
MODULE 12

SAMPLING DISTRIBUTIONS

Objectives:

1. Describe the concept of sampling variability.
2. Describe why sampling variability must be dealt with to make inferences.
3. Describe what a sampling distribution represents.
4. Identify how a sampling distribution differs from a population distribution.
5. Describe what a standard error is.
6. Identify how a standard error differs from a standard deviation.
7. Describe how and why sampling distributions are simulated.
8. Explain the concepts of precision, accuracy, and bias as it relates to statistics and parameters.
9. Describe the theoretical distribution of the sampling distribution of the sample means.
10. Gain some belief that the theoretical distribution actually represents the sampling distribution of the sample means.
11. Use the sampling distribution of sample means to compute the probability of particular sets of means.

Contents

12.1 Definition and Characteristics	147
12.2 Simulating	153
12.3 Central Limit Theorem	154
12.4 Probability Calculations	160
12.5 Accuracy and Precision	164

STATISTICAL INFERENCE IS THE PROCESS of making a conclusion about the parameter of a population based on the statistic computed from a sample. This process is difficult because statistics are random variables (i.e., the exact value of the statistic depends on the individuals in the sample from which it was computed). For example, recall from Section 2.1 that the mean length of fish differed among the four samples of fish “taken” from Square Lake. Thus, to make conclusions about the population from the sample, the distribution of the statistic computed from all possible samples must be understood. In other words, to adequately consider sampling variability when making inferences, the shape, center, and dispersion of the statistic among samples must be understood.¹ In this module, the distribution of statistics from all possible samples is explored and generalizations used to make inferences are identified. In subsequent modules, this information along with results from a single sample, will be used to make specific inferences about the population.

◇ Making statistical inferences requires a consideration of sampling variability.

12.1 Definition and Characteristics

A **Sampling distribution** is the distribution of the values of a particular statistic computed from all possible samples of the same size from the same population. The discussion of sampling distributions and all subsequent theories related to statistical inference are based on repeated samples from the same population. As these theories are developed, we will consider taking multiple samples; however, after the theories have been developed, then only one sample will be taken with the theory then being applied to those results. Thus, it is important to note that only one sample is ever actually taken from a population.

△ **Sampling Distribution:** The distribution of the values of a particular statistic computed from all possible samples of the same size from the same population.

Actual sampling distributions can only be computed for very small populations.² Thus, to illustrate the concept of a sampling distribution, consider a population of six students that have scored 6, 6, 4, 5, 7 and 8 points, respectively, on an 8-point quiz. The mean of this population is $\mu = 6.000$ points and the standard deviation is $\sigma = 1.414$ points. Suppose that every sample of size $n = 2$ is extracted from this population and the sample mean is computed for each sample (Table 12.1).³ The histogram of these 15 means is the sampling distribution of the sample mean from samples of $n = 2$ from this population (Figure 12.1).⁴

Table 12.1. All possible samples of $n = 2$ and the corresponding sample mean from the simple population of quiz scores.

Scores	Mean	Scores	Mean	Scores	Mean	Scores	Mean	Scores	Mean
6,6	6.0	6,7	6.5	6,5	5.5	4,5	4.5	5,7	6.0
6,4	5.0	6,8	7	6,7	6.5	4,7	5.5	5,8	6.5
6,5	5.5	6,4	5	6,8	7.0	4,8	6.0	7,8	7.5

¹See Module 1 for a review of sampling variability.

²See Section 12.2 for how sampling distributions for larger populations are simulated.

³These samples are found by putting the values into a vector with `vals <- c(6,6,4,5,7,8)` and then using `combn(vals,2)`. The means are found with `mns <- as.numeric(combn(vals,2,mean))`.

⁴The histogram is constructed with `hist(~mns,w=0.5)`.

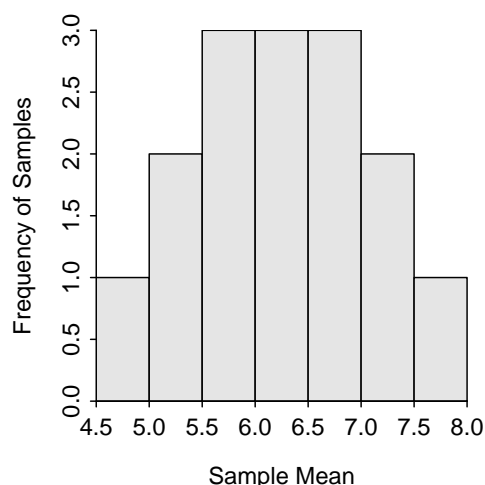


Figure 12.1. Sampling distribution of mean quiz scores from samples of $n = 2$ from the simple population of quiz scores.

The mean and standard deviation of the 15 sample means are measures of center and dispersion for the sampling distribution. The mean and standard deviation of the 15 sample means are 6.000 and 0.845, respectively. The standard deviation of the statistics (i.e., the dispersion of the sampling distribution) is generally referred to as the **standard error of the statistic** (abbreviated as SE_{stat}). This new terminology is used to help keep the dispersion of the sampling distribution separate from the dispersion of the individuals in the population, which is measured by the standard deviation. Thus, the standard deviation of all possible sample means is generally referred to as the standard error of the sample means, or SE. Thus, the SE in this example is 0.845.

△ Standard Error: The numerical measure of dispersion used for sampling distributions – i.e., measures the dispersion among statistics from all possible samples.

This simple example illustrates three major concepts concerning sampling distributions. First, the sampling distribution of the statistic will more closely resemble a normal distribution than the original population distribution (unless, of course, the population distribution was normal).

Second, the center (i.e., mean) of the sampling distribution of a statistic will equal the parameter that the statistic was intended to estimate (e.g., a sample mean is intended to be an estimate of the population mean). In this example, the mean of all possible sample means (= 6.0 points) is equal to the mean of the original population ($\mu = 6.0$ points). A statistic is said to be **unbiased** if the center (mean) of its sampling distribution equals the parameter it was intended to estimate. This example illustrates that the sample mean is an unbiased estimator of the population mean.

△ Unbiased Statistic: A statistic in which the center of its sampling distribution equals the parameter it is intended to estimate.

◇ All statistics in this course are unbiased.

Third, the standard error of the statistic is less than the standard deviation of the original population. In

other words, the dispersion of statistics is less than the dispersion of individuals in the population. For example, the dispersion of individuals in the population is $\sigma = 1.414$ points, whereas the dispersion of statistics from all possible samples is $SE_{\bar{x}} = 0.845$ points.

◇ The sampling distribution will be more normal than the original population distribution.

◇ The mean of the statistics in a sampling distribution will (generally) equal the parameter that the statistic was intended to estimate.

◇ The dispersion of the sampling distribution will be less than the dispersion of the original population distribution.

Review Exercises

12.1 Use the simple population of quiz scores from the previous section (i.e., 6, 6, 4, 5, 7, and 8) to answer the questions below. [Answer](#)

- Construct a table similar to Table 12.1 that shows the values and the mean of those values for all possible samples of size $n = 4$. Note: there are 15 such samples.
- Construct a histogram of the means from all possible samples. Describe its general shape.
- Compute the mean of the means from all possible samples. How does this compare to the mean of all six individuals in the population?
- Compute the standard error of the means from all possible samples. How does this compare to the standard deviation of all six individuals in the population? How does this compare to the standard error of the means of all possible samples of $n = 2$ shown in Table 12.1 and for all possible samples of $n = 3$ shown in Table 12.2 (later in this module)? Can you make a general statement about how the standard error of the means is related to the size of the sample used to construct the means?

12.2 Suppose the individuals in a simple population have the following “values” for a simple binomial categorical variable – Y, Y, N, Y, Y, N, and N. Use this to answer the questions below. [Answer](#)

- Construct a table similar to Table 12.1 that shows the “values” of the individuals and the proportion of “yeses” for all possible samples of size $n = 3$. Note: there are 35 such samples.
- Construct a histogram of the proportions from all possible samples. Describe its general shape.
- Construct the mean of the proportions from all possible samples. How does this compare to the proportion of “yeses” for all seven individuals in the population?
- Construct the standard error of the proportions from all possible samples.

12.1.1 Critical Distinction

Three distributions are considered in statistics. The sampling distribution is the distribution of a statistic computed from all possible samples of the same size from the same population., the population distribution is the distribution of all individuals in a population (see Module 7), and the sample distribution is the

distribution of all individuals in a sample (see histograms in Module 5). The sampling distribution is about **statistics**, whereas the population and sample distributions are about **individuals**. For inferential statistics, it is important to distinguish between the population and sampling distributions. Keep in mind that one (population) is the distribution of individuals and the other (sampling) is the distribution of statistics.

Just as importantly, remember that a standard error measures the dispersion among statistics (i.e., sampling variability), whereas a standard deviation measures dispersion among individuals (i.e., natural variability). Specifically, the population standard deviation measures dispersion among all individuals in the population and the sample standard deviation measures the dispersion of all individuals in a sample. In contrast, the standard error measures the dispersion among statistics computed from all possible samples. The population standard deviation is the dispersion on a population distribution, whereas the standard error is the dispersion on a sampling distribution.

◇ Sampling distributions represent the distribution of statistics from all possible samples, whereas population distributions represent the distribution of all individuals in a population.

◇ Standard error measures dispersion among statistics, whereas standard deviation measures dispersion among individuals.

◇ Standard error measures sampling variability, whereas the standard deviation measures natural variability.

Review Exercises

- 12.3 What type of distribution is blood serum level for every individual in a population? [Answer](#)
- 12.4 What type of distribution is mean cholesterol level computed from all possible samples of $n = 15$ patients for a clinic? [Answer](#)
- 12.5 What type of distribution is water discharge amounts for Bay City Creek for every day in 2005 assuming that all days in 2005 was the population of interest? [Answer](#)
- 12.6 What type of distribution is water discharge amounts for Bay City Creek for every day in 2005 if the population of interest is all days in the 21st century? [Answer](#)
- 12.7 What type of distribution is the proportion of days where the water discharge from Bay City Creek is near negligible calculated from all samples of $n = 30$ days. [Answer](#)
- 12.8 On average, the mean length of $n = 30$ cicadas is 2.9 mm away from the overall average. Is this a standard deviation or a standard error? [Answer](#)

12.9 On average, the number of litter items found along the Escarpment Trail in the Porcupine Mountains on a single day is 12 items different than the overall mean. Is this a standard deviation or a standard error?

[Answer](#)

12.1.2 Dependencies

The sampling distribution of sample means from samples of $n = 2$ from the population of quizzes was shown above. The sampling distribution will look different if any other sample size is used. For example, the samples and means from each sample of $n = 3$ are shown in Table 12.2. The mean of these means is 6.000, the standard error is 0.592, and the sampling distribution is symmetric, perhaps approximately normal (Figure 12.2). The three major characteristics of sampling distributions noted in Section 12.1 are still true: the sampling distribution is still more normal than the original population, the sample mean is still unbiased (i.e., the mean of the means is equal to μ), and the standard error is smaller than the standard deviation of the original population. However, also take note that the standard error of the sample mean is smaller from samples of $n = 3$ than from $n = 2$.⁵

Table 12.2. All possible samples of $n = 3$ and the corresponding sample means from the simple population of quiz scores.

Scores	Mean	Scores	Mean	Scores	Mean	Scores	Mean	Scores	Mean
6,6,4	5.3	6,6,5	5.7	6,6,7	6.3	6,6,8	6.7	4,5,7	5.3
6,4,5	5.0	6,4,7	5.7	6,4,8	6.0	6,5,7	6.0	4,5,8	5.7
6,5,8	6.3	6,7,8	7.0	6,4,5	5.0	6,4,7	5.7	4,7,8	6.3
6,4,8	6.0	6,5,7	6.0	6,5,8	6.3	6,7,8	7.0	5,7,8	6.7

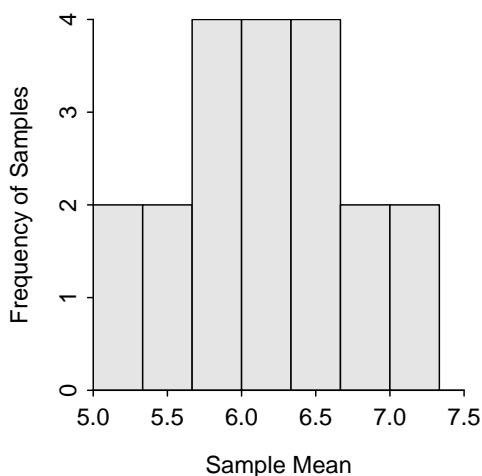


Figure 12.2. Sampling distribution of mean quiz scores from samples of $n = 3$ from the simple population of quiz scores.

◇ Sampling distributions differ for samples of different sizes. In particular the distribution will be “more” normal and the standard error will be smaller as sample size increases.

⁵One should also look at the results from $n = 4$ in Review Exercise 12.1.

The sampling distribution will also be different if the statistic changes; e.g, if the sample median rather than sample mean is computed in each sample. Before showing the results of each sample, note that the population median (i.e., the median of the individuals in the population — 6, 6, 4, 5, 7, and 8) is 6.0 points. The sample median from each sample is shown in Table 12.3 and the actual sampling distribution is shown in Figure 12.3. Note that the sampling distribution of the sample medians is still “more” normal than the original population distribution, the mean of the sample medians (=6.000 points) still equals the parameter (population median) that the sample median is intended to estimate (thus the sample median is also unbiased), and this sampling distribution differs from the sampling distribution of sample means from samples of $n = 3$.

Table 12.3. All possible samples of $n = 3$ and the corresponding sample medians from the simple population of quiz scores.

Scores	Median	Scores	Median	Scores	Median	Scores	Median	Scores	Median
6,6,4	6	6,6,5	6	6,6,7	6	6,6,8	6	4,5,7	5
6,4,5	5	6,4,7	6	6,4,8	6	6,5,7	6	4,5,8	5
6,5,8	6	6,7,8	7	6,4,5	5	6,4,7	6	4,7,8	7
6,4,8	6	6,5,7	6	6,5,8	6	6,7,8	7	5,7,8	7

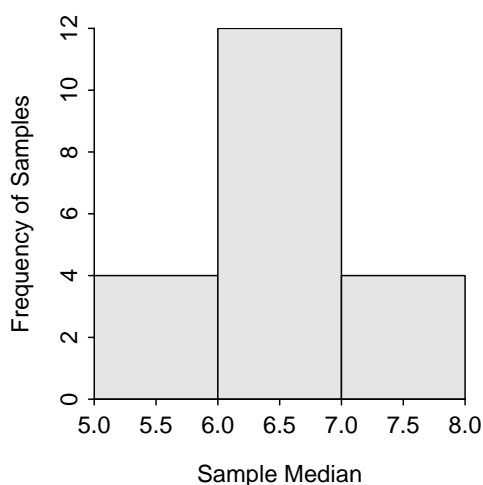


Figure 12.3. Sampling distribution of median quiz scores from samples of $n = 3$ from the simple population of quiz scores.

◇ Sampling distributions for different statistics are different.

These examples demonstrate that the naming of a sampling distribution must be specific. For example, the first sampling distribution in this module should be described as the “sampling distribution of sample means from samples of $n=2$.” This last example should be described as the “sampling distribution of sample medians from samples of $n=3$.” Doing this with each distribution reinforces the point that sampling distributions depend on the sample size and the statistic calculated.

◇ Each sampling distribution should be specifically labeled with the statistic calculated and the sample size of the samples.

12.2 Simulating

In Section 12.1, exact sampling distributions were computed for very small samples taken from a small population. Exact sampling distributions are difficult to show for even moderate sample sizes from moderately-sized populations. For example, there are 15504 unique samples of $n = 5$ from a population of 20 individuals. How are sampling distributions examined in these larger cases?

There are two ways to examine sampling distributions in situations with large sample and population sizes. First, the computer can take many (hundreds or thousands) samples and compute the statistic for each. These statistics can then be summarized to give an indication of what the actual sampling distribution would look like. This process is called “simulating a sampling distribution” and is the subject of this section. Second, theorems exist that describe the specifics of sampling distributions under certain conditions. One such theorem is described in Section 12.3. These theorems will be relied upon in subsequent modules.

◇ The approximate shape of sampling distributions from large samples or large populations can be obtained from (1) theorems or (2) computer simulations.

Sampling distributions are simulated by drawing many samples from a population, computing the statistic of interest for each sample, and constructing a histogram of these statistics (Figure 12.4). The computer is helpful with this simulation; however, keep in mind that the computer is basically following the same process as used in Section 12.1, with the exception that not every sample is taken.

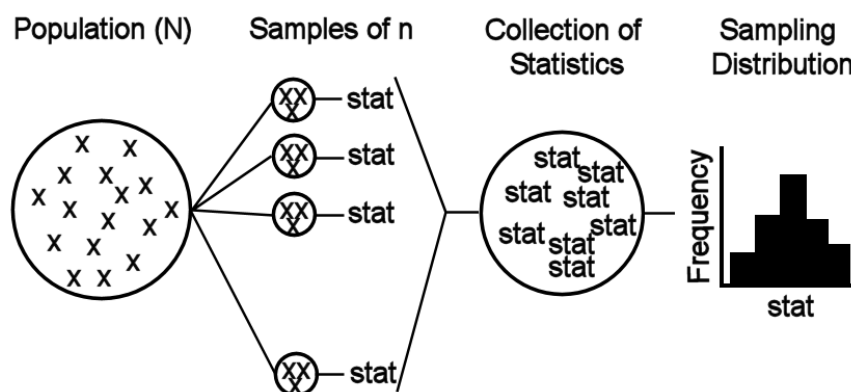


Figure 12.4. Schematic representation of the process for simulating sampling distributions.

◇ Sampling distributions can be simulated by drawing many samples from a population, computing the statistic of interest for each sample, and constructing a histogram of the values of the statistic.

To illustrate the simulation of a sampling distribution, let's return to the Square Lake fish population explored in Section 2.1. Recall that this is a hypothetical population with 1015 fish, a population distribution shown in Figure 2.1, and parameters shown in Table 2.1. Further recall that four samples of $n = 50$ were removed from this population and summarized in Table 2.2 and Table 2.3. Suppose, that an additional 996 samples of $n = 50$ were extracted in exactly the same way as the first four, the sample mean was computed in each sample, and the 1000 sample means were collected to form the histogram in Figure 12.5. This histogram is a simulated sampling distribution of sample means because it represents the distribution of sample means from 1000, rather than all possible, samples.

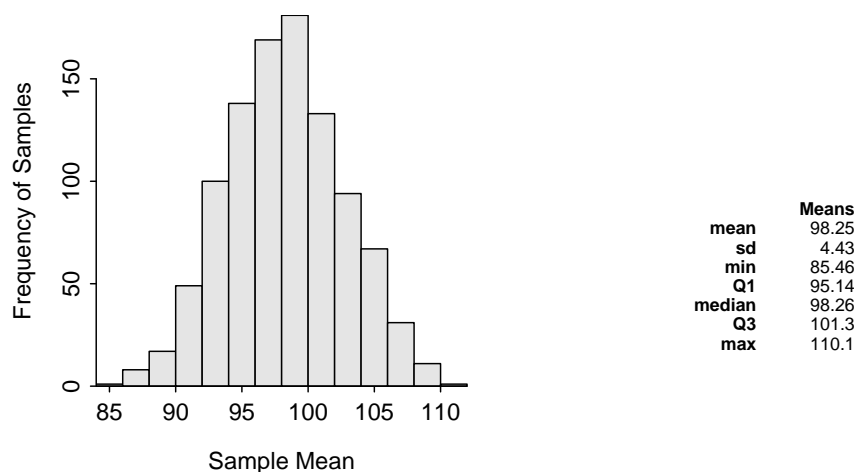


Figure 12.5. Histogram (**Left**) and summary statistics (**Right**) from 1000 sample mean total lengths computed from samples of $n = 50$ from the Square Lake fish population.

As with the actual sampling distributions discussed previously, three characteristics (shape, center, and dispersion) are examined with simulated sampling distributions. First, this sampling distribution looks at least approximately normally distributed. Second, the mean of the 1000 means in the sampling distribution ($=98.25$) is approximately equal to the mean of the original 1015 fish in Square Lake ($=98.06$). These two values are not exactly the same because the simulated sampling distribution was constructed from only a “few” samples rather than all possible samples. Third, the standard error of the sample means ($=4.43$) is much less than the standard deviation of individuals in the original population ($=31.49$). So, within reasonable approximation, the concepts identified with actual sampling distributions also appear to hold for simulated sampling distributions.

As before, computing a different statistic on each sample results in a different sampling distribution. This is illustrated by comparing the sampling distributions of a variety of statistics from the same 1000 samples of size $n=50$ taken above (Figure 12.6).

Simulating a sampling distribution by taking many samples of the same size from a population is powerful for two reasons. First, it reinforces the ideas of sampling variability – i.e., each sample results in a slightly different statistic. Second, the entire concept of inferential statistics is based on theoretical sampling distributions. Simulating sampling distributions will allow us to check this theory and better visualize the theoretical concepts. From this module forward, though, remember that sampling distributions are simulated primarily as a check of theoretical concepts. In real-life, only one sample is taken from the population and the theory is used to identify the specifics of the sampling distribution.

◇ Simulating sampling distributions is a tool for checking the theory concerning sampling distributions; however, in “real-life” only one sample from the population is needed.

12.3 Central Limit Theorem

The sampling distribution of the sample mean was examined in the previous sections by taking all possible samples from a small population (Section 12.1) or taking a large number of samples from a large population

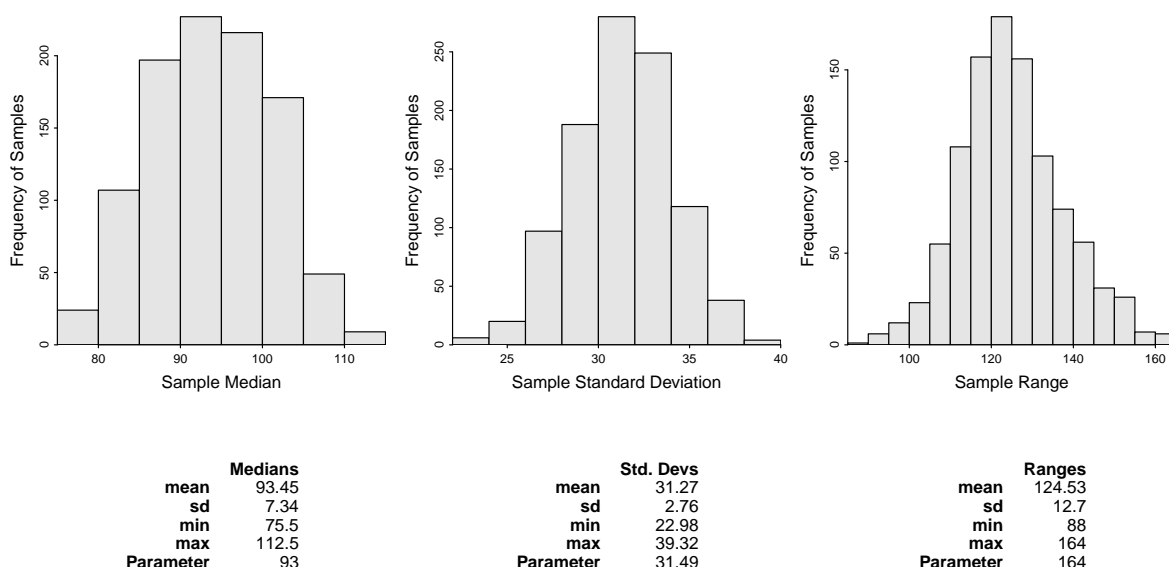


Figure 12.6. Histograms from 1000 sample median (**Left**), standard deviation (**Center**), and range (**Right**) of total lengths computed from samples of $n = 50$ from the Square Lake fish population. Note that the value in the parameter row is the value computed from the entire population.

(Section 12.2). In both instances, it was observed that the sampling distribution of the sample mean was approximately normally distributed, centered on the true mean of the population, and had a standard error that was smaller than the standard deviation of the population and decreased as n increased. In this section, the Central Limit Theorem (CLT) is introduced and explored as a method for identifying the specific characteristics of the sampling distribution of the sample mean without going through the process of extracting multiple samples from the population.

The CLT specifically addresses the shape, center, and dispersion of the sampling distribution of the sample means by stating that $\bar{x} \sim N\left(\mu, \frac{\sigma}{\sqrt{n}}\right)$ as long as

- $n \geq 30$,
- $n \geq 15$ and the population distribution is not strongly skewed, **or**
- the population distribution is normally distributed.

Thus, the sampling distribution of \bar{x} should be normally distributed **no matter what the shape of the population distribution is** as long as $n \geq 30$. The CLT also suggests that \bar{x} is unbiased and that the formula for the $SE_{\bar{x}}$ is $\frac{\sigma}{\sqrt{n}}$ regardless of the size of n . In other words, n impacts the shape of the sampling distribution of the sample means, but not the center or formula for computing the standard error.

Δ Central Limit Theorem: If a variable x has a population distribution with a mean, μ , and a standard deviation, σ , then the sampling distribution of the sample means (\bar{x}) from random samples of size n , will have a mean equal to μ , a standard error equal to $\frac{\sigma}{\sqrt{n}}$, and a shape that will tend to be normal as n becomes “large.”

12.3.1 Exploring CLT

The validity of the CLT can be examined by again simulating several (with different n) sampling distributions of \bar{x} from the Square Lake population (Figure 12.7). Recall from Section 2.1 that the population distribution (Figure 2.1) and several parameters (Table 2.1) are known and the sampling distribution from $n = 50$ is in Figure 12.5.

Several observations about the CLT can be made from the results in Figure 12.7. First, the sampling distribution is approximately normal even for very small n because the population distribution is only slightly skewed (Figure 2.1). If the population distribution had been strongly skewed, then the sampling distributions would only approximate normality for larger n (see next paragraph). Second, the means of all sampling distributions are approximately equal to $\mu = 98.06$, regardless of n . Third, the dispersion of the sampling distributions (i.e., the SE of the means) becomes smaller with increasing n . Furthermore, the SE from the simulated results closely match the SE expected from the CLT (i.e., $\frac{34.19}{\sqrt{n}}$).

To illustrate that the CLT is not true just for the Square Lake population, similar results from uniform (i.e., rectangular) and strongly right-skewed population distributions are in Figures 12.8 and 12.9, respectively. For each figure, note how (1) each distribution becomes more “normal” as n increases, (2) the sampling distributions from the uniform distribution become normal at smaller n , (3) each sampling distribution remains centered on approximately the same value for all n (approximately 0.5 for the uniform and 1 for the skewed population distributions), (4) each sampling distribution becomes narrower as n increases (i.e., SE gets smaller), and (5) the observed SE is approximately equal to the SE expected from the CLT.

Review Exercises

- 12.10** Assume that the population distribution is $\sim N(100, 20)$ and you take samples of $n = 50$. [Answer](#)
- (a) What shape would you expect the sampling distribution of the sample means to be?
 - (b) What do you expect the center of the sampling distribution of the sample means to equal?
 - (c) What do you expect the standard deviation of the sampling distribution of the sample means to equal?
 - (d) What do you expect the standard error of \bar{x} to equal?
- 12.11** Assume that the population distribution is skewed to the right with $\mu = 500$ and $\sigma = 60$. Further suppose that samples of $n = 100$ are taken. [Answer](#)
- (a) What shape would you expect the sampling distribution of the sample means to be?
 - (b) What do you expect the center of the sampling distribution of the sample means to equal?
 - (c) What do you expect the standard deviation of the sampling distribution of the means to equal?
 - (d) What do you expect the standard error of \bar{x} to equal?
- 12.12** Assume that the population distribution is slightly skewed to the right with $\mu = 500$ and $\sigma = 60$. Further suppose that samples of $n = 20$ are taken. [Answer](#)
- (a) What shape would you expect the sampling distribution of the sample means to be?
 - (b) What do you expect the center of the sampling distribution of the sample means to equal?
 - (c) What do you expect the standard deviation of the sampling distribution of the means to equal?
 - (d) What do you expect the standard error of \bar{x} to equal?

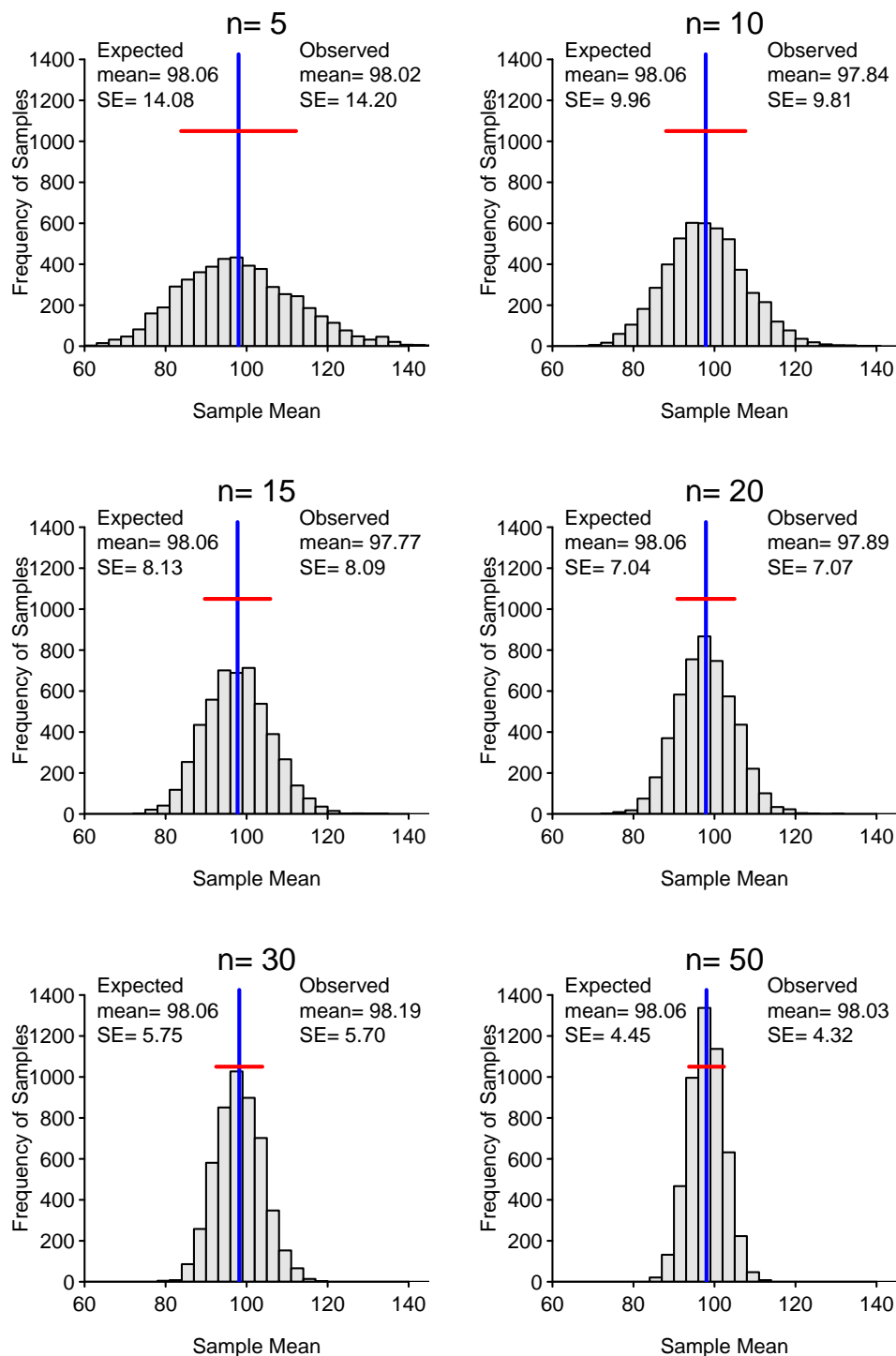


Figure 12.7. Sampling distribution of the sample mean TL simulated from 5000 samples of six different sample sizes extracted from the Square Lake fish population. The vertical blue line is the mean of the 5000 means and the horizontal red line represents $\pm 1SE$ from the mean.

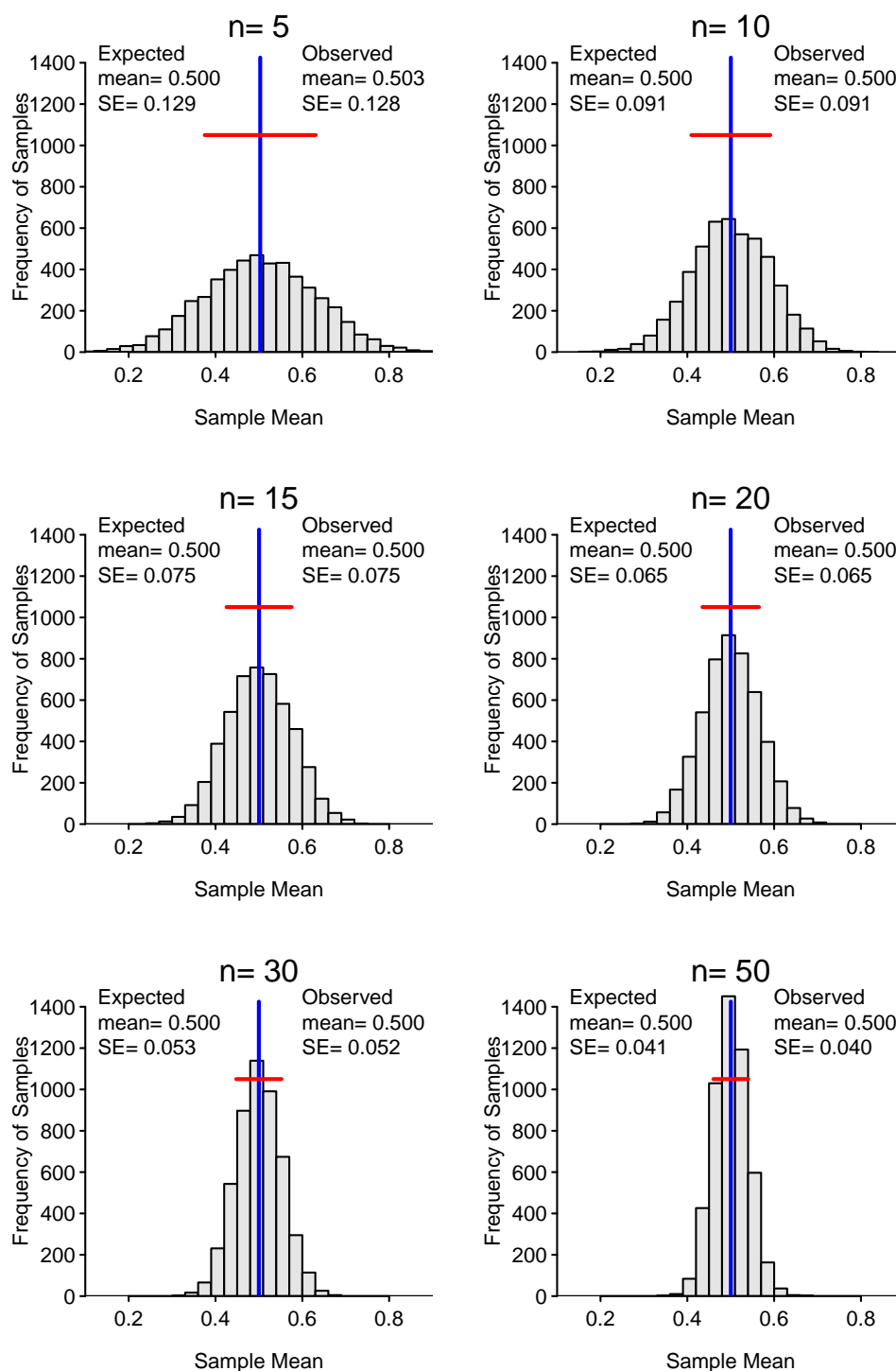


Figure 12.8. Sampling distribution of the sample mean simulated from 5000 samples of six different sample sizes extracted from a uniform population distribution. The vertical blue line is the mean of the 5000 means and the horizontal red line represents $\pm 1SE$ from the mean.

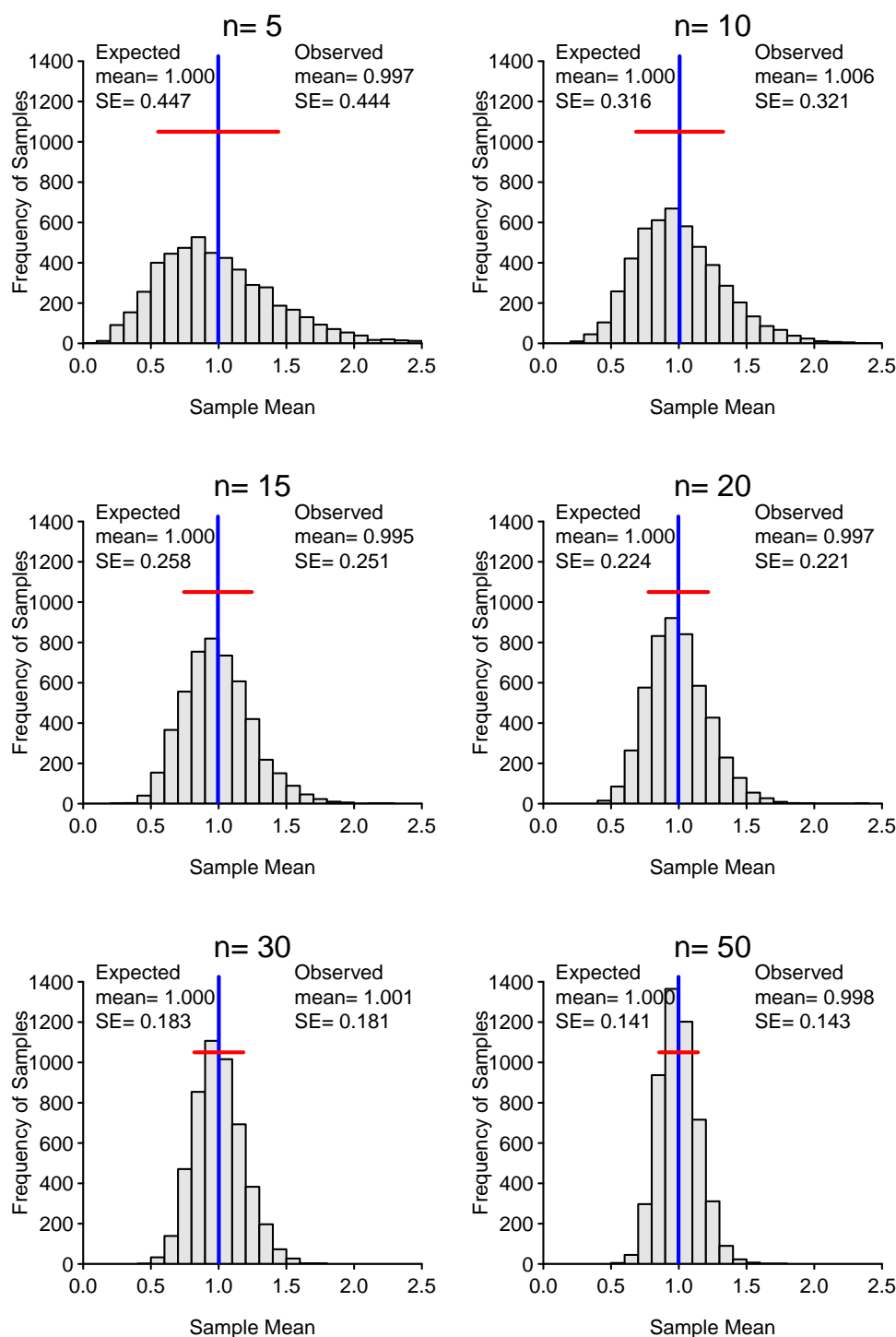


Figure 12.9. Sampling distribution of the sample mean simulated from 5000 samples of six different sample sizes extracted from an exponential population distribution ($\lambda = 1$). The vertical blue line is the mean of the 5000 means and the horizontal red line represents ± 1 SE from the mean.

12.4 Probability Calculations

If the sample size is large enough, then the CLT states that the sampling distribution of sample means is approximately normally distributed. If the sampling distribution is normal, then the methods from Module 7 may be used to compute probabilities. Thus, if the sampling distribution of the sample means is normally distributed, then questions such as “what is the probability of observing a sample mean of less than 95 mm from a sample of $n = 50$ from Square Lake?” can be answered. In other words, questions related to the probability of **statistics** can be answered.

The question above is answered by first recalling that, for the length of all fish in Square Lake, $\mu = 98.06$ and $\sigma = 31.49$. Because $n = 50$ is greater than 30, the CLT says that the distribution of the sample means from samples of $n = 50$ is $\bar{x} \sim N(98.06, \frac{31.49}{\sqrt{50}})$ or $\bar{x} \sim N(98.06, 4.835)$. Thus, the proportion of samples of $n = 50$ from Square Lake with an $\bar{x} < 95$ mm is 0.2634, which comes from computing the area less than 95 on a $N(98.06, 4.835)$ distribution (Figure 12.10).⁶

```
> ( distrib(95,mean=98.06,sd=31.49/sqrt(50)) )
[1] 0.2634127
```

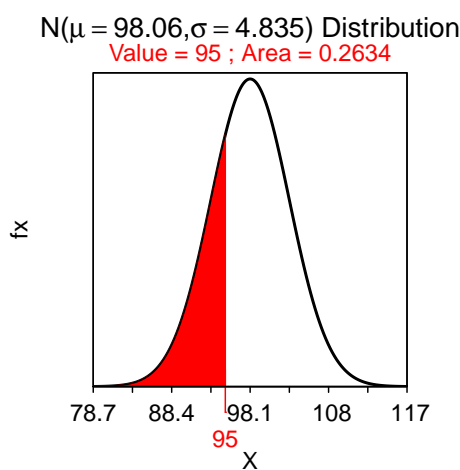


Figure 12.10. Proportion of sample means less than 95 mm on a $N(98.06, 4.84)$ distribution.

◇ Calculating the probability of a set of means is as simple as computing areas on a normal distribution as long as the assumptions of the CLT hold true (i.e., n is large enough).

Consider another question – “what is the probability of observing a sample mean of more than 95 mm in a sample of $n = 40$ from Square Lake?” At first glance it may appear that this question can be answered from the work done for the previous question. However, the sample sizes differ between the two questions and, because the sampling distribution depends on the sample size, a different sampling distribution is used here. Because $n > 30$ the sampling distribution will be $\bar{x} \sim N(98.06, \frac{31.49}{\sqrt{40}})$ or $\bar{x} \sim N(98.06, 5.406)$ (Note the different value of the SE). Thus, the answer to this question is the area to the right of 95 on a $N(98.06, 5.406)$ or 0.7143 (Figure 12.11).

⁶Notice that the standard error of \bar{x} is put into the `sd=` argument of `distrib()`. Recall that a standard error really is a standard deviation, it is just named differently (see Section 12.1). R has no way of knowing whether the question is about an individual or a statistic; it requires the dispersion in either case and calls both of them `sd=`.

```
> ( distrib(95,mean=98.06,sd=34.19/sqrt(40),lower.tail=FALSE) )
[1] 0.714319
```

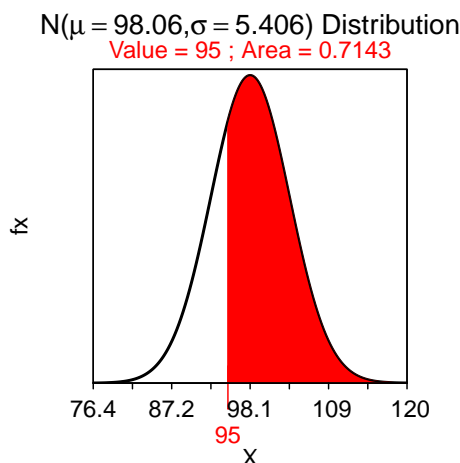


Figure 12.11. Proportion of sample means greater than 95 mm on a $N(98.06, 5, 41)$ distribution.

◇ Always check what sample size is being used – if the sample size changes, then the sampling distribution changes.

Consider two more Square Lake example questions. First, “what is the probability of observing a sample mean of more than 95 mm in a sample of $n = 10$ from Square Lake?” This question is again about a statistic, but because $n < 15$ and the population is not known to be normal it is not known that the sampling distribution will be normal. Thus, this question cannot be answered. Second, “What is the probability that a fish will have a length less than 85 mm?” This question is about an individual, not a statistic as in the previous questions. Thus, the population distribution, NOT the sampling distribution, is appropriate here. However, this question also cannot be answered because the population distribution is not known to be normally distributed.

Two points are illustrated with the last two questions. First, population distributions are used for questions about individuals and sampling distributions are used for questions about statistics. Second, if the distribution is not known to be normal, no matter which distribution is used, then the probability cannot be computed.⁷


◇ If the question refers to individuals, then use the population distribution. If the question refers to a statistic, then use a sampling distribution.

◇ If the distribution needed to answer a question is not normal, then normal distribution calculations cannot be used to answer the question. The proper answer to the question in this case is to say “I cannot compute the probability because the required distribution is not known to be normal.”


⁷At least with the techniques in this course.

One issue you may have noticed is that these calculations require knowing the mean, standard deviation, and shape (if $n < 30$) of the population. However, the population usually cannot be “seen” (recall Module 1) and, thus, it is uncomfortable to assume so much to be known about the population. The only appropriate response to this concern is that we are building towards being able to make inferences with statements based on probabilities that take into account sampling variability. To make these probabilistic statements we need to fully understand sampling distributions. These questions, while not yet realistic, will help you to better understand sampling distributions for when they are needed to make inferences in later modules.


Review Exercises

12.13  Assume that it is known that the distribution of time spent hunting (hours) by an individual Minnesota moose (*Alces alces*) hunter is approximately symmetric in shape with a mean of 40 hours and a standard deviation of 15 hours. Use this information to answer the questions below. [Answer](#)


- Describe what an individual is in this problem.
- List the variable or variables in this problem and identify the type of variable for each.
- What is the probability that a hunter will spend more than 55 hrs hunting moose?
- What is the probability that the average hours spent hunting by a sample of 25 hunters is greater than 48 hrs?

12.14  Facilities management is interested in the mean relative weight (= actual weight / predicted weight; W_r) of fish in the portion of Bay City Creek that runs through the Northland campus. For each question below assume that W_r for fish in the population is $\sim N(1, 0.2)$. [Answer](#)


- What is the population of interest (be very specific)?
- What is the parameter of interest?
- What is the value of the parameter of interest?
- What statistic should be computed to estimate this parameter?
- We can take a random sample of either 25 or 36 fish. Which sample, if either, would tend to produce the most accurate statistic? Why?
- Which sample ($n = 25$ or $= 36$), if either, would tend to produce the most precise statistic? Why?
- What is the exact distribution of the statistic for the n you chose to produce the most precise estimate?
- A mean W_r under 0.95 is indicative of a stressed population. What is the probability of observing a mean W_r that is indicative of a stressed population in Bay City Creek? Use your chosen sample size (here and in the next two questions).
- What are the lower and upper bounds for the most common 95% of W_r values?
- What is the range for the most common 90% of mean W_r values?

12.15  The WI Department of Natural Resources is examining the amount of domestic corn consumed by raccoons per week. Assume that the amount eaten is slightly right-skewed, with a mean of 8 kg, and a standard deviation of 2 kg. [Answer](#)


- What is the probability that a raccoon consumes more than 13 kg per week?
- What is the probability that a sample of 25 raccoons have a mean corn consumption of more than 10 kg per week?
- What is the probability that a sample of 60 raccoons have a TOTAL corn consumption of more than 510 kg per week?

12.16  Suppose that it is known that the number of yards gained per game for the primary running back on a National Football League team is slightly left-skewed with a mean of 82 yards and a standard deviation of 26 yards. [Answer](#)


- (a) What is the probability that a running back will gain more than 100 yards in a single game?
- (b) What is the probability that a running back will average more than 100 yards per game in a 16-game season?
- (c) What is the probability that a running back will average between 70 and 90 yards per game in a 16-game season?
- (d) What is the probability that a running back will average more than 70 yards per game over two 16-game seasons?
- (e) What is the top 25% of yards gained by a running back in a single game?
- (f) What is the top 5% of mean yards gained by a running back in a 16-game season?

12.17  Suppose that the average annual rate of return for a wide array of available stocks is approximately normally distributed with a mean of 4.2% with a standard deviation of 4.9%. [Answer](#)


- (a) What is the probability that five randomly selected stocks produce a positive average rate of return?
- (b) What is the probability that a randomly selected stock produces a positive rate of return?
- (c) What is the probability that ten randomly selected stocks produce a less than 2% average rate of return?
- (d) The top 10% of stocks produce what rate of return?
- (e) The top 10% of random samples of 10 stocks produce what average rate of return?

12.18  Renner (1970) examined the content of hydroxymethylfurfural (HMF) in honey. HMF is an organic compound derived from cellulose without the use of fermentation and is a potential “carbon-neutral” source for fuels. This study found that the distribution of HMF in honey was extremely strongly right-skewed with a mean of 9.5 g/kg and a standard deviation of 13.5 g/kg. [Answer](#)

- (a) What is the probability that one kg of honey have more than 20 g of HMF?
- (b) What is the probability that 20 samples of one kg of honey have an average of more than 20 g of HMF?
- (c) What is the probability that 50 samples of one kg of honey have an average of less than 10 g of HMF?
- (d) What are the 20% least common average amounts of HMF in 50 samples of one kg of honey?

12.19  Allanson (1992) examined the size of farms in England in 1939 and 1989. He found the distribution of farm sizes in 1989 to be very right-skewed with a mean of 65.13 ha and a standard deviation of 108.71 ha. [Answer](#)

- (a) What are the 10% most common sizes of farms in England?
- (b) What are the 10% most common average sizes in samples of 60 farms from England?
- (c) What is the probability that the average size of 60 farms from England is less than 50 ha?
- (d) What is the probability that a farm from England is greater than 50 ha?

12.20  Janzen and Morjan (2002) examined the size of male and female painted turtles (*Chrysemys picta*) at hatching. They found in a sample of 77 turtles that size at hatching was very slightly right-skewed with a mean of 4.46 g with a standard deviation of 0.13 g. Assume that the results of this sample extend to the population to answer the questions below. [Answer](#)

- (a) What is the probability that a turtle will hatch in more than 7 days?
- (b) What is the probability that a sample of 20 turtles will have an average number of days until hatching that is greater than 4.5 days?

- (c) What is the probability that a sample of 50 turtles will have an average number of days until hatching that is greater than 4.5 days?
- (d) What is the mean number of days until hatching such that 20% of samples of 50 turtles have a smaller mean?
- (e) What are the most common 80% of times to hatching?

12.5 Accuracy and Precision

Accuracy and **precision** are often used to describe characteristics of a sampling distribution. Accuracy refers to how closely a statistic estimates the intended parameter. If, **on average**, a statistic is approximately equal to the parameter it was intended to estimate, then the statistic is considered **accurate**. Unbiased statistics are also accurate statistics. Precision refers to the repeatability of a statistic. A statistic is considered to be **precise** if multiple samples produce similar statistics. The standard error is a measure of precision; i.e., a high SE means low precision and a low SE means high precision.

The concepts of accuracy and precision are illustrated in Figure 12.12. The targets in Figure 12.12 provide an intuitive interpretation of accuracy and precision, whereas the sampling distributions (i.e., histograms) are what statisticians look at to identify accuracy and precision. Targets in which the blue plus (i.e., mean of the means) is close to the bullseye are considered accurate (i.e., unbiased). Similarly, sampling distributions where the observed center (i.e., blue vertical line) is very close to the actual parameter (i.e., black tick labeled with a “T”) are considered accurate. Targets in which the red dots are closely clustered are considered precise. Similarly, sampling distributions that exhibit little variability (low dispersion) are considered precise.

△ **Accuracy:** The tendency of a statistic to come close to the parameter it was intended to estimate.

△ **Precision:** The tendency to have values clustered closely together. Precision is inversely related to the standard error – the smaller the standard error, the greater the precision.

Review Exercises

12.21 Suppose that it is known that a population has $\mu=10$. Use this to answer the questions below. [Answer](#)

- (a) Which is more accurate – four samples with means of 9,10,11, and 9 or means of 6,8,7, and 9?
- (b) Which is more accurate – four samples with means of 6,14,8, and 12 or means of 8,7,9, and 8?
- (c) Which is more precise – four samples with means of 7,14,8, and 11 or means of 7,7,9, and 8?
- (d) How would you judge the accuracy and precision of four samples with means of 2,8,12, and 18?
- (e) How would you judge the accuracy and precision of four samples with means of 9,10,11, and 10?
- (f) How would you judge the accuracy and precision of four samples with means of 1,7,8, and 19?

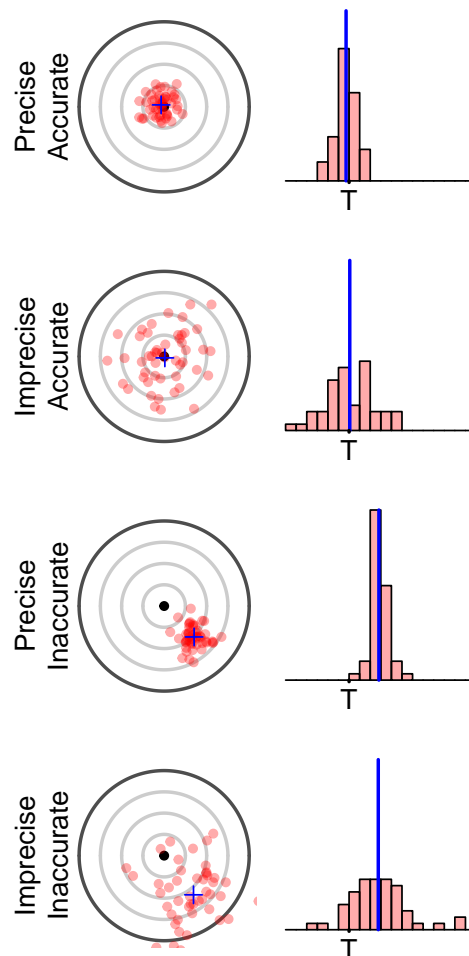


Figure 12.12. Model used to demonstrate accuracy, precision, and bias. The center of each target (i.e., the bullseye) and the point marked with a “T” (for “truth”) represent the parameter of interest. Each dot on the target represents a statistic computed from a single sample and, thus, the many red dots on each target represent repeated samplings from the same population. The center of the samples (analogous to the center of the sampling distribution) is denoted by a blue plus-sign on the target and a blue vertical line on the histogram. The target concept is modified from Ratti and Garton (1994).