MODULE 13

SAMPLING DISTRIBUTIONS

Contents

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 becauses statistics depend on the specific individuals in the sample and, thus, vary from sample to sample. For example, recall from Section 2.1 that the mean length of fish differed among four samples "taken" from Square Lake. Thus, to make conclusions about the population from the sample, the distribution (i.e., shape, center, and dispersion) of the statistic computed from all possible samples must be understood.[1] In this module, the distribution of statistics from all possible samples is explored and generalizations are defined that can be used to make inferences. In subsequent modules, this information, along with results from a single sample, will be used to make specific inferences about the population.

◇ **Statistical inference requires considering sampling variability.**

[1] See Module 1 for a review of sampling variability.

13.1 What is a Sampling Distribution?

13.1.1 Definitions and Characteristics

A **Sampling distribution** is the distribution of 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.

The concept of a sampling distribution is illustrated with a population of six students that 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 that the sample mean is computed for each sample (Table 13.1).[2] The sampling distribution of the sample mean from samples of $n = 2$ from this population (Figure 13.1) is the histogram of means from these 15 samples.[3]

Table 13.1. All possible samples of $n = 2$ and corresponding sample mean from the quiz score population.

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

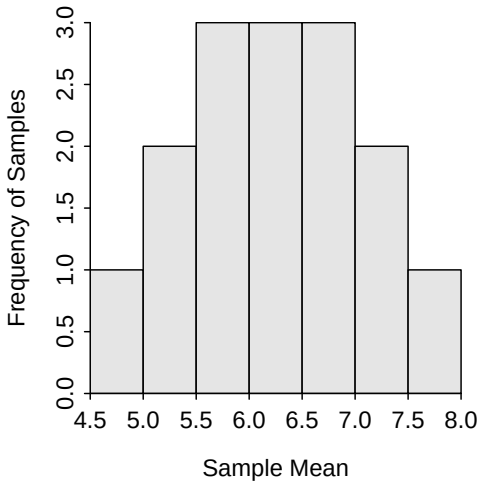


Figure 13.1. Sampling distribution of mean quiz scores from samples of $n = 2$ from the quiz score population.

The mean (=6.000) and standard deviation (=0.845) of the 15 sample means are measures of center and dispersion for the sampling distribution. The standard deviation of statistics (i.e., 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 keep the dispersion of the sampling distribution separate from the dispersion of individuals in the population, which is measured by the standard deviation. Thus, the standard deviation

[2] These 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))`.

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

of all possible sample means is referred to as the standard error of the sample means (SE). The SE in this example is 0.845. The standard deviation is the dispersion of individuals in the population and, in this example, is 1.414.

This example illustrates three major concepts concerning sampling distributions. First, the sampling distribution 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 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 estimate of the population mean.

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.

◇ **All statistics in this course are unbiased.**

13.1.2 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 8), 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 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 dispersion of all individuals in a sample. In contrast, the standard error measures 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.

13.1.3 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 13.2. The mean of these means is 6.000, the standard error is 0.592, and the sampling distribution is symmetric, perhaps approximately normal (Figure 13.2). The three major characteristics of sampling distributions noted in Section 13.1.1 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 $n = 2$.[4]

[4] One should also look at the results from $n = 4$ in one of the online Review Exercises.

Table 13.2. All possible samples of $n = 3$ and corresponding sample means from the quiz score population.

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

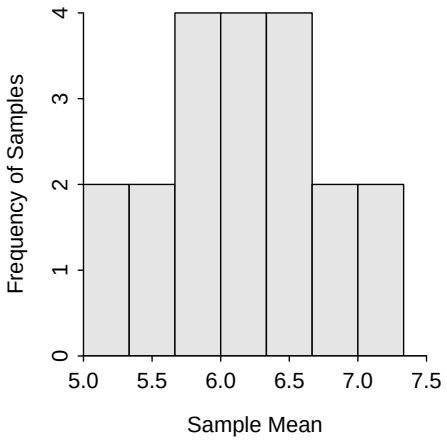


Figure 13.2. Sampling distribution of mean quiz scores from samples of $n = 3$ from the quiz score population.

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 13.3 and the actual sampling distribution is shown in Figure 13.3. 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 13.3. All possible samples of $n = 3$ and corresponding sample medians from the quiz score population.

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

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.**

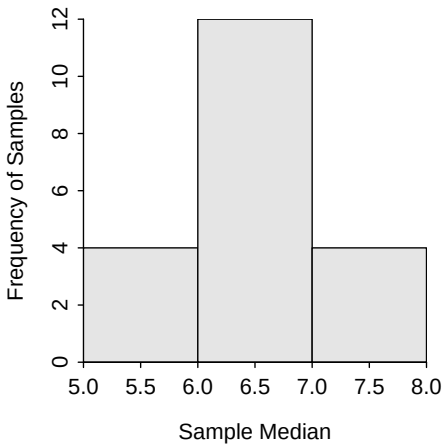


Figure 13.3. Sampling distribution of median quiz scores from $n = 3$ samples from the quiz score population.

13.1.4 Simulating

Exact sampling distributions can only be 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 and even larger situations?

There are two ways to examine sampling distributions in situations with large sample and population sizes. First, theorems exist that describe the specifics of sampling distributions under certain conditions. One such theorem is described in Section 13.2. Second, 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." We will simulate some sampling distributions here so that the theorem will be easier to understand.

Sampling distributions are simulated by drawing many samples from a population, computing the statistic of interest for each sample, and constructing a histogram of those statistics (Figure 13.4). The computer is helpful with this simulation; however, keep in mind that the computer is basically following the same process as used in Section 13.1.1, with the exception that not every sample is taken.

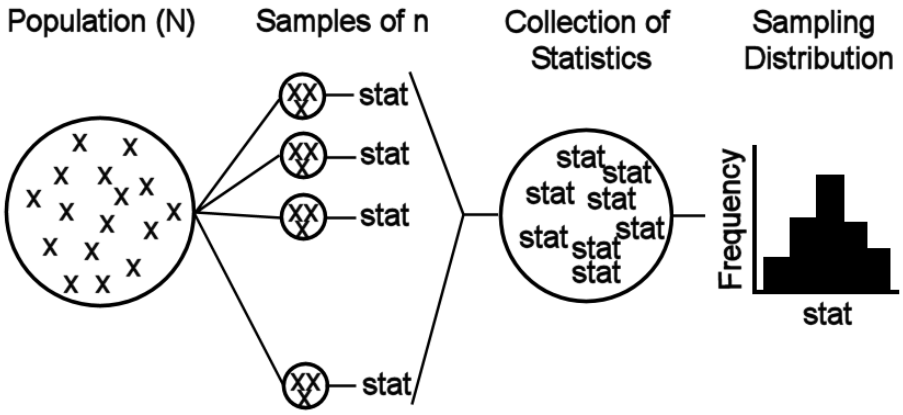
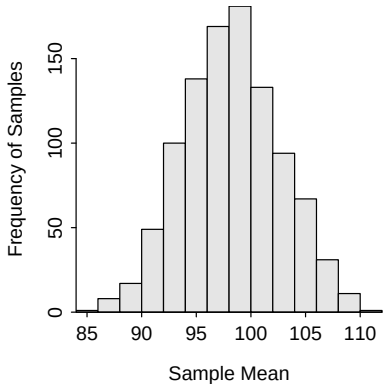


Figure 13.4. Schematic representation of the process for simulating a sampling distribution.

Let's return to the Square Lake fish population from Section 2.1 to illustrate simulating a sampling distribution. 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 13.5. 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.



	Means
mean	98.25
sd	4.43
min	85.46
Q1	95.14
median	98.26
Q3	101.27
max	110.1

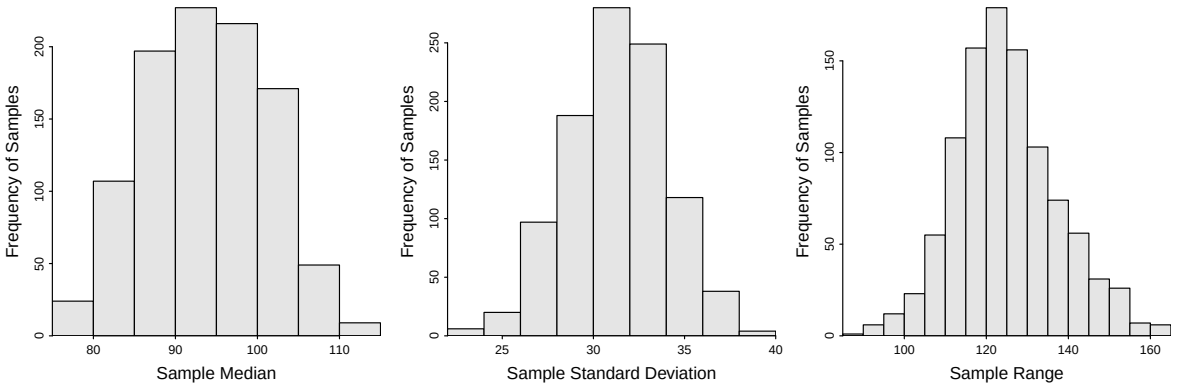
Figure 13.5. Histogram (**Left**) and summary statistics (**Right**) from 1000 sample mean total lengths computed from samples of $n = 50$ from the Square Lake fish population.

As with the actual sampling distributions discussed previously, three characteristics (shape, center, and dispersion) are examined with simulated sampling distributions. First, Figure 13.5 looks at least approximately normally distributed. Second, the mean of the 1000 means (=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" 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 13.6).

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.**



	Medians
mean	93.45
sd	7.34
min	75.5
max	112.5
Parameter	93

	Std. Devs
mean	31.27
sd	2.76
min	22.98
max	39.32
Parameter	31.49

	Ranges
mean	124.53
sd	12.7
min	88
max	164
Parameter	164

Figure 13.6. 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.

13.2 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 13.1.1) or taking a large number of samples from a large population (Section 13.1.4). 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 to identify 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.

The validity of the CLT can be examined by simulating several (with different n) sampling distributions of $\bar{x}$ from the Square Lake population and from a strongly right-skewed exponential distribution (Figure 13.7). Several observations about the CLT can be made from Figure 13.7. First, the sampling distribution is approximately normal for $n \geq 30$ for both scenarios and is approximately normal for smaller n for the Square Lake example because that population is only slightly skewed. Second, the means of all sampling

distributions in both examples are approximately equal to μ, 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.

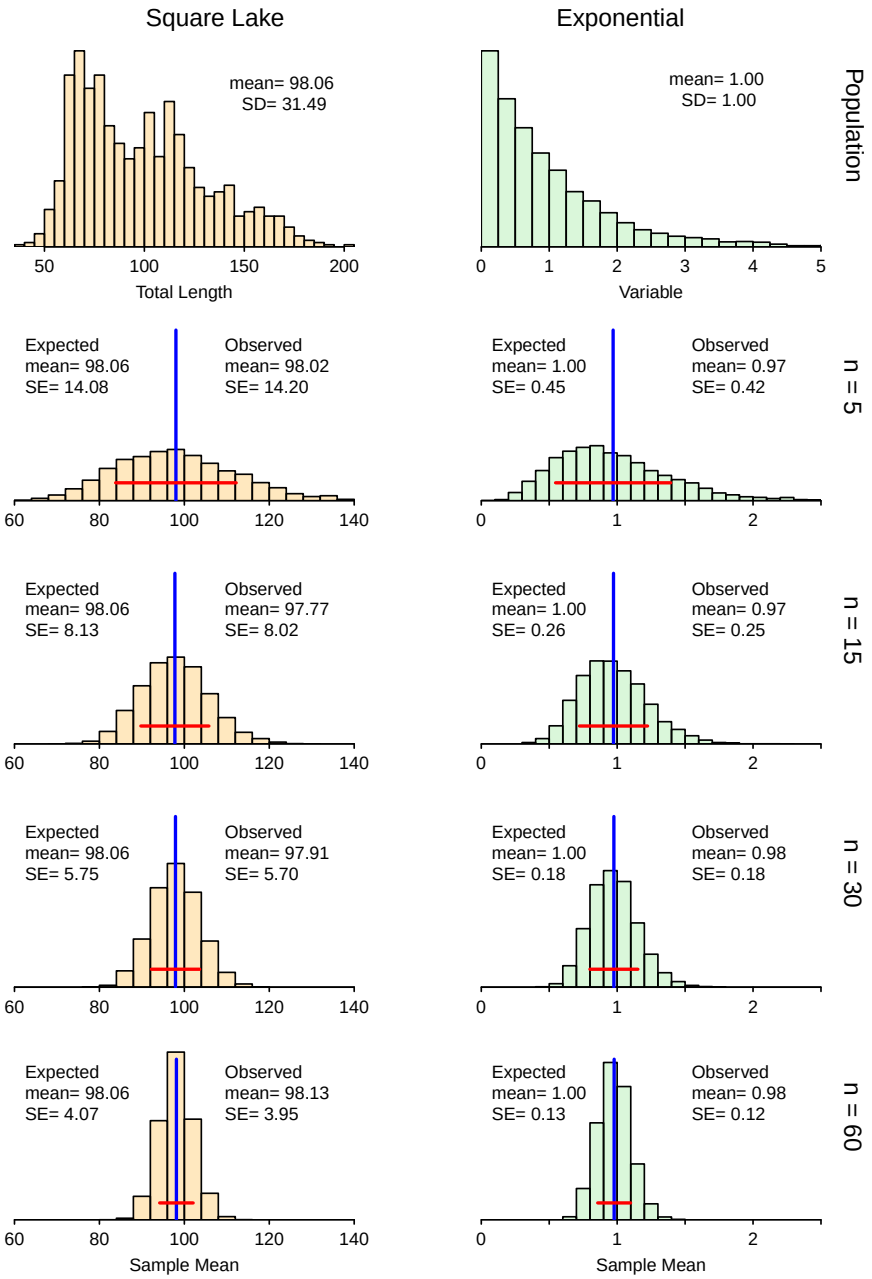


Figure 13.7. Sampling distribution of the sample mean simulated from 5000 samples of four different sample sizes extracted from the Square Lake fish population (Left) and an exponential population (Right). The shapes of the populations are shown in the top histogram. On each simulated sampling distribution, the vertical blue line is the mean of the 5000 means and the horizontal red line represents ±1SE from the mean.

13.3 Probability Calculations

If the sample size is large enough, then the CLT states that the sampling distribution of sample means is approximately normal. If the sampling distribution is normal, then the methods from Module 8 may be used to compute probabilities. Therefore 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 probabilitiy of **statistics** can be answered.

The question above is answered by first recalling that, for the length of all fish in Square Lake, μ =98.06 and σ =31.49. Because $n = 50$ is greater than 30, the CLT says that the distribution of the sample means from these samples is $\bar{x} \sim N(98.06, \frac{34.19}{\sqrt{50}})$ 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 13.8-Left).[5]

```
> ( distrib(95,mean=98.06,sd=34.19/sqrt(50)) )
[1] 0.2634127
```

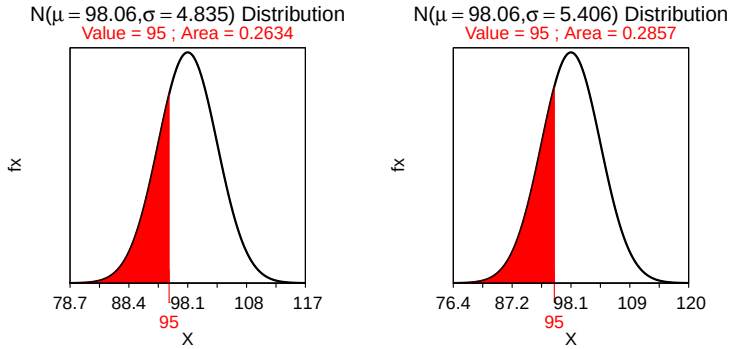


Figure 13.8. Proportion of sample means less than 95 mm on a $N(98.06, 4.84)$ (Left) and $N(98.06, 5.406)$ (Right) distribution.

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 $\bar{x} \sim N(98.06, \frac{34.19}{\sqrt{40}})$ 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 13.8-Right).

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

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

[5] Notice 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 13.1.1). 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=`.

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 questions 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.[6]

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 is 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 distibutions for when they are needed to make inferences in later modules.

13.4 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 13.9. The targets in Figure 13.9 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.

[6] At least with the techniques in this course.

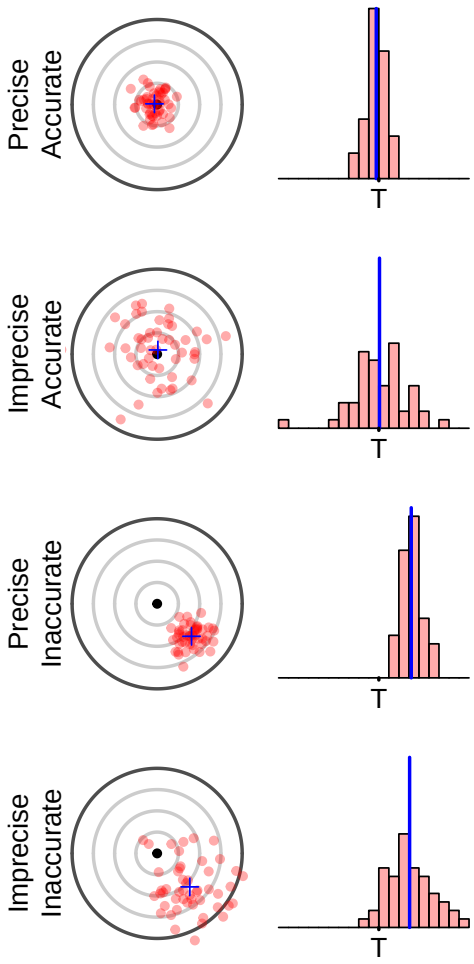


Figure 13.9. Model used to demonstrate accuracy, precision, and bias. The center of each target (i.e., the bullseye) and the point marked with a "T" (for "truth") represent the parameter of interest. Each dot on the target represents a statistic computed from a single sample and, thus, the many red dots 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.