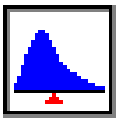
CHAPTER 8



Visualizing Properties of Estimators

CONCEPTS

- Estimator, Properties, Parameter, Unbiased Estimator, Relatively Efficient Estimator, Consistent Estimator, Asymptotically Unbiased Estimator, Sufficient Estimator, Sampling Distribution, Empirical Sampling Distribution

OBJECTIVES

- Recognize how the distribution of an estimator is affected by sample size and the shape of the distribution being sampled
- Understand the properties of unbiasedness, asymptotic unbiasedness, relative efficiency, consistency, and sufficiency
- Realize that there are a variety of estimators for a parameter
- Recognize that not all estimators for a parameter are equally good in terms of their properties

Overview of Concepts

An **estimator** or sample statistic is a formula that enables the statistician to estimate an unknown **parameter** based upon a sample of data. The value obtained is called an estimate. This estimate provides the statistician with a possible value for the unknown parameter. However, because a variety of alternative estimators may be used to estimate an unknown parameter, statisticians have developed a list of desirable **properties** that can be used to evaluate the different estimators.

The most fundamental of these properties is called *unbiasedness*. An **unbiased estimator** is one whose expected value equals the unknown parameter. This means that if a very large number of samples are drawn and an estimate of the unknown parameter is obtained from each sample, the average value of those estimates will equal the parameter value. A related property is called *asymptotic unbiasedness*. An **asymptotically unbiased estimator** is a biased estimator whose bias goes to zero as the sample size approaches infinity.

Another important property is called *efficiency*. An efficient estimator is the estimator that has the smallest variance out of *all* unbiased estimators for the parameter. Because this is generally impossible to show (since there may be a very large number of unbiased estimators), statisticians have developed sophisticated methods to prove that an estimator is efficient. A less-encompassing but related property is called *relative efficiency*. Estimator A is a **relatively efficient estimator** compared with estimator B if A has a smaller variance than B *and* both A and B are unbiased estimators for the parameter.

Another asymptotic property is called *consistency*. A **consistent estimator** is an estimator whose probability of being close to the parameter increases as the sample size increases. The probability approaches 1 as the sample size approaches infinity. This property is often demonstrated by showing that an unbiased or asymptotically unbiased estimator has a standard error that decreases as the sample size increases. As a result, the estimator collapses on the parameter value as the sample size approaches infinity. This property is called mean squared error consistency. An estimator that is mean squared error consistent is a consistent estimator; however, it is possible for an estimator to be consistent even if it is not mean squared error consistent. This atypical situation will not be considered in this module.

The final statistical property considered here is *sufficiency*. A **sufficient estimator** is an estimator that uses all available information in a sample about the parameter it is estimating. Statisticians prefer sufficient estimators because they usually have a smaller variance.

In this module you can conduct experiments to construct the **empirical sampling distribution** for a variety of estimators for μ (population mean) and σ^2 (population variance). This empirical sampling distribution can be compared with the theoretical **sampling distribution** of either $\overline{X}$ or s^2. In addition, the mean of the empirical sampling distribution can be compared with the true population parameter, and the variance of the empirical sampling distribution can be examined to see if it decreases as the sample size is increased.

With the development of high speed computers, an entire class of new estimators has been developed. They are called resampling or bootstrap estimators (originally developed by Brad Efron and others since 1979). These estimators base the estimate on the average result after you resample your original sample K times. Using this technique, the bootstrap estimator for the variance was developed. This has turned out to be a relatively efficient estimator compared with the traditional sample variance estimator, s^2.

Illustration of Concepts

The **estimator** $\bar{X}$(sample mean) is an **unbiased estimator** for the **parameter** μ (population mean). This can be verified by conducting a simulation experiment. Consider light bulb failures. Many light bulb packages state that a light bulb will last an average of 1,000 hours (μ = 1000). This can be modeled using an exponential distribution with λ = 0.001 (reciprocal of the mean) as shown in Figure 1.

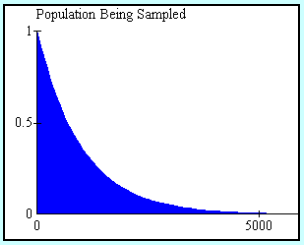


Figure 1: Exponential Distribution

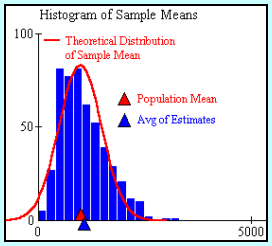


Figure 2: 500 Means with n = 4

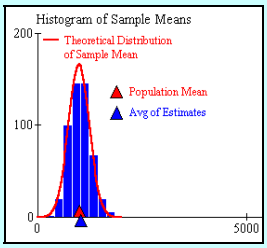


Figure 3: 500 Means with n = 16

We take 500 samples of size 4 from this exponential population, find the mean of each sample, and create a histogram of the 500 means to get an **empirical sampling distribution**. The Central Limit Theorem says that $\bar{X}$ will have a normal distribution if the sample size is large enough. If we superimpose a normal **sampling distribution** (Figure 2) with a mean of 1000 and a variance of $\sigma^2/n = 1000^2/4 = 250{,}000$ on our histogram, we see that the histogram of means is somewhat skewed (although much less so than the exponential population). This suggests that n = 4 is not large enough for the sampling distribution to be normal when sampling an exponential distribution. For this experiment, the mean of the 500 sample means (980.4) was very close to the population mean (μ = 1000), providing evidence that the sample mean is an unbiased estimator for μ (the difference is due to sample variation).

The sample mean is a **consistent estimator** of the population mean that will collapse on the true mean μ when n is large. For an unbiased or an **asymptotically unbiased estimator,** consistency can be demonstrated by showing that the estimator's standard error decreases as the sample size increases. To do this, we repeat the previous experiment using a larger sample size of n = 16, obtaining the histogram shown in Figure 3. Despite the skewed population, the histogram for n = 16 is nearly symmetric, is narrower than for n = 4, and is almost normally distributed, as suggested by the Central Limit Theorem. The 500 sample means have a mean of 1,005.1 and a variance of 61,504 (quite close to the theoretical values of μ = 1,000 and $\sigma^2/n = 1000^2/16 = 62{,}500$). Because the variance of the sampling distribution decreased as the sample size increased, we see evidence that the sample mean is a consistent estimator for μ. that This is an *asymptotic* **property** because the sample mean slowly collapses on the population mean as the sample size approaches infinity.

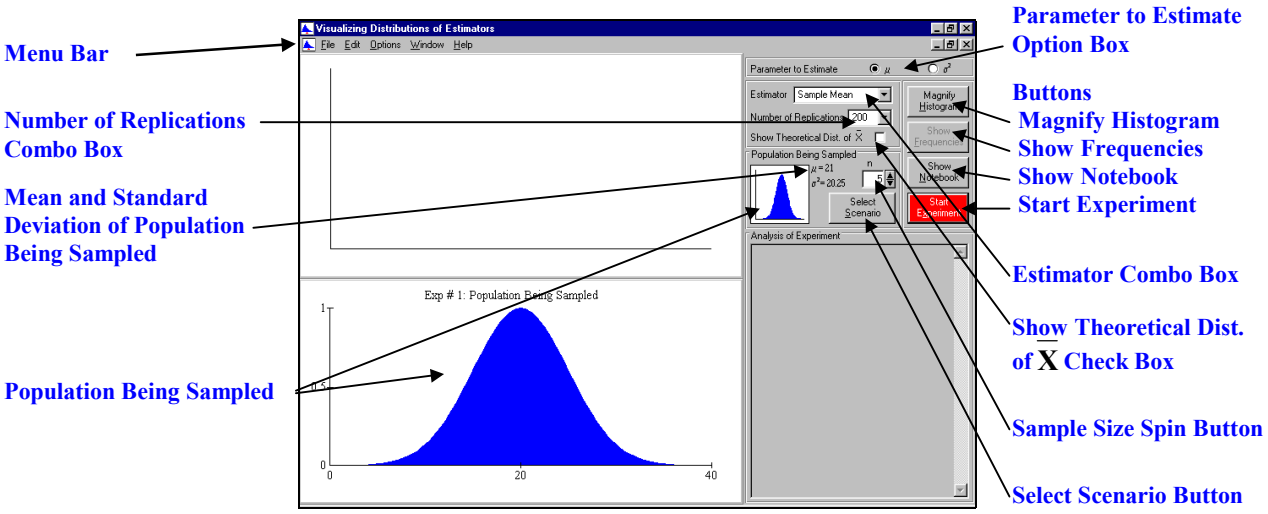
The sample mean is a **sufficient estimator** because it uses *all* the information in the sample (i.e., all n sample points were used in its calculation). But to show that the sample mean is *efficient* requires showing that the sample mean has a smaller variance than *all* unbiased estimators for μ, which is impossible to do. However, we can show that it is a **relatively efficient estimator** compared with another unbiased estimator, the average of a subsample of size 2. Using this estimator, 500 estimates were calculated using a sample size of 16. The mean and variance of this sampling distribution were 998.1 and 495,630, respectively. Because the variance of the 500 sample means was only 61,504, we see that the sample mean is relatively efficient compared with the subsample of size 2 estimator for the population mean.

Orientation to Basic Features

This module conducts an experiment to create sampling distributions for six estimators for μ or five estimators for σ^2. The population being sampled may be selected from a scenario, chosen from a template of distributions, or specified by the user. The module illustrates the statistical properties of unbiasedness, relative efficiency, and consistency.

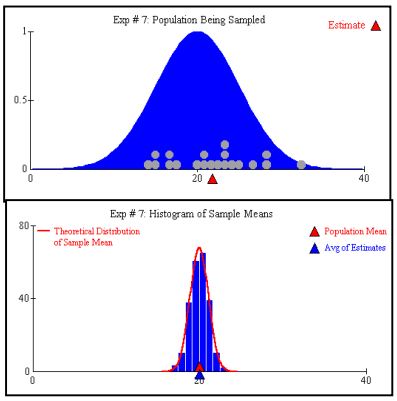
1. Opening Screen

Start the module by clicking on the module's icon, title, or chapter number in the *Visual Statistics* menu and pressing the Run Module button. When the module is loaded, you will be on the introduction page of the Notebook. Read the questions and then click the Concepts tab to see the concepts that you will learn. Click on the Scenarios tab. Select Sample mean from the list of choices under Estimators for Mu. Select a scenario, read it, and press OK. The upper left quadrant of the display is empty, waiting to create a histogram of sample estimates. The upper right quadrant of the display contains the Control Panel. In the lower left quadrant is the distribution of the population being sampled, and to its right will appear an analysis of the experiment after it has been completed.



2. Start Experiment Button

Click on the Start Experiment button. The experiment will sample from the population and estimate the sample mean 200 times. Click the Pause Experiment button. Superimposed on the population distribution is a dot plot of a sample and its sample mean (in red) as shown in the figure to the right. The sample mean is also shown on the histogram in the upper left screen. Click the Continue Experiment button. At the end of the experiment a histogram of the sampling distribution has been created. It can be compared with the theoretical distribution of the sample mean (normal with mean of μ and a variance of σ^2/n) superimposed in red (see figure to the right). During the experiment, if you do not want the graph updated one sample at a time, click the Finish Experiment button. Below the Control Panel appears an analysis of the experiment.



3. **Control Panel**
   a. The Number of Replications combo box sets the number of samples to be drawn. In general, 200 or more replications are needed to create a reliable representation of the true sampling distribution.
   b. Push the Show Frequencies button to see the number of estimates in each histogram class.
   c. Click the n spin button to change the sample size (or click inside its box and enter a number from your keyboard).
   d. Click on the Select Scenario button to change the scenario. This will return you to the Scenario section of the Notebook. You can browse through the scenarios by clicking on Next page or Previous page.
   e. Click the Show Theoretical Dist. of $\bar{X}$ check box. The theoretical sampling distribution of the sample mean appears (based on the Central Limit Theorem).
   f. Press the Magnify Histogram button to redraw the histogram using the entire left half of the screen and rescale the histogram so that its shape is more easily seen. Click the Reduce Histogram button to return to the original display.
   g. Click on the Estimator combo box. Six estimators for the mean are listed. Change the estimator to Sample Median. Start the experiment. The histogram displays the distribution of sample medians when you sample from the distribution in the lower left quadrant. Six estimators for the mean are listed.
   h. Select σ^2 using the Parameter to Estimate option buttons. The Estimator combo box lists five estimators of the variance. See Help for definitions of any estimator. The Show Theoretical Dist. of $\bar{X}$ check box is relabeled Show Theoretical Dist. of s^2.

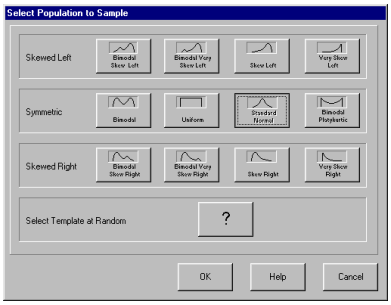
4. **Copying a Display**
   Click on the graph you wish to copy or the Analysis panel. Black handles appear indicating it has been selected. Select Copy from the Edit menu (on the menu bar at the top of the screen) or press Ctrl-C to copy the display. It can then be pasted into other applications, such as Word or WordPerfect, so it can be printed or made part of a report.

5. **Help**
   Click on Help on the menu bar at the top of the screen. Search for Help lets you search a topic index, Contents shows a table of contents, Using Help gives instructions on how to use Help, and About gives licensing and copyright information about this *Visual Statistics* module.

6. **Exit**
   Close the module by selecting Exit in the File menu (or click  in the upper right-hand corner of the window). You will be returned to the *Visual Statistics* main menu.

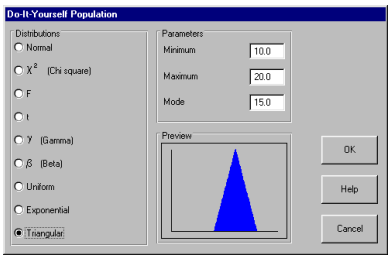
Orientation to Additional Features

1. **Population Templates**



Press the Show Notebook button to return to the Notebook. Select the Pop. Templates tab and click OK. The display on the right will appear. Press the Very Skewed Left button (upper right corner) and click OK. The new distribution being sampled appears in the lower left quadrant and in the Population Being Sampled panel along with its mean and variance. The Select Scenario button has been replaced with a Select Templates button providing a short-cut to the templates display.

2. **Do-It-Yourself Distributions**



Press the Show Notebook button and select the Do-It-Yourself tab. Select Known Distributions or User-Created Distributions. Pressing OK will bring up a display enabling you to create or select a distribution from which to sample. Accept the distribution by clicking OK. The Select Templates button will be replaced by a Select Distribution button if Known Distributions is selected or by a Show DIY Controls button if User-Created Distributions is selected.

3. **Second Display**

Select Copy Current Window from the Windows menu on the menu bar. You can now run two experiments and compare the results.

4. **Options**

Click on Options on the menu bar and select Change Title to retitle the current display. This can be useful if you create a second display and need to keep track of them.

Basic Learning Exercises Name _____

Replication Experiment

1. Press the Show Notebook button, select the Scenarios tab, click on Sample mean under Estimators for Mu, and select Golf Professional. Read the scenario and click on OK. What is the mean and variance of the population being sampled (look on the Population Being Sampled panel on the Control Panel)? What parameter is being estimated? What estimator is being used? What is the equation for this estimator? **Hint:** Select Help on the menu bar, click Search for Help, and type "sample mean" for its definition.

   Population being Sampled: Mean _____ Variance _____

2. Set the Number of Replications combo box to 50. Press the Start Experiment button. Press the Pause Experiment button after the experiment starts. a) Describe what is being displayed on the bottom graph. b) What does the triangle represent? c) What is being displayed on the top graph?

3. Press Continue Experiment to restart the experiment and watch the histogram build. Press Finish Experiment to suspend displaying each replication and complete the experiment. Read the Analysis of Experiment panel. What is the average of the sample means and the standard error of the sample means? What should each be according to the Central Limit Theorem (CLT)? Are the experiment results close to these predictions? What do the blue and red triangles on the top diagram represent? What do their relative positions tell you?

   Average of Sample Means _____ Standard Error of Sample Means _____
   CLT Predicted Mean _____ CLT Predicted Standard Error _____

Central Limit Theorem

4. Give a formal statement of the Central Limit Theorem. **Hint:** Use Help if you need it.

5. Increase the sample size to 100 by using the n spin button or typing the number directly into the box and pressing the Enter key on your computer. What is the standard deviation of the population being sampled? According to the CLT, what will be the standard error of the sampling distribution? What formula did you use to calculate the sampling distribution? What does the standard deviation measure? What does the standard error measure?

   Standard Deviation _____ CLT Predicted Standard Error _____

6. Select 500 from the Number of Replications combo box. Press the Start Experiment button and then the Finish Experiment button. What is the average of the sample means and the standard error of the sample means? Read the Analysis of Experiment. How do these results and those in exercise 3 illustrate the Central Limit Theorem?

   Average of Sample Means _____ Standard Error of Sample Means _____

7. Press the Select Scenarios button. Select the Smoke Stack Emissions scenario and read it. Press OK. Reduce the sample size from 30 to 10. Press the Start Experiment button. Read the Analysis of Experiment. Is the sampling distribution normally distributed? **Hint:** Press the Magnify Histogram button and make sure that Show Theoretical Dist. of $\bar{X}$ is checked in order to evaluate the histogram visually. When finished, press the Reduce Histogram button and uncheck Show Theoretical Dist. of $\bar{X}$ by clicking on it.

8. Increase the sample size to 30 and press Start Experiment. Is the sampling distribution normally distributed? How does exercise 7 illustrate the Central Limit Theorem?

Intermediate Learning Exercises Name _____

Unbiased Estimators/Sufficient Estimators

9. What does it mean for an estimator to be unbiased? **Hint:** Click Search for Help in the Help menu and type "unbiased" if you are not sure.

10. How does the Central Limit Theorem prove that the sample mean is an unbiased estimator?

11. Press the Select Scenario button and click on Next page. Select Accuracy of Pistol Shooter and read the scenario. Click OK. What is the equation (estimator) for the sample median? Is the median a sufficient estimator? **Hint:** Use Help or the Glossary for definitions of terms.

12. Press the Start Experiment button. Does the median provide an unbiased estimator for μ in this case? Why or why not? **Hint:** Use the largest number of replications possible when you do an experiment. If the experiment took less than 15 seconds to finish, increase the number of replications from 500 to 1000 (or decrease the number of replications to 200 if it took longer than 30 seconds).

13. Press the Select Scenario button. Select the SAT Verbal Test Scores scenario and read it. Click OK. Press the Start Experiment button. Does the median provide an unbiased estimator for μ in this case? Explain.

14. Why does the median provide an unbiased estimator in exercise 13 but not in exercise 12?

Consistent Estimator

15. Push the Select Scenario button and click on Previous page. Select The Sum of Two Random Numbers scenario. Read the scenario and click OK. Press the Start Experiment button. Record the average of the sample means and their standard error. Increase n to 30 and repeat the experiment. Increase n to 90 and repeat the experiment.

n	Average of Sample Means	Standard Error of Sample Means
10	_____	_____
30	_____	_____
90	_____	_____

16. a) What is the definition of a consistent estimator? b) How can you show that an estimator is consistent? c) How do the results you recorded in exercise 15 demonstrate that the sample mean is a consistent estimator for μ? **Hint:** Use Help or the Glossary if you need guidance.

Relatively Efficient Estimator

17. Define a relatively efficient estimator. How would you show that an estimator is relatively efficient? **Hint:** Use Help or the Glossary if you need guidance.

18. Change the estimator to Sample Median using the Estimator combo box. Set the sample size to 10 and press the Start Experiment button. Record the average of the sample medians and their standard error. Increase n to 30 and repeat the experiment.

n	Average of Sample Medians	Standard Error of Sample Medians
10	_____	_____
30	_____	_____

19. Compare the results of exercises 17 and 18. How do these results show that the sample mean is relatively efficient compared with the sample median if a population with a symmetric distribution is being sampled?

Advanced Learning Exercises Name _____

Unbiased Estimator for the Variance

Push the Show Notebook button, select the Do-It-Yourself tab, and click on Known Distributions. Read the description and press OK. Select the Exponential option button and enter 0.1 in the Scale parameter box. A sketch of the distribution appears in the window. Press OK. You are now sampling from the exponential distribution you specified. Click the σ^2 option button in the Parameter to Estimate panel. Make sure that the Sample Variance is appearing in the Estimator combo box. Use the Number of Replications combo box to specify the largest number of replications possible given the speed of your computer. The n spin button should be 9.

20. Press the Start Experiment button. a) Define the sample variance estimator. b) Read the Analysis of Experiment panel. What is the value of σ^2 in the population? c) What is the average of the sample variances? d) Is the sample variance an unbiased estimator for σ^2? e) How can you tell by looking at the Histogram of Sample Variances display?

21. Press the Select Distribution button. Select Normal distribution. Enter 5 for standard deviation and whatever mean you desire. Click OK. You are now sampling from the specified normal distribution. Press the Start Experiment button. What is the value of σ^2 in the population? What is the average of the sample variances? Given these results and those in exercise 20, does the distribution of the population being sampled cause the sample variance to become a biased estimator for σ^2?

Distribution of Sample Variance

22. Make sure the Show Theoretical Dist. of s^2 box is checked. The sample variance estimator has a χ^2 distribution that is scaled by $\sigma^2/(n-1)$. Notice how closely the histogram's shape corresponds to this theoretical distribution. Press the Select Distribution button. Select the t distribution and type 10 into the Degrees of freedom box. Although not normally distributed, the distribution has an approximate bell-shape. Press OK. Press the Start Experiment button. Read the Analysis of Experiment panel. Does the histogram have the scaled χ^2 distribution? Is the sample variance still an unbiased estimator?

23. Increase the sample size to 100 using the n spin button. Press the Start Experiment button. Read the Analysis of Experiment panel. Does the histogram have the scaled χ^2 distribution? The Central Limit Theorem says that the sample mean will have a normal distribution if the sample size is large enough. Given these results, do you think a similar theorem could be proven for the sample variance?

Asymptotically Unbiased Estimator

24. Press the Select Distribution button and select Normal. Set its standard deviation to 9 and its mean to 20. Press OK. Change the Estimator combo box to Sample MSD. Set the sample size to 4. Press the Start Experiment button. a) Define the sample MSD estimator. b) What is the variance of the population you are sampling? c) What is the average of the sample MSDs? d) Is the sample MSD an unbiased estimator for σ^2?

25. Increase the sample size to 100. Press the Start Experiment button. Read the Analysis of Experiment panel. What is the average of the sample MSDs? An estimator is asymptotically unbiased if the bias disappears as the sample size increases. Is sample MSD an asymptotically unbiased estimator for σ^2?

Bootstrap Estimator for σ^2

26. Change the Estimator combo box to Bootstrap-Variance and enter 50 in the dialog box. a) Define the bootstrap estimator for the variance. b) What does the 50 represent? c) Set the sample size to 20. Press the Start Experiment button. Why does this experiment take longer than the other experiments? d) Read the Analysis of Experiment panel. Why is the bootstrap estimator for σ^2 relatively efficient compared with the sample variance estimator?

Individual Learning Projects

Write a report on one of the three topics listed below. Use the cut-and-paste facilities of the module to place the appropriate graphs in your report. For each investigation use at least 200 replications (1000 is recommended).

1. Investigate three estimators for either μ or σ^2 and determine which are unbiased estimators. For each estimator investigate its bias by drawing a very small sample from three different distributions (symmetric, skewed, and very skewed). Define each estimator. To be unbiased, an estimator's experimental sampling distribution must be centered around the true parameter (μ or σ^2) for each investigation.
2. Investigate one estimator for the mean (not the sample mean) and one estimator for the variance (not the sample variance). Define each estimator and explain how it is calculated. Determine whether the estimator for the mean is unbiased (or asymptotically unbiased), whether it is consistent, and assess its relative efficiency compared with the sample mean. Determine whether the estimator for the variance is unbiased, whether it is consistent, and assess its relative efficiency compared with the sample variance. In your investigation, be sure to sample both from a symmetric and a very asymmetric distribution.
3. The Central Limit Theorem says that the sample mean is an unbiased estimator for the population mean and that its standard error is $\sigma/\sqrt{n}$ regardless of the distribution of the population being sampled. It goes on to say that the sampling distribution of the sample mean is normally distributed if the sample size is large enough. Illustrate the Central Limit Theorem. Select at least three different distributions to sample (they should be increasingly asymmetric). For each distribution, select three different sample sizes, ranging from very small (fewer than 8) to large (greater than 80). Discuss how these nine experiments illustrate the Central Limit Theorem.

Team Learning Projects

Select one of the three projects listed below. In each case, produce a team project that is suitable for an oral presentation. Use presentation software or large poster boards to display your results. Graphs should be large enough for your audience to see. Each team member should be responsible for producing some of the graphs. Ask your instructor if a written report is also expected. For each investigation use at least 200 replications (1000 is recommended).

1. A team of two to four should evaluate the relative efficiency of each estimator for either μ or σ^2. An estimator is relatively efficient compared with another estimator if both are unbiased and one has a smaller standard error than the other. For each estimator, investigate its relative efficiency by drawing a very small sample from four different distributions. Define each estimator. Tell which are unbiased and rank the relative efficiency of those that are unbiased. Support all findings with illustrations.
2. A team of three or four should evaluate the consistency of each estimator for either μ or σ^2. An estimator is consistent if it is unbiased or asymptotically unbiased (its bias disappears as n increases) and if the standard error of the sampling distribution decreases as n increases. For each estimator, investigate its consistency by drawing two samples, one very small and one moderate, from three different distributions (symmetric, skewed, and very skewed). Define each estimator and tell which are unbiased, asymptotically unbiased, and consistent. Support all findings with illustrations.
3. A team of two to four should evaluate the properties of estimators for the mean and variance. Each team member should select an estimator for μ and an estimator for σ^2 to evaluate. No one should select the sample mean or the sample variance. Define each estimator selected and explain how it is calculated. For each chosen estimator for μ, determine whether it is unbiased (or asymptotically unbiased), whether it is consistent, and assess its relative efficiency compared to the sample mean. For each chosen estimator for σ^2, determine whether it is unbiased, whether it is consistent, and assess its relative efficiency compared to the sample variance. For each estimator investigated, be sure to sample both from a symmetric and very asymmetric distribution. Support all findings with illustrations.

Self-Evaluation Quiz

1. Which of the following is an *unbiased* estimator of the population mean?
   a. The sample mean, regardless of the population sampled.
   b. The sample midrange, if the distribution of the population is symmetric.
   c. The sample median, if the distribution of the population is symmetric.
   d. The sample median, if the distribution of the population is skewed.
   e. All of the above are unbiased estimators.

2. The sample median
   a. is an unbiased estimator of μ if the population sampled is symmetric.
   b. is an asymptotically unbiased estimator of μ.
   c. is not efficient since it is not always an unbiased estimator of μ.
   d. has more than one of the above characteristics.
   e. has none of the above characteristics.

3. Relative efficiency requires an examination of
   a. the variability of a sample estimator.
   b. the bias of a sample estimator.
   c. both the bias and the variability of a sample estimator.
   d. neither the bias nor the variability of a sample estimator.
   e. none of the above.

4. If estimator A is relatively efficient compared with estimator B, then which is most likely?
   a. Estimator A is probably based on a smaller sample than is estimator B.
   b. Estimator A has a smaller bias than does estimator B.
   c. Estimator A is less sensitive to population shape than is estimator B.
   d. Estimator A has a smaller variance than does estimator B.
   e. Estimator A is based on a better sampling method than that of estimator B.

5. If you sample from a population with a symmetric distribution, which of the following is the relatively efficient estimator of the population mean?
   a. The average of a subsample of size K of the n sample observations ($K < n$).
   b. The sample midrange.
   c. The sample median.
   d. The sample mean.
   e. More than one of the above.

6. In a sample of size n, which is *not* a consistent estimator of the population mean?
   a. Subsample of size K of the n sample observations ($K < n$).
   b. Subsample of K% of the n sample observations ($K < 100$).
   c. The sample mean.
   d. All of the above are consistent estimator of the mean.
   e. None of the above is a consistent estimator of the mean.

7. Which is a correct statement about the Empirical Rules?
   a. The Empirical Rule of 4 is based on the $\mu \pm 2\sigma$ normal area (95.44 percent).
   b. The Empirical Rule of 6 is based on the $\mu \pm 3\sigma$ normal area (99.74 percent).
   c. The Empirical Rules are biased estimators even if the population is symmetric.
   d. The Empirical Rules are inconsistent and inefficient estimators.
   e. The Empirical Rules have all of the above characteristics.

8. The bootstrap estimator
   a. uses nonlinear variable transformations to improve accuracy.
   b. was discovered long before powerful computers were available.
   c. was initially developed by statistician Brad Efron (1979).
   d. was first applied to quality sampling in the shoemaking industry.
   e. provides biased but efficient estimates of a parameter.

9. Which of the following is a biased estimator of the population variance?
   a. The sample variance.
   b. The sample MSD.
   c. The sample Empirical Rule of 4.
   d. The sample bootstrap estimator.
   e. More than one of the above is a biased estimator.

10. Which of the following is the most efficient estimator of the population variance?
    a. The sample MSD.
    b. The sample Empirical Rule of 4 estimator.
    c. The sample Empirical Rule of 6 estimator.
    d. The sample bootstrap estimator.
    e. None of the above is an efficient estimator.

11. Which of the following is an asymptotically unbiased estimator of the population variance?
    a. The sample MSD.
    b. The sample Empirical Rule of 4 estimator.
    c. The sample Empirical Rule of 4 estimator.
    d. The sample bootstrap estimator.
    e. None of the above is asymptotically unbiased.

12. Which is *not* true of the bootstrap estimator for the variance?
    a. It often has a smaller variance than other estimators for the variance.
    b. It is somewhat more difficult to calculate than other estimators.
    c. It is widely used in the footwear industry to improve product quality.
    d. It is growing in popularity partly because it requires few assumptions.
    e. It may yield different results if the number of times the sample is resampled is not large.

Glossary of Terms

Asymmetric distribution Distribution that is skewed right or left. See **Symmetric**.

Asymptotic Refers to large sample sizes. Formally it refers to a limit as the sample size approaches infinity (denoted $n \to \infty$).

Asymptotic properties Properties of an estimator that can be evaluated only as the sample size approaches infinity. Essentially, this is a mathematical question, though it can be investigated using simulation of sampling as sample size is increased.

Asymptotically unbiased estimator Biased estimator whose bias decreases toward zero as the sample size increases (i.e., the probability of a given difference between the estimator and the parameter approaches zero as $n \to \infty$).

Average of K observations Same as the sample mean except that only K observations in the sample are used. It is unbiased but is not consistent (because K does not change as n increases) and has a larger variance than the sample mean. Because it is a mean, it is vulnerable to the presence of outliers in the sample.

Average of K% of the observations Same as the sample mean except that only K percent of the observations are included in the sample. It is unbiased and consistent, but is has a larger variance than the sample mean. Because it is a mean, it is sensitive to outliers in a sample.

Biased estimator An estimator whose expected value is not equal to the corresponding population parameter. See **Unbiased estimator**.

Bootstrap estimator Once the sample is taken, the sample is resampled K times, calculating the sample mean each time. The bootstrap estimator for the mean is the average of these sample means. It is unbiased and consistent. It has the same sampling distribution as the sample mean, and as a result is efficient. The bootstrap estimator of the variance is the square root of the variance of the K resampled means (an estimator of the sample standard error of the mean) multiplied by $\sqrt{n}$ and squared. It is unbiased and consistent. It has a smaller spread than the sample variance and is therefore relatively efficient compared with the sample variance.

Central Limit Theorem This famous theorem says that, if the sample size is large enough, sample means from any population will follow a normal distribution with the same mean as the population but with a smaller variance. More formally, if $x_1, x_2, \ldots, x_n$ are identically distributed, independent, random variables from any population with finite mean μ and finite variance σ^2, then the sample mean, $\overline{X} = (x_1 + x_2 + \ldots + x_n) / n$, will have a sampling distribution with mean μ and variance σ^2 / n. Further, the sampling distribution of the sample mean $\overline{X}$ is normally distributed as n approaches infinity.

Consistent estimator Estimator whose probability of being close to the parameter being estimated increases as the sample size increases. One can demonstrate that an estimator is consistent by showing that it is unbiased (or asymptotically unbiased) and that its standard error decreases as the sample size increases. This type of consistency is known as mean squared error consistency. The Central Limit Theorem implies that the sample mean is a consistent estimator of μ since, as the sample size increases, the distribution of the sample mean is centered at μ with a variance of σ^2/n. Therefore, as $n \to \infty$, the variance approaches zero and the sample mean collapses on μ.

Efficient estimator Estimator that has the smallest variance among *all* unbiased estimators for a parameter.

Empirical Rule of 4 $\left(\frac{X_{Max} - X_{Min}}{4}\right)^2$. This estimator of the variance is biased and is not consistent. Since it is biased it cannot be efficient. It is sensitive to outliers. It is based on the fact that 95.44 percent of the observations lie within 2 standard deviations of the mean in a normal distribution.

Empirical Rule of 6 $\left(\frac{X_{Max} - X_{Min}}{6}\right)^2$. This estimator of the variance is biased and is not consistent. Since it is biased it cannot be efficient. It is sensitive to outliers. It is based on the fact that 99.73 percent of the observations lie within 3 standard deviations of the mean in a normal distribution.

Empirical sampling distribution The empirical distribution created by applying the formula for an estimator to m different samples and producing a histogram of the m estimates.

Estimate Value of an estimator that results from applying its formula to sample data.

Estimator Any sample statistic that is used to estimate a population parameter of interest.

Known distributions A distribution that has been studied by statisticians and has been given a name. Known distributions that can be used in this module are the normal, Student's t, F, chi square, uniform, beta, gamma, exponential, and triangular.

Mean of sampling distribution The theoretical mean of a sampling distribution or the average of the estimates from an empirical sampling distribution.

Mean square deviation $\sum_{i=1}^{n} \frac{\left(x_i - \overline{X}\right)^2}{n}$. This estimator of the variance is biased but asymptotically unbiased (the bias decreases as n becomes large). It is consistent. It has slightly less spread than the sample variance, but since it is biased it is not efficient. It is sensitive to non-normal populations. See **Sample MSD**.

Normal distribution Standard bell-shaped or Gaussian distribution. It has two parameters called the mean μ and variance σ^2.

Parameter Population characteristic that determines the probability distribution of a particular population distribution.

Population distribution The probability density function f(x) that is defined over a set of values of X. The ordinate shows the probability associated with each X value. For a continuous random variable, the area under the entire distribution is 1.

Properties Desirable statistical attributes of an estimator that help a statistician evaluate a variety of estimators for an unknown parameter. See **Asymptotic properties**.

Relatively efficient estimator Unbiased estimator with the smallest variance. Suppose we have two unbiased estimators A and B. For a given sample size, if the variance of estimator A is smaller than the variance of estimator B, then estimator A is relatively more efficient than estimator B.

Sample mean $\overline{X} = \frac{1}{n}\sum_{i=1}^{n} x_i$ (the average of n sample observations). It is unbiased, consistent, and efficient (no other estimator has a smaller variance). It is sensitive to outliers in a sample. Its theoretical distribution is normal if the population is normal or if n is large enough.

Sample median The middle sample observation (if n is odd) or the average of the two middle observations (if n is even). It is unbiased if a symmetric distribution is being sampled and is consistent but has a larger variance than the sample mean. It is relatively insensitive to outliers.

Sample midrange The average of the smallest and largest sample observations. It is unbiased if a symmetric distribution is being sampled, but it is not consistent and has a larger variance than the sample mean. It is very sensitive to outliers.

Sample MSD $\sum_{i=1}^{n}\frac{\left(x_i-\overline{X}\right)^2}{n}$. This estimator of the variance is biased. However, the bias disappears as n approaches infinity and is therefore asymptotically unbiased. It is consistent. Its distribution is sensitive to non-normal populations. See **Mean squared deviation**.

Sample variance $\sum_{i=1}^{n}\frac{\left(x_i-\overline{X}\right)^2}{n-1}$. This estimator of the variance is unbiased, consistent, and only the bootstrap estimator is more efficient. Its distribution is sensitive to non-normal populations. Its theoretical distribution is chi-square distribution with n – 1 degrees of freedom if the population sampled is normal.

Sampling distribution The theoretical distribution of an estimator derived from a sample, such as the sample mean. According to the Central Limit Theorem, regardless of the population being sampled, the sample mean's sampling distribution has a mean μ and standard deviation $\sigma/\sqrt{n}$. It is normally distributed if the sample size is large enough. See **Empirical sampling distribution**.

Standard deviation Denoted σ, it is the square root of the population variance. It is a measure of dispersion about the mean. Larger σ values indicate greater dispersion.

Standard error The square root of the variance of the sampling distribution. According to the Central Limit Theorem, the theoretical standard error of the sample mean is $\sigma/\sqrt{n}$, where σ is the population standard deviation and n is the sample size. It is generally written as $\sigma_{\overline{X}}$.

Sufficient estimator Estimator that uses all the information from a sample.

Symmetric distribution A distribution whose mirror image looks the same on either side of its mean. The normal, uniform, and Student's t are symmetric distributions. Depending upon the parameter values used, the beta and triangular distributions may also be symmetric.

Unbiased estimator Estimator whose expected value is equal to the parameter it is estimating. It is centered on the true population parameter (neither under- nor over-estimates the parameter). For example, the Central Limit Theorem says that the sample mean is an unbiased estimator of μ. Therefore, the average of a large number of sample means should equal μ.

Variance In a population, the variance is the expected value of $(X - \mu)^2$ and is denoted σ^2. In a sample, the variance is the sum of the squared deviations about the sample mean divided by $n - 1$ and is denoted s^2.

Variance of Sampling Distribution The theoretical variance of a sampling distribution or the variance of the estimates from an empirical sampling distribution.

Solutions to Self-Evaluation Quiz

1. e Do Exercises 9–14. Do Individual Learning Project 1. Read the Overview of Concepts, Illustration of Concepts, and Glossary.
2. d Do Exercises 9–14. Do Individual Learning Project 1. Read the Overview of Concepts, Illustration of Concepts, and Glossary.
3. c Do Exercises 17–19. Do Individual Learning Project 2. Read the Overview of Concepts, Illustration of Concepts, and Glossary.
4. d Read the Overview of Concepts.
5. d Do Exercises 17–19. Do Individual Learning Project 2. Read the Overview of Concepts, Illustration of Concepts, and Glossary.
6. d Do Exercises 15–16. Do Individual Learning Project 2. Read the Overview of Concepts, Illustration of Concepts, and Glossary.
7. e Do Exercises 10–18 and 19–23.
8. c Read the Glossary.
9. e Do Exercises 20 and 21. Do Individual Learning Project 1. Read the Overview of Concepts, Illustration of Concepts, and Glossary.
10. d Do Individual Learning Project 2. Read the Overview of Concepts, Illustration of Concepts, and Glossary.
11. a Do Exercises 24 and 25. Read the Overview of Concepts, Illustration of Concepts, and Glossary.
12. c Do Exercise 26. Read the Overview of Concepts and Glossary.