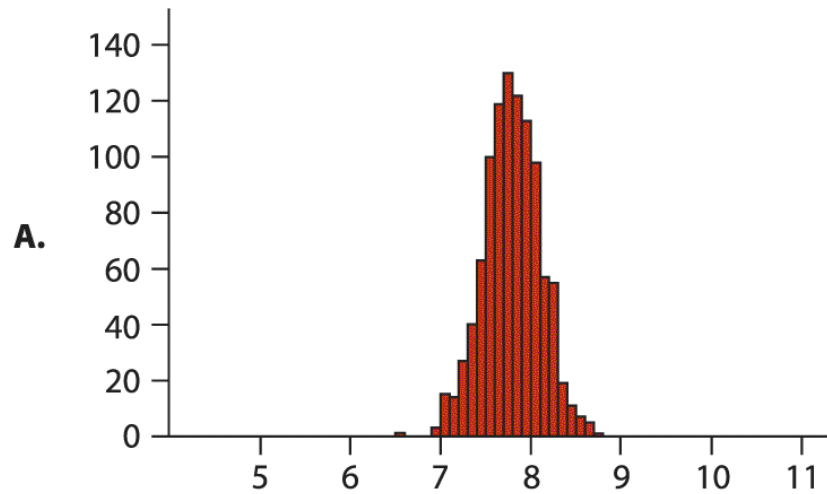
The vertical axis is labeled “Concentration”, ranging from 0 to 7 with increments of 1. The error bar 1 ranges from 1 to 3 point 5 with median 2. The error bar 2 ranges from 4 to 6 with median 5. The error bar 3 ranges from 1 to 4 with median 2 point 5.

18. *Amorphophallus johnsonii* is a plant growing in West Africa, and it is better known as a “corpse flower.” Its common name comes from the fact that when it flowers, it gives off a “powerful aroma of rotting fish and faeces” ([Beath 1996](#)). The flowers smell this way because their principal pollinators are carrion beetles, who are attracted to such a smell. [Beath \(1996\)](#) observed the number of carrion beetles (*Phaeochrous amplus*) that arrive per night to flowers of this species. The data are as follows:

51, 51, 45, 45, 61, 61, 76, 76, 11, 11, 117, 117, 7, 7, 132, 132, 52, 52, 149, 149

- What is the mean and standard deviation of beetles per flower?
- What is the standard error of this estimate of the mean?
- Give an approximate 95% confidence interval of the mean. Provide lower and upper limits.

- d. If you had been given 25 data points instead of 10, would you expect the mean to be greater than, less than, or about the same as the mean of this sample?
  - e. If you had been given 25 data points instead of 10, would you have expected the standard deviation to be greater than, less than, or about the same as this sample?
  - f. If you had been given 25 data points instead of 10, would you have expected the standard error of the mean to be greater than, less than, or about the same as this sample?
19. The following three histograms (A, B, and C) plot information about the number of hours of sleep adult Europeans get per night ([Roenneberg 2012](#)). One of them shows the frequency distribution of individual values in a random sample. Another shows the distribution of sample means for samples of size 10 taken from the same population. Another shows the distribution of sample means for samples of size 100.



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

**Description**

The approximate data in the first plot, A, are as follows. Most of the rectangular bars are drawn between 7 and 8 point 8 in such a way that the edges of the bars can be connected by a concave downward curve. The maximum frequency is at 130.

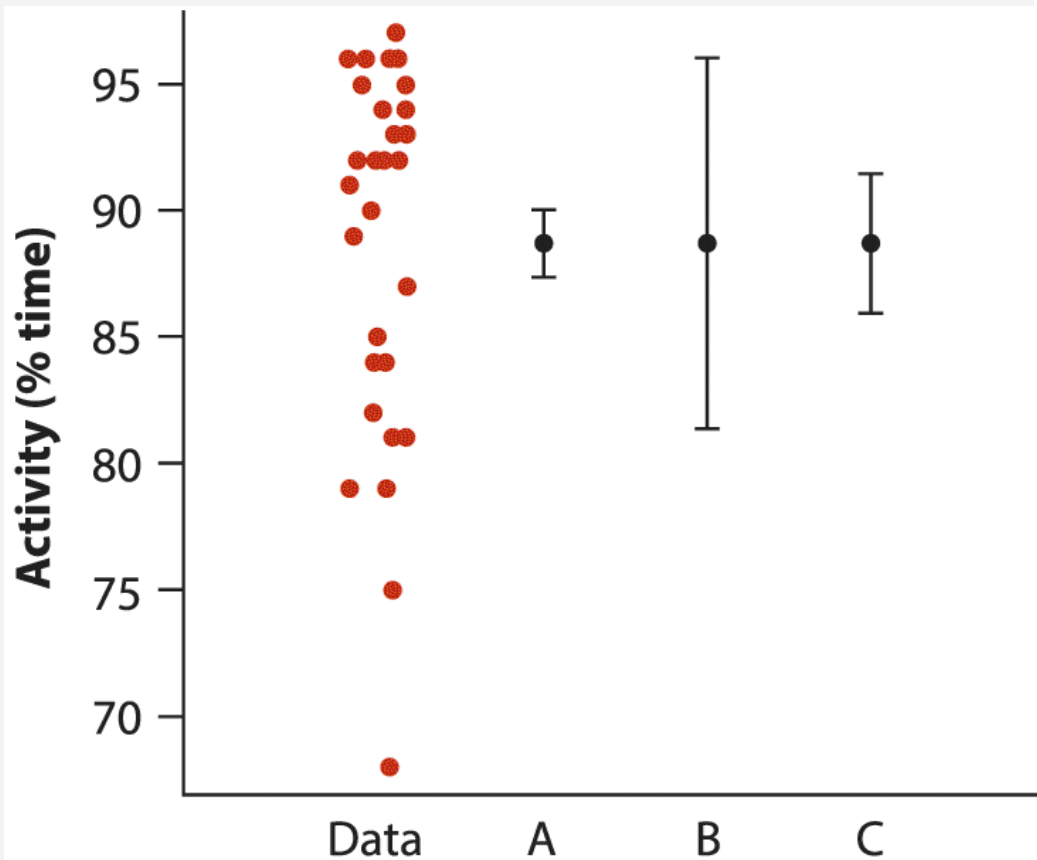
The approximate data in the first plot, B, are as follows. Most of the rectangular bars are drawn between 4 and 11 point 5 in such a way that the edges of the bars can almost be completely connected by a concave downward curve. The maximum frequency is at 2250.

The approximate data in the first plot, C, are as follows. Most of the rectangular bars are drawn between 7 point 5 and 8 in such a way that the edges of the bars can be connected by a concave downward curve. The maximum frequency is at 190.

- a. Identify which graph goes with which distribution.

- b. What features of these distributions allowed you to distinguish which was which?
  - c. Estimate by eye the approximate population mean of the number of hours of sleep using the distribution for the data.
  - d. Estimate by eye the approximate mean of the distributions of sample means.
20. The HIV virus has a high mutation rate compared with bacteria and multicellular life. Many of these mutations are bad for the virus, causing it to replicate more slowly. [Theys et al. \(2018\)](#) assembled a very large database of known HIV mutants and their replication rates.
- We took a random sample of 100 mutants from this database using the computer and calculated the sample median replication rate and the sample mean rate. We repeated this process a vast number of times, calculating the sample median and mean each time. The standard deviation of the distribution of sample means was 0.00073. The standard deviation of the distribution of sample medians was 0.0224.
- a. What is the standard error of the mean replication rate?
  - b. What is the standard error of the median replication rate?
  - c. Which measure of the central tendency (mean or median) is likely to be estimated with greater precision if a random sample of 100 data mutants is sampled from the database: the mean or the median? Explain your answer.
21. Is sleep necessary? To investigate, [Lesku et al. \(2012\)](#) measured the activity patterns of breeding pectoral sandpipers (*Calidris melanotos*) in the high Arctic in summer, when the sun never sets. The accompanying figure shows the observed percent time that individual males were awake and active in a 2008 field study. The data are on the left. To the right of the data are the sample mean (filled circle) and error bars for the

standard deviation, the standard error of the mean, and a 95% confidence interval for the mean (in no particular order).



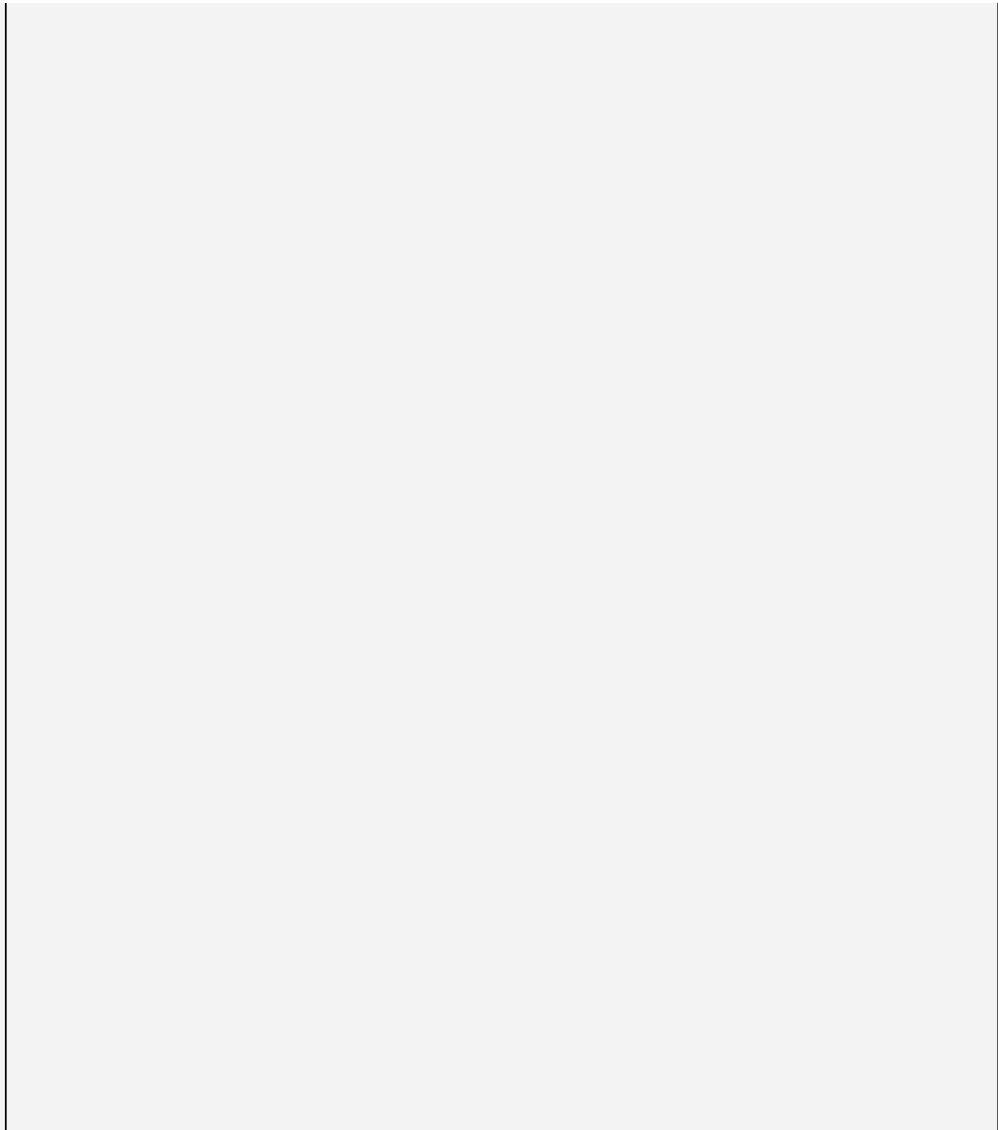
Whitlock & Schluter, *The Analysis of Biological Data*, 3e  
 © 2020 W. H. Freeman and Company

#### Description

The horizontal axis marks Data, A, B, and C. The vertical axis is labeled “Activity in percent time”, ranging from 70 to 95 with increments of 5. The approximate data are as follows. The clusters of points are most dense between 79 and 97 against Data.

The error bar against A indicates: minimum activity time as 88; maximum activity time as 90; median as 89. The error bar against B indicates: minimum activity time as 82; maximum activity time as 96; median as 89. The error bar against C indicates: minimum activity time as 86; maximum activity time as 92; median as 89.





- a. Which of the error bars indicates the standard deviation?
- b. Which error bar indicates the standard error of the mean?
- c. Which error bar indicates a 95% confidence interval for the mean?
- d. Estimate by eye the smallest value in the 95% confidence interval for the mean activity (% time) of male pectoral sandpipers. Using this value, calculate approximately the maximum number of hours (out of 24 hours) that males spend inactive or asleep that is consistent with this confidence interval.

22. How long do you hug somebody? [Nagy \(2011\)](#) measured the duration of spontaneous embraces at the 2008 Summer Olympic Games in Beijing, China. The data are the durations of hugs, in seconds, of athletes immediately after competing in the finals of an event. Hugs were either with their coach, a supporter (e.g., a team member), or a competitor. Descriptive statistics calculated from the data are in the following table;  $n$  refers to the sample size.

Relationship	Mean	Standard deviation	$n$
Coach	3.77	3.96	77
Supporter	3.16	2.76	75
Competitor	1.81	1.13	33

- According to the values in the table, which relationship group gets the longest hugs, on average, and which gets the briefest hugs? Do the values shown represent parameters or sample estimates? Explain.
- Using the numbers in the table, calculate the standard error of the mean hug duration for each relationship group. What do these values measure?
- What assumption(s) about the samples are you making in (b)?
- Using the numbers in the table, calculate an approximate 95% confidence interval for the mean hug duration when athletes embrace competitors. Provide the lower and upper limits of the confidence interval.
- In light of your results in (d), consider the 95% confidence interval for the mean duration of hugs with competitors in the population of athletes. Is 2 seconds among the most plausible values for the population mean hug duration?

f. For which of the relationship groups is the possibility of a 3-second mean hug duration in the population within the 95% confidence interval?

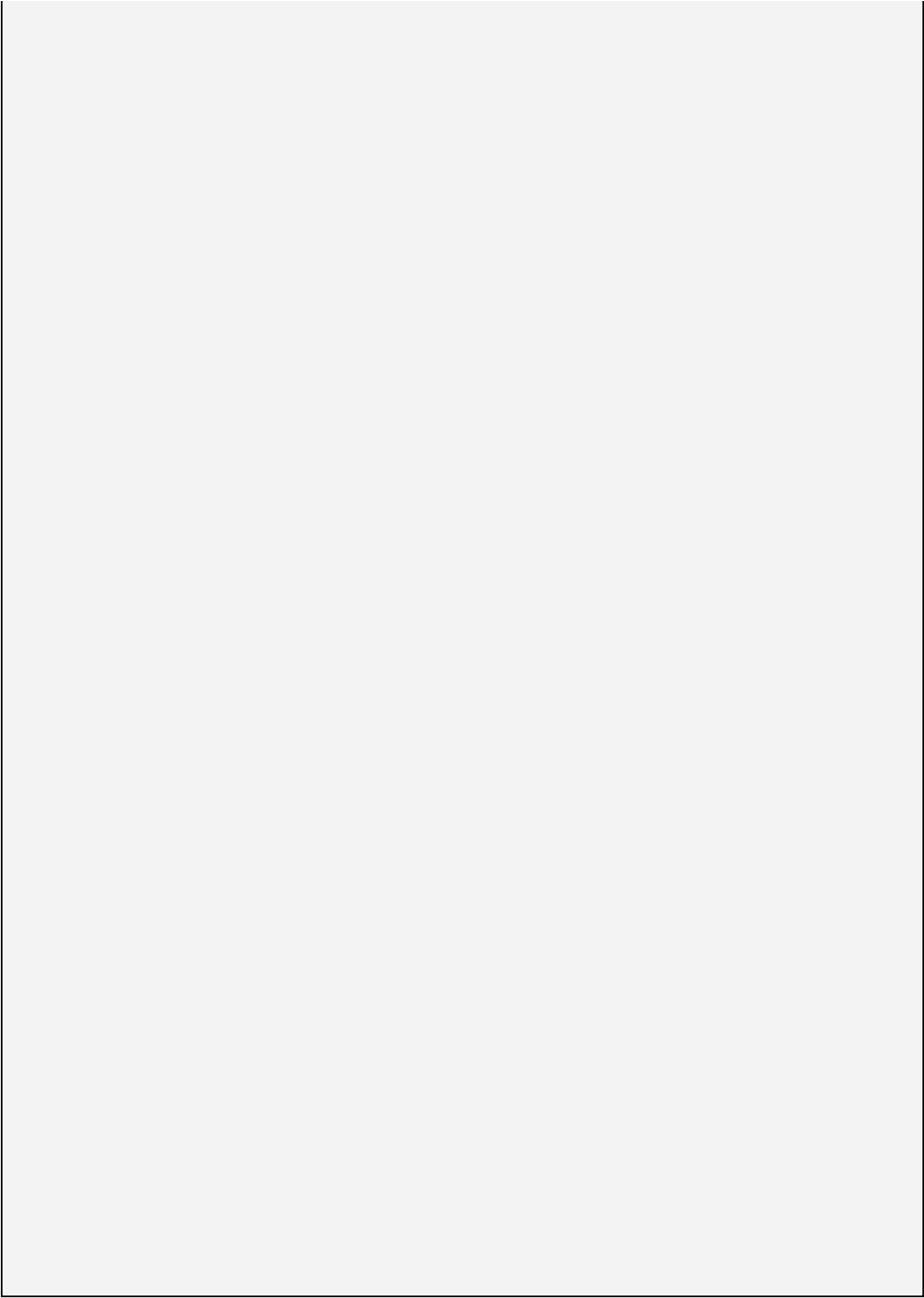
23. Pitcher plants of the genus *Nepenthes* are typically carnivorous, obtaining a great deal of their nutrition from insects that become trapped in the pitcher, die, and decay. *N. lowii*, a pitcher plant from Borneo, produces a second type of pitcher that attracts tree shrews (*Tupaia montana*), which provide nutrients by defecating into the pitcher<sup>10</sup> while they feed on a substance secreted by the plant. Based on measurements of 20 plants, [Clarke et al. \(2009\)](#) calculated a 95% confidence interval for the mean fraction of total leaf nitrogen in the plant species derived from tree shrews:  $0.57 < \mu < 1.0$ .  $0.57 < \mu < 1.0$ .



Scubazoo Images/Getty Images

**Description**

-

- 
- a. Does this result imply that individual plants receive between 57% and 100% of their leaf nitrogen from tree shrews? Explain.
  - b. Is the confidence interval meant to bracket the sample mean or the population mean?

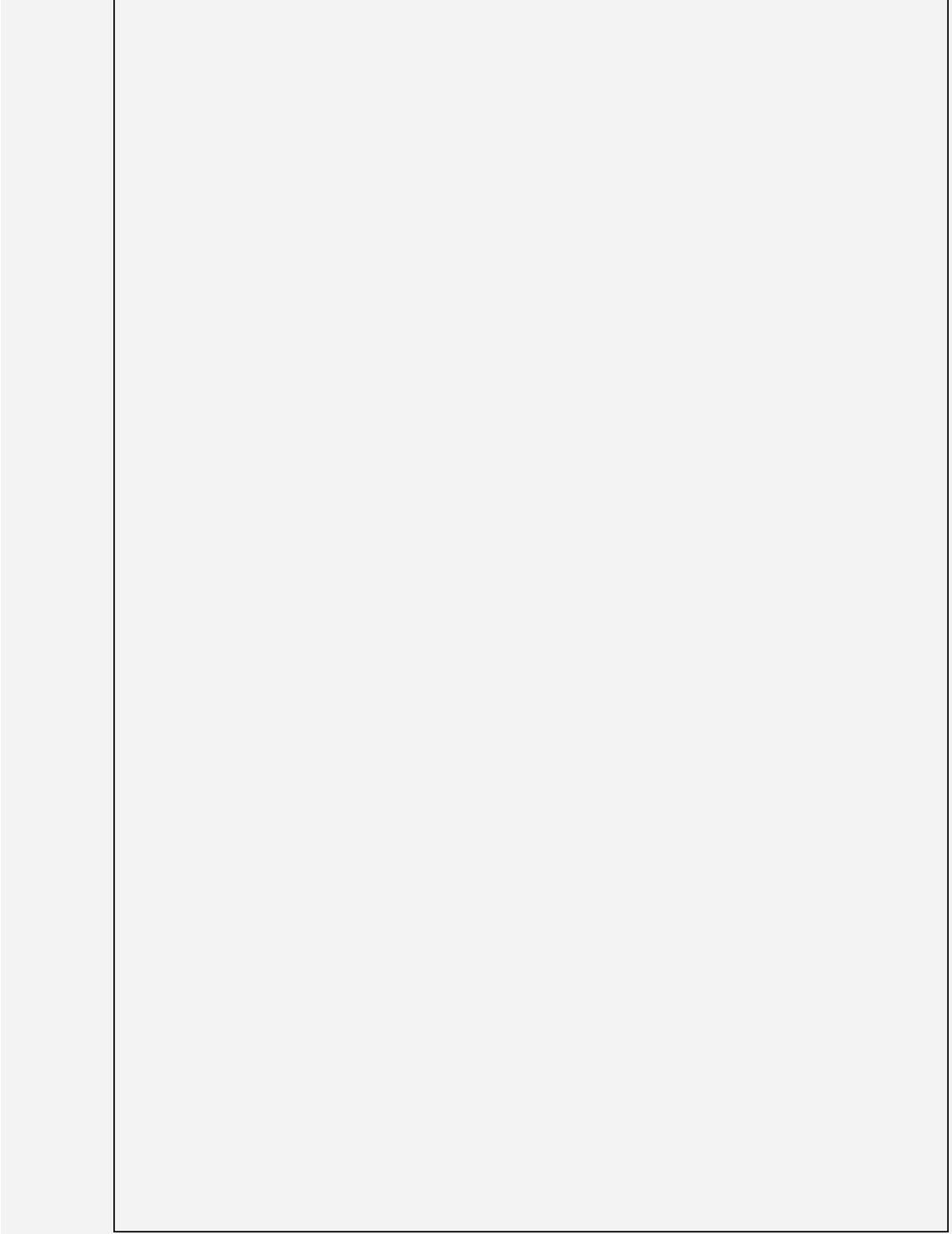
- c. Identify two values for the mean shrew fraction of total leaf nitrogen that the analysis suggests are among the most plausible.
  - d. Identify two values for the mean shrew fraction of total leaf nitrogen that the analysis suggests are less plausible.
24. [Hagen et al. \(2011\)](#) estimated the home range sizes of four bumblebees (*Bombus*) by fitting them with tiny radio transmitters and tracking their positions by plane and ground surveys. They estimated the mean home range size to be  $20.7 \pm 11.6$  ha, where the number after the  $\pm$  sign refers to standard error of the mean.



*tr3gin/Shutterstock.com*

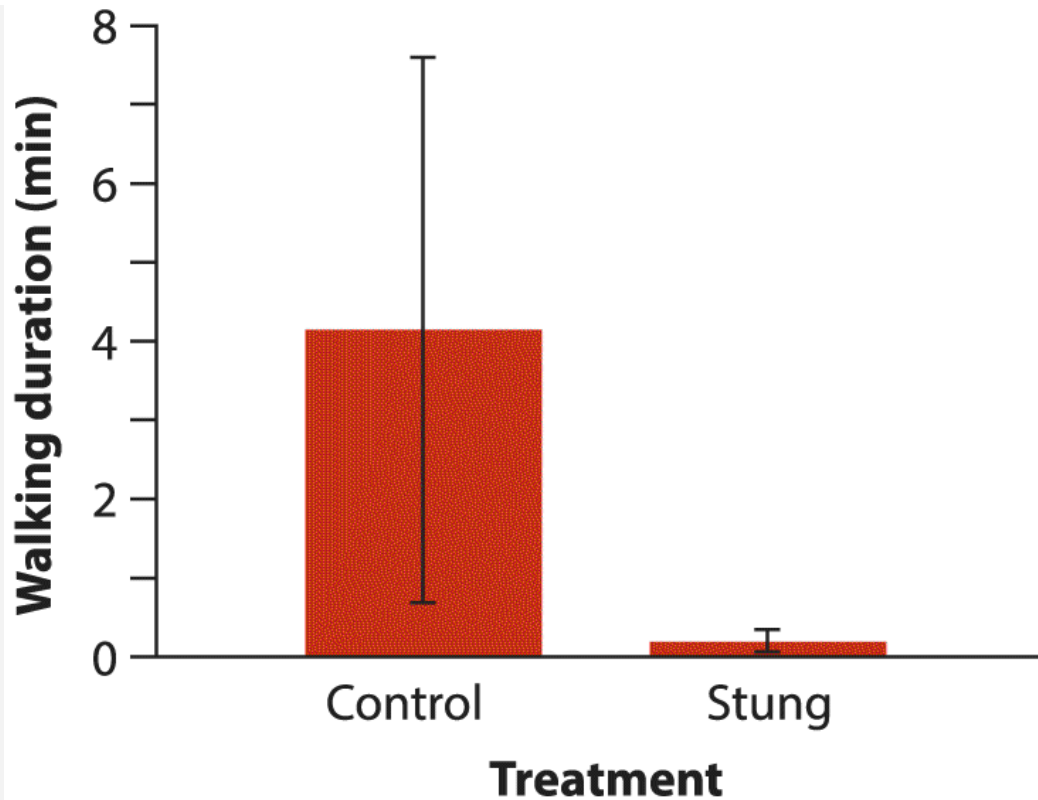
**Description**

-

- 
- a. Provide a description for the standard error. What does it measure?
  - b. What assumption are we making when calculating the standard error of the mean?



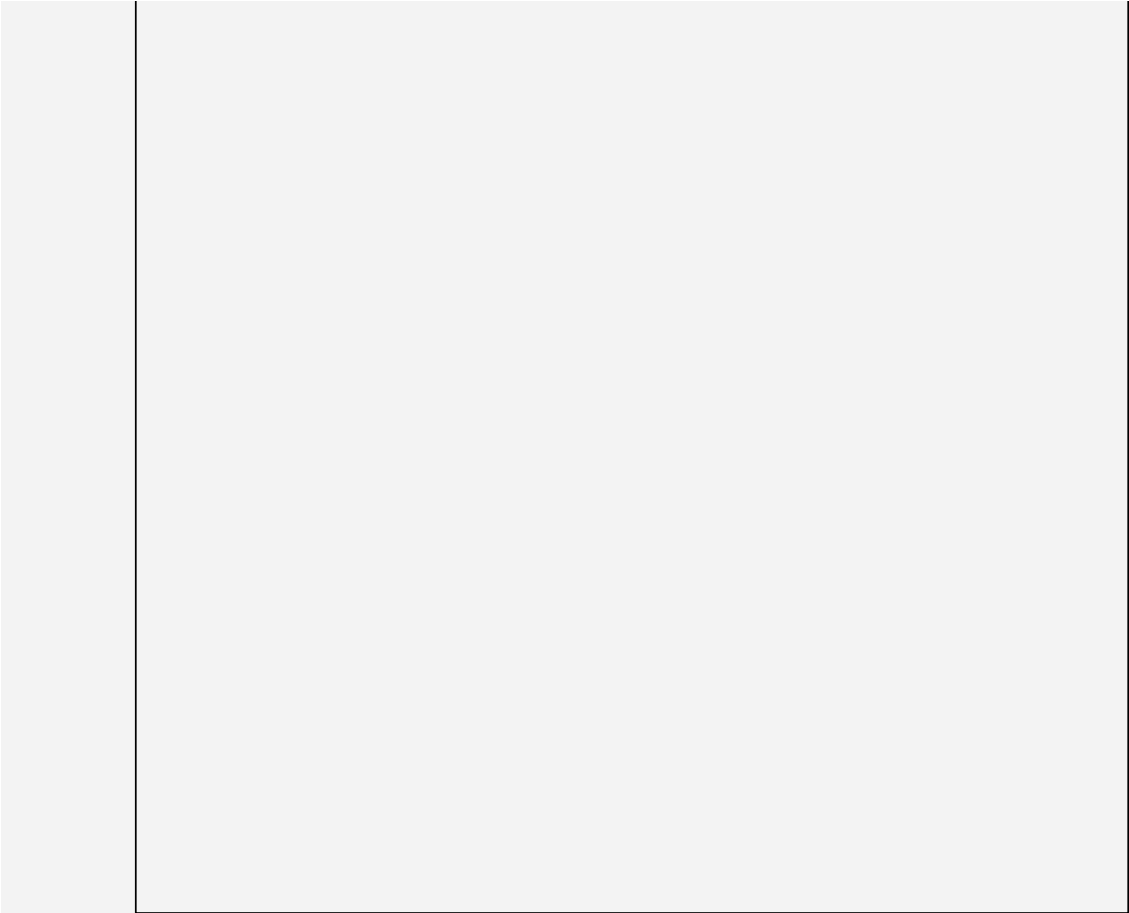
- c. What would you recommend the researchers do next to reduce the standard error of their estimate of the mean home range size?
25. The following definition of a confidence interval was found on a web page at the National Institute of Standards and Technology. “*Confidence intervals are constructed at a confidence level, such as 95%, selected by the user. . . . It means that if the same population is sampled on numerous occasions and interval estimates are made on each occasion, the resulting intervals would bracket the true population parameter in approximately 95% of the cases.*”<sup>11</sup> Is this a correct interpretation of the confidence interval? Explain.
26. When a female jewel wasp (*Ampulex compressa*) encounters a cockroach (*Periplaneta americana*), she stings and injects neurotoxins into its head that render the insect unable to initiate self-movement but not paralyzed. The wasp then holds the compliant (zombie) cockroach by the antenna and leads it to her nest, where it will become live food for her larval offspring. The following graph (data from [Gal and Libersat 2010](#)) compares the mean self-initiated walking duration of stung and control cockroaches during the first 30 minutes after treatment. The error bars indicate approximate 95% confidence intervals;  $n=5$  in each group.



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

#### Description

The horizontal axis is labeled “Treatment”, with data Control and Stung. The vertical axis is labeled “Walking duration in minutes”, ranging from 0 to 8 with increments of 2. The approximate data are as follows. Control, 4 and an error bar ranges from zero point five to seven point five; Stung, zero point two five and an error bar ranges from zero point zero five to zero point four five.

- 
- a. Estimate the lower and upper limits of the confidence intervals for the control group.
  - b. Approximate the value of the standard error for the control group.
  - c. Identify two values of the population mean duration for the control group that are among the most plausible.
  - d. Identify two values of the population mean duration for the stung group that are less plausible.
  - e. Identify the main weakness in the construction of the graph. How would you improve it?

## INTERLEAF 2

# Pseudoreplication

Most statistical techniques assume that the data are a random sample from the population, in which individuals are sampled independently and with equal probability. Unfortunately, many experiments are conducted and analyzed in a way that violates the assumption of independence.

For example, imagine that a team is trying to measure the average tusk length of contemporary African elephants, after decades of hunting and poaching have removed the animals with the largest tusks. For each of 12 individual elephants, a measurement of both the left and right tusk was taken, a total of 24 tusks. A confidence interval for the mean tusk length was calculated based on these 24 tusks. The result was a narrow confidence interval for the mean length.

**Pseudoreplication is probably the single most common fault in the design and analysis of ecological field experiments. It is at least as common in many other areas of research.**

—Stuart Hurlbert

But something has gone wrong. The statistical analysis was carried out as though the 24 tusk measurements comprised an independent sample. However, the 24 measurements of tusk length are not all independent, because they came from only 12 animals. The left and right tusks of the same elephant are not equivalent to two tusks picked at random from the elephant population. The real independent sample here is that of the 12 elephants, not of the 24 tusks. To ensure independence, the analysis should therefore use only 12 measurements: a single tusk per individual elephant, or the average of the two tusks per individual.

This hypothetical analysis of 24 tusks is an example of pseudoreplication. **Pseudoreplication** occurs whenever individual measurements that are not independent are analyzed as if they were independent. In the elephant study, the 24 measurements of tusk length were treated as 24 independent data points, when they were not independent. In general, measurements that are grouped within a randomly sampled unit, such as the two tusks per independent elephant, are likely to be more similar to one another than measurements obtained by randomly sampling from the population.

Replication by itself is good. It refers to the sampling of multiple units from a population, which makes it possible to estimate population characteristics and the precision of those estimates. In general, the higher the level of replication, the greater is our confidence in our results. But when we inflate the number of data points by taking repeated measurements from each independent unit, the data points are not independent. If we then analyze the non-

independent data points as if they were an independent sample, we are making a false claim about the amount of replication, hence the “pseudo-” in *pseudoreplication*.<sup>1</sup>

Pseudoreplication can occur in many forms, which can make it challenging to detect. For example, imagine an experiment to test the effects of temperature on plant growth, conducted in two growth chambers. One of the chambers is set to a high temperature, and the other is set to a lower level. This study design is a worst-case scenario because it includes no replication of the “low” and “high” treatments at all. All of the plants in a given treatment share whatever effects the individual chamber causes on growth—over and above the effects of temperature. This design leads to pseudoreplication if the multiple subjects within each chamber are measured and then analyzed as though they represented an independent sample of treatment subjects. The chamber, and not the subject, is the independent unit. The same mistake would be made by a researcher who designs a study to compare the cell division rates of cancer cells and healthy cells by measuring multiple cells from a single cell line of cancer cells and multiple cells from a single line of healthy cells. Pretending that the multiple measurements of cells from a single line are independent in analysis of the data would be pseudoreplication. Cells from the same cell line are not independent of each other, and they do not represent a random sample of all possible cancer cells. A well-designed experiment would compare cells from multiple cancerous and healthy cell lines and treat the cell line as the independent unit.

Most statistical techniques, including almost everything in this book, assume that each data point is independent of the others.

Independence is, after all, part of the definition of a random sample.

If two data points are not independent, treating them as independent assumes that we have more information than we really do. As a result, we would calculate confidence intervals that were too narrow and P-values *P-values* (see [Section 6.2](#)) that were too small.

Pseudoreplication is often subtle, and it remains a major source of mistakes in the analysis of experiments. The rate of pseudoreplication has been estimated to be about one in every eight field studies in ecology ([Hurlbert and White 1993](#); [Heffner et al. 1996](#)) and over 40% for animal experiments ([Lazic et al. 2018](#))!

When reading the scientific literature, keep in mind the possibility of pseudoreplication. Be on the lookout for features that group individual data points during the sampling process. Watch out for multiple measurements taken on the same individuals or the same experimental unit. If the number of measurements, and not the number of independent individuals, is counted as the sample size in a statistical analysis, there could be a problem. Also, when reporting results from your own studies, always report the sample sizes and the degrees of freedom for every analysis. That way, others can see that you are reporting results from analyses that used the correct number of independent units.