Sampling Distributions and Estimations

Ellen Lawsky, (EllenL)
Brenda Meery, (BrendaM)

AUTHORS
Ellen Lawsky, (EllenL)
Brenda Meery, (BrendaM)

To access a customizable version of this book, as well as other interactive content, visit www.ck12.org

CK-12 Foundation is a non-profit organization with a mission to reduce the cost of textbook materials for the K-12 market both in the U.S. and worldwide. Using an open-content, web-based collaborative model termed the **FlexBook®**, CK-12 intends to pioneer the generation and distribution of high-quality educational content that will serve both as core text as well as provide an adaptive environment for learning, powered through the **FlexBook Platform®**.

Printed: September 14, 2013

flexbook
next generation textbooks

CHAPTER 1 Sampling Distributions and Estimations

Chapter Outline

1.1 Sampling Distribution

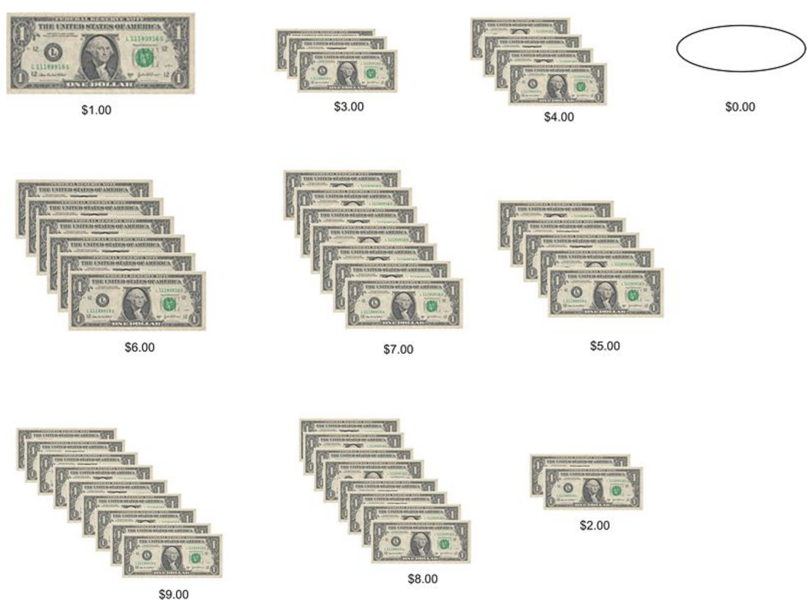
Learning Objectives

- Understand the inferential relationship between a sampling distribution and a population parameter.
- Graph a frequency distribution of sample means using a data set.
- Understand the relationship between sample size and the distribution of sample means.
- Understand sampling error.

Introduction

Have you ever wondered how the mean, or average, amount of money per person in a population is determined? It would be impossible to contact 100% of the population, so there must be a statistical way to estimate the mean number of dollars per person in the population.

Suppose, more simply, that we are interested in the mean number of dollars that are in each of the pockets of ten people on a busy street corner. The diagram below reveals the amount of money that each person in the group of ten has in his/her pocket. We will investigate this scenario later in the lesson.



Sampling Distributions

In previous chapters, you have examined methods that are good for the exploration and description of data. In this section, we will discuss how collecting data by random sampling helps us to draw more rigorous conclusions about the data.

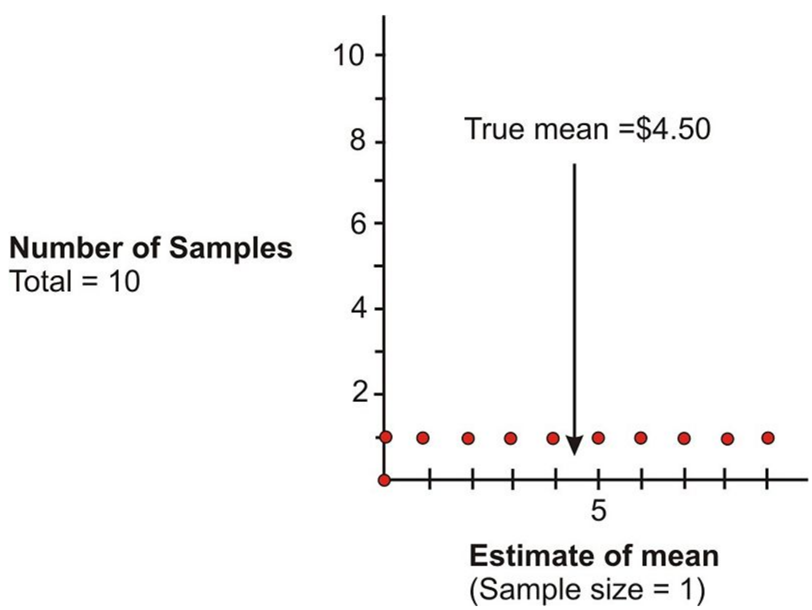
The purpose of sampling is to select a set of units, or elements, from a population that we can use to estimate the parameters of the population. Random sampling is one special type of probability sampling. Random sampling erases the danger of a researcher consciously or unconsciously introducing bias when selecting a sample. In addition, random sampling allows us to use tools from probability theory that provide the basis for estimating the characteristics of the population, as well as for estimating the accuracy of the samples.

Probability theory is the branch of mathematics that provides the tools researchers need to make statistical conclusions about sets of data based on samples. As previously stated, it also helps statisticians estimate the parameters of a population. A *parameter* is a summary description of a given variable in a population. A population mean is an example of a parameter. When researchers generalize from a sample, they're using sample observations to estimate population parameters. Probability theory enables them to both make these estimates and to judge how likely it is that the estimates accurately represent the actual parameters of the population.

Probability theory accomplishes this by way of the concept of *sampling distributions*. A single sample selected from a population will give an estimate of the population parameters. Other samples would give the same, or slightly different, estimates. Probability theory helps us understand how to make estimates of the actual population parameters based on such samples.

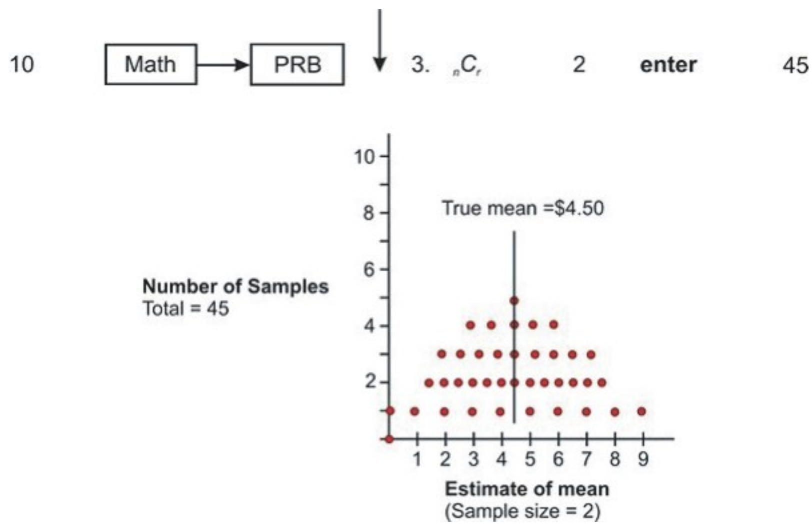
In the scenario that was presented in the introduction to this lesson, the assumption was made that in the case of a population of size ten, one person had no money, another had \$1.00, another had \$2.00, and so on. until we reached the person who had \$9.00.

The purpose of the task was to determine the average amount of money per person in this population. If you total the money of the ten people, you will find that the sum is \$45.00, thus yielding a mean of \$4.50. However, suppose you couldn't count the money of all ten people at once. In this case, to complete the task of determining the mean number of dollars per person of this population, it is necessary to select random samples from the population and to use the means of these samples to estimate the mean of the whole population. To start, suppose you were to randomly select a sample of only one person from the ten. The ten possible samples are represented in the diagram in the introduction, which shows the dollar bills possessed by each sample. Since samples of one are being taken, they also represent the means you would get as estimates of the population. The graph below shows the results:



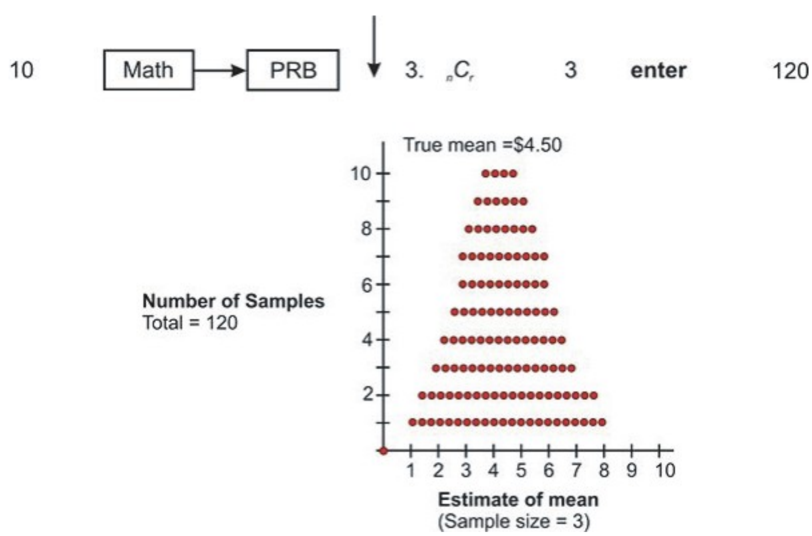
The distribution of the dots on the graph is an example of a sampling distribution. As can be seen, selecting a sample of one is not very good, since the group's mean can be estimated to be anywhere from \$0.00 to \$9.00, and the true mean of \$4.50 could be missed by quite a bit.

What happens if we take samples of two? From a population of 10, in how many ways can two be selected if the order of the two does not matter? The answer, which is 45, can be found by using a graphing calculator as shown in the figure below. When selecting samples of size two from the population, the sampling distribution is as follows:

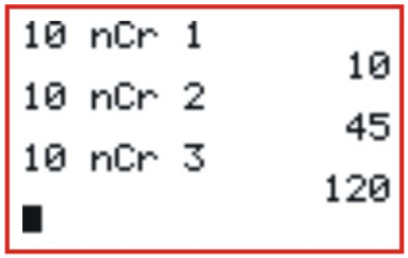


Increasing the sample size has improved your estimates. There are now 45 possible samples, such as ($0, $1), ($0, $2), ($7, $8), ($8, $9), and so on, and some of these samples produce the same means. For example, ($0, $6), ($1, $5), and ($2, $4) all produce means of $3. The three dots above the mean of 3 represent these three samples. In addition, the 45 means are not evenly distributed, as they were when the sample size was one. Instead, they are more clustered around the true mean of $4.50. ($0, $1) and ($8, $9) are the only two samples whose means deviate by as much as $4.00. Also, five of the samples yield the true estimate of $4.50, and another eight deviate by only plus or minus 50 cents.

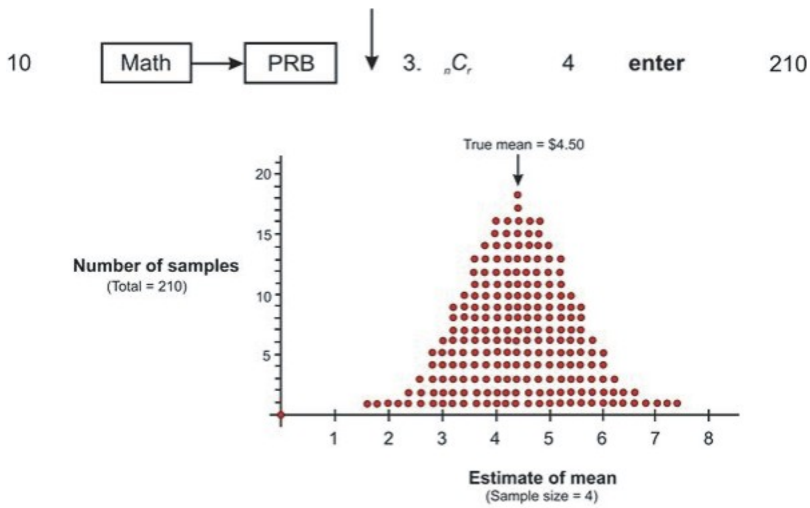
If three people are randomly selected from the population of 10 for each sample, there are 120 possible samples, which can be calculated with a graphing calculator as shown below. The sampling distribution in this case is as follows:

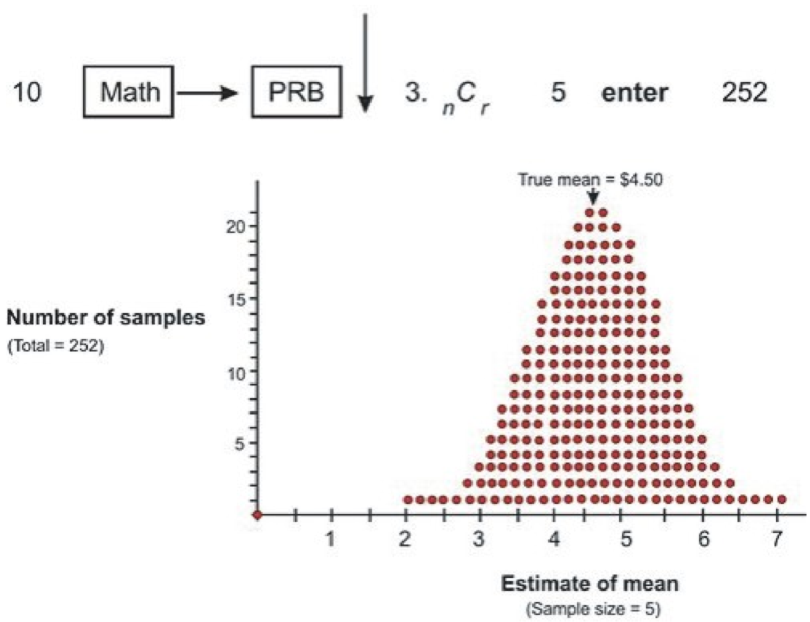


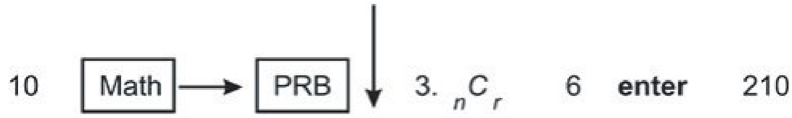
Here are screen shots from a graphing calculator for the results of randomly selecting 1, 2, and 3 people from the population of 10. The 10, 45, and 120 represent the total number of possible samples that are generated by increasing the sample size by 1 each time.

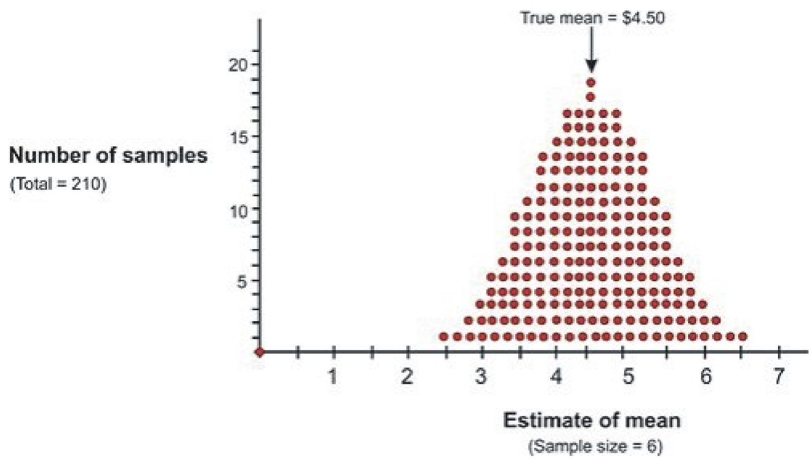


Next, the sampling distributions for sample sizes of 4, 5, and 6 are shown:





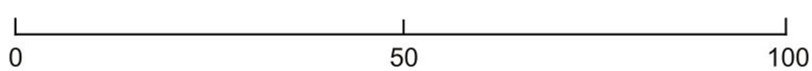




From the graphs above, it is obvious that increasing the size of the samples chosen from the population of size 10 resulted in a distribution of the means that was more closely clustered around the true mean. If a sample of size 10 were selected, there would be only one possible sample, and it would yield the true mean of $4.50. Also, the sampling distribution of the *sample means* is approximately normal, as can be seen by the bell shape in each of the graphs.

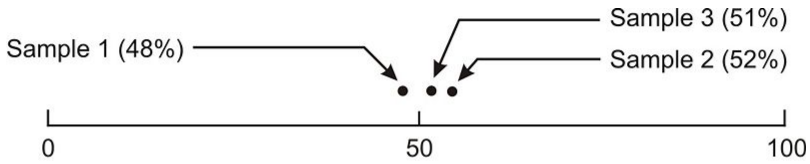
Now that you have been introduced to sampling distributions and how the sample size affects the distribution of the sample means, it is time to investigate a more realistic sampling situation. Assume you want to study the student population of a university to determine approval or disapproval of a student dress code proposed by the administration. The study's population will be the 18,000 students who attend the school, and the elements will be the individual students. A random sample of 100 students will be selected for the purpose of estimating the opinion of the entire student body, and attitudes toward the dress code will be the variable under consideration. For simplicity's sake, assume that the attitude variable has two variations: approve and disapprove. As you know from the last chapter, a scenario such as this in which a variable has two attributes is called binomial.

The following figure shows the range of possible sample study results. It presents all possible values of the parameter in question by representing a range of 0 percent to 100 percent of students approving of the dress code. The number 50 represents the midpoint, or 50 percent of the students approving of the dress code and 50 percent disapproving. Since the sample size is 100, at the midpoint, half of the students would be approving of the dress code, and the other half would be disapproving.



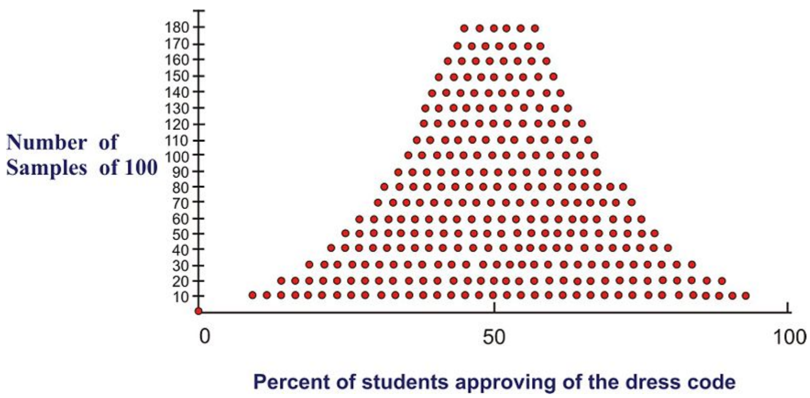
Percent of students approving of the dress code

To randomly select the sample of 100 students, every student is presented with a number from 1 to 18,000, and the sample is randomly chosen from a drum containing all of the numbers. Each member of the sample is then asked whether he or she approves or disapproves of the dress code. If this procedure gives 48 students who approve of the dress code and 52 who disapprove, the result would be recorded on the figure by placing a dot at 48%. This statistic is the *sample proportion*. Let's assume that the process was repeated, and it resulted in 52 students approving of the dress code. Let's also assume that a third sample of 100 resulted in 51 students approving of the dress code. The results are shown in the figure below.



Percent of students approving of the dress code

In this figure, the three different sample statistics representing the percentages of students who approved of the dress code are shown. The three random samples chosen from the population give estimates of the parameter that exists for the entire population. In particular, each of the random samples gives an estimate of the percentage of students in the total student body of 18,000 who approve of the dress code. Assume for simplicity's sake that the true proportion for the population is 50%. This would mean that the estimates are close to the true proportion. To more precisely estimate the true proportion, it would be necessary to continue choosing samples of 100 students and to record all of the results in a summary graph as shown:



Notice that the statistics resulting from the samples are distributed around the population parameter. Although there is a wide range of estimates, most of them lie close to the 50% area of the graph. Therefore, the true value is likely to be in the vicinity of 50%. In addition, probability theory gives a formula for estimating how closely the sample statistics are clustered around the true value. In other words, it is possible to estimate the sampling error, or the degree of error expected for a given sample design. The formula $s = \sqrt{\frac{p(1-p)}{n}}$ contains three variables: the parameter, p, the sample size, n, and the *standard error*, s.

The symbols p and $1-p$ in the formula represent the population parameters. For example, if 60 percent of the student body approves of the dress code and 40% disapproves, p and $1-p$ would be 0.6 and 0.4,, respectively. The square root of the product of p and $1-p$ is the population standard deviation. As previously stated, the symbol n represents the number of cases in each sample, and s is the standard error.

If the assumption is made that the true population parameters are 0.50 approving of the dress code and 0.50 disapproving of the dress code, when selecting samples of 100, the standard error obtained from the formula equals 0.05:

$$s = \sqrt{\frac{(0.5)(0.5)}{100}} = 0.05$$

This calculation indicates how tightly the sample estimates are distributed around the population parameter. In this case, the standard error is the standard deviation of the sampling distribution.

The Empirical Rule states that certain proportions of the sample estimates will fall within defined increments, each increment being one standard error from the population parameter. According to this rule, 34% of the sample estimates will fall within one standard error above the population parameter, and another 34% will fall within one standard error below the population parameter. In the above example, you have calculated the standard error to be 0.05, so you know that 34% of the samples will yield estimates of student approval between 0.50 (the population parameter) and 0.55 (one standard error above the population parameter). Likewise, another 34% of the samples will give estimates between 0.5 and 0.45 (one standard error below the population parameter). Therefore, you know that 68% of the samples will give estimates between 0.45 and 0.55. In addition, probability theory says that 95% of the samples will fall within two standard errors of the true value, and 99.7% will fall within three standard errors. In this example, you can say that only three samples out of one thousand would give an estimate of student approval below 0.35 or above 0.65.

The size of the standard error is a function of the population parameter. By looking at the formula $s = \sqrt{\frac{p(1-p)}{n}}$, it is obvious that the standard error will increase as the quantity $p\ (1-p)$ increases. Referring back to our example, the maximum for this product occurred when there was an even split in the population. When $p = 0.5,\ p(1-p) = (0.5)(0.5) = 0.25$. If $p = 0.6$, then $p(1-p) = (0.6)(0.4) = 0.24$. Likewise, if $p = 0.8$, then $p(1-p) = (0.8)(0.2) = 0.16$. If p were either 0 or 1 (none or all of the student body approves of the dress code), then the standard error would be 0. This means that there would be no variation, and every sample would give the same estimate.

The standard error is also a function of the sample size. In other words, as the sample size increases, the standard error decreases, or the bigger the sample size, the more closely the samples will be clustered around the true value. Therefore, this is an inverse relationship. The last point about that formula that is obvious is emphasized by the square root operation. That is, the standard error will be reduced by one-half as the sample size is quadrupled.

On the Web

http://tinyurl.com/294stkw Explore the result of changing the population parameter, the sample size, and the number of samples taken for the proportion of Reese's Pieces that are brown or yellow.

Lesson Summary

In this lesson, we have learned about probability sampling, which is the key sampling method used in survey research. In the example presented above, the elements were chosen for study from a population by random sampling. The sample size had a direct effect on the distribution of estimates of the population parameter. The larger the sample size, the closer the sampling distribution was to a normal distribution.

Points to Consider

- Does the mean of the sampling distribution equal the mean of the population?
- If the sampling distribution is normally distributed, is the population normally distributed?
- Are there any restrictions on the size of the sample that is used to estimate the parameters of a population?
- Are there any other components of sampling error estimates?

Review Questions

The following activity could be done in the classroom, with the students working in pairs or small groups. Before doing the activity, students could put their pennies into a jar and save them as a class, with the teacher also contributing. In a class of 30 students, groups of 5 students could work together, and the various tasks could be divided among those in each group.

1. If you had 100 pennies and were asked to record the age of each penny, predict the shape of the distribution. (The age of a penny is the current year minus the date on the coin.)
2. Construct a histogram of the ages of the pennies.
3. Calculate the mean of the ages of the pennies.

Have each student in each group randomly select a sample of 5 pennies from the 100 coins and calculate the mean of the five ages of the coins chosen. Have the students then record their means on a number line. Have the students repeat this process until all of the coins have been chosen.

4. How does the mean of the samples compare to the mean of the population (100 ages)? Repeat step 4 using a sample size of 10 pennies. (As before, allow the students to work in groups.)
5. What is happening to the shape of the sampling distribution of the sample means as the sample size increases?

1.2 The z-Score and the Central Limit Theorem

Learning Objectives

- Understand the Central Limit Theorem and calculate a sampling distribution using the mean and standard deviation of a normally distributed random variable.
- Understand the relationship between the Central Limit Theorem and the normal approximation of a sampling distribution.

Introduction

In the previous lesson, you learned that sampling is an important tool for determining the characteristics of a population. Although the parameters of the population (mean, standard deviation, etc.) were unknown, random sampling was used to yield reliable estimates of these values. The estimates were plotted on graphs to provide a visual representation of the distribution of the sample means for various sample sizes. It is now time to define some properties of a sampling distribution of sample means and to examine what we can conclude about the entire population based on these properties.

Central Limit Theorem

The *Central Limit Theorem* is a very important theorem in statistics. It basically confirms what might be an intuitive truth to you: that as you increase the sample size for a random variable, the distribution of the sample means better approximates a normal distribution.

Before going any further, you should become familiar with (or reacquaint yourself with) the symbols that are commonly used when dealing with properties of the sampling distribution of sample means. These symbols are shown in the table below:

TABLE 1.1:

	Population Parameter	**Sample Statistic**	**Sampling Distribution**
Mean	μ	$\bar{x}$	$\mu_{\bar{x}}$
Standard Deviation	σ	s	$S_{\bar{x}}$ or $\sigma_{\bar{x}}$
Size	N	n	

As the sample size, n, increases, the resulting sampling distribution would approach a normal distribution with the same mean as the population and with $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$. The notation $\sigma_{\bar{x}}$ reminds you that this is the standard deviation of the distribution of sample means and not the standard deviation of a single observation.

The Central Limit Theorem states the following:

If samples of size n are drawn at random from any population with a finite mean and standard deviation, then the sampling distribution of the sample means, $\bar{x}$, approximates a normal distribution as n increases.

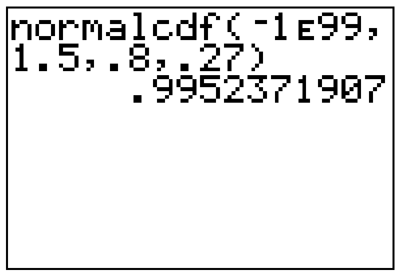
The mean of this sampling distribution approximates the population mean, and the standard deviation of this sampling distribution approximates the standard deviation of the population divided by the square root of the sample size: $\mu_{\bar{x}} = \mu$ and $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$.

These properties of the sampling distribution of sample means can be applied to determining probabilities. If the sample size is sufficiently large (> 30), the sampling distribution of sample means can be assumed to be approximately normal, even if the population is not normally distributed.

Example: Suppose you wanted to answer the question, "What is the probability that a random sample of 20 families in Canada will have an average of 1.5 pets or fewer?" where the mean of the population is 0.8 and the standard deviation of the population is 1.2.

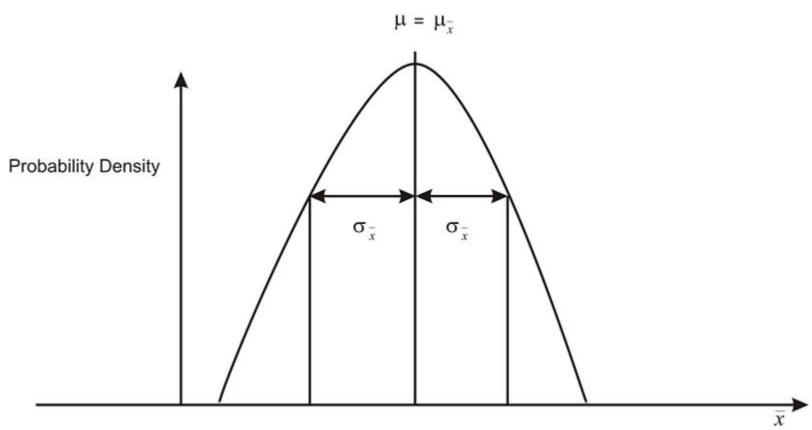
For the sampling distribution, $\mu_{\bar{x}} = \mu = 0.8$ and $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} = \frac{1.2}{\sqrt{20}} = 0.268$.

We can use a graphing calculator as follows:



Therefore, the probability that the sample mean will be below 1.5 is 0.9952. In other words, with a random sample of 20 families, it is almost definite that the average number of pets per family will be less than 1.5.

The properties associated with the Central Limit Theorem are displayed in the diagram below:



The vertical axis now reads probability density, rather than frequency, since frequency can only be used when you are dealing with a finite number of sample means. Sampling distributions, on the other hand, are theoretical depictions of an infinite number of sample means, and probability density is the relative density of the selections from within this set.

Example: A random sample of size 40 is selected from a known population with a mean of 23.5 and a standard deviation of 4.3. Samples of the same size are repeatedly collected, allowing a sampling distribution of sample means to be drawn.

a) What is the expected shape of the resulting distribution?

b) Where is the sampling distribution of sample means centered?

c) What is the approximate standard deviation of the sample means?

The question indicates that multiple samples of size 40 are being collected from a known population, multiple sample means are being calculated, and then the sampling distribution of the sample means is being studied. Therefore, an understanding of the Central Limit Theorem is necessary to answer the question.

a) The sampling distribution of the sample means will be approximately bell-shaped.

b) The sampling distribution of the sample means will be centered about the population mean of 23.5.

c) The approximate standard deviation of the sample means is 0.68, which can be calculated as shown below:

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$$
$$\sigma_{\bar{x}} = \frac{4.3}{\sqrt{40}}$$
$$\sigma_{\bar{x}} = 0.68$$

Example: Multiple samples with a sample size of 40 are taken from a known population, where $\mu = 25$ and $\sigma = 4$. The following chart displays the sample means:

25	25	26	26	26	24	25	25	24	25
26	25	26	25	24	25	25	25	25	25
24	24	24	24	26	26	26	25	25	25
25	25	24	24	25	25	25	24	25	25
25	24	25	25	24	26	24	26	24	26

a) What is the population mean?

b) Using technology, determine the mean of the sample means.

c) What is the population standard deviation?

d) Using technology, determine the standard deviation of the sample means.

e) As the sample size increases, what value will the mean of the sample means approach?

f) As the sample size increases, what value will the the standard deviation of the sample means approach?

a) The population mean of 25 was given in the question: $\mu = 25$.

b) The mean of the sample means is 24.94 and is determined by using '1 Vars Stat' on the TI-83/84 calculator: $\mu_{\bar{x}} = 24.94$.

c) The population standard deviation of 4 was given in the question: $\sigma = 4$.

d) The standard deviation of the sample means is 0.71 and is determined by using '1 Vars Stat' on the TI-83/84 calculator: $S_{\bar{x}} = 0.71$. Note that the Central Limit Theorem states that the standard deviation should be approximately $\frac{4}{\sqrt{40}} = 0.63$.

e) The mean of the sample means will approach 25 and is determined by a property of the Central Limit Theorem: $\mu_{\bar{x}} = 25$.

f) The standard deviation of the sample means will approach $\frac{4}{\sqrt{n}}$ and is determined by a property of the Central Limit Theorem: $\sigma_{\bar{x}} = \frac{4}{\sqrt{n}}$.

On the Web

http://tinyurl.com/2f969wj Explore how the sample size and the number of samples affect the mean and standard deviation of the distribution of sample means.

Lesson Summary

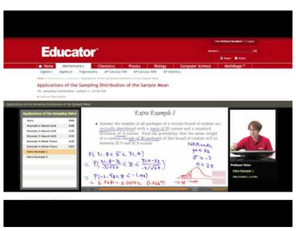
The Central Limit Theorem confirms the intuitive notion that as the sample size increases for a random variable, the distribution of the sample means will begin to approximate a normal distribution, with the mean equal to the mean of the underlying population and the standard deviation equal to the standard deviation of the population divided by the square root of the sample size, n.

Point to Consider

- How does sample size affect the variation in sample results?

Multimedia Links

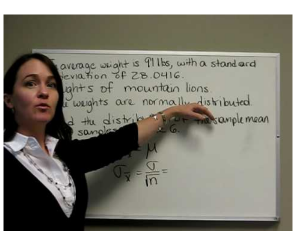
For an example using the sampling distribution of x-bar **(15.0)(16.0)**, see EducatorVids, Statistics: Sampling Distribution of the Sample Mean (2:15).



MEDIA

Click image to the left for more content.

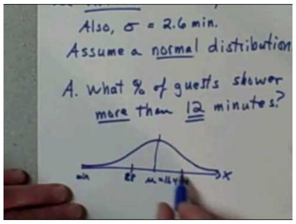
For another example of the sampling distribution of x-bar **(15.0)(16.0)**, see tcreelmuw, Distribution of Sample Mean (2:22).



MEDIA

Click image to the left for more content.

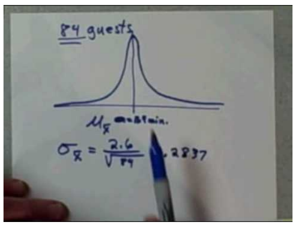
For an example of using the Central Limit Theorem **(9.0)**, see jsnider3675,Applicationof the Central Limit Theorem, Part 1 (5:44).



MEDIA

Click image to the left for more content.

For the continuation of an example using the Central Limit Theorem **(9.0)**, see jsnider3675,Applicationof the Central Limit Theorem, Part 2 (6:38).



MEDIA

Click image to the left for more content.

Review Questions

1. A random sample of size 30 is selected from a known population with a mean of 13.2 and a standard deviation of 2.1. Samples of the same size are repeatedly collected, allowing a sampling distribution of sample means to be drawn.
   a. What is the expected shape of the resulting distribution?
   b. Where is the sampling distribution of sample means centered?
   c. What is the approximate standard deviation of the sample means?
2. What is the probability that a random sample of 40 families will have an average of 0.5 pets or fewer where the mean of the population is 0.8 and the standard deviation of the population is 1.2?
3. The scores of students on a college entrance exam were normally distributed with a mean of 19.4 and a standard deviation of 6.3.
   a. If a sample of 70 students who took the test (who have the same distribution as all scores) is collected, what are the mean and standard deviation of the sample mean for the 70 students?
   b. What is the probability that a random sample of 50 students will have an average score of 22 or higher?
4. The lifetimes of a certain type of calculator battery are normally distributed. The mean lifetime is 400 days, with a standard deviation of 50 days. For a sample of 6000 new batteries, determine how many batteries will last:
   a. between 360 and 460 days.
   b. more than 320 days.
   c. less than 280 days.

1.3 Confidence Intervals

Learning Objectives

- Calculate the mean of a sample as a point estimate of the population mean.
- Construct a confidence interval for a population mean based on a sample mean.
- Calculate a sample proportion as a point estimate of the population proportion.
- Construct a confidence interval for a population proportion based on a sample proportion.
- Calculate the margin of error for a point estimate as a function of sample mean or proportion and size.
- Understand the logic of confidence intervals, as well as the meaning of confidence level and confidence intervals.

Introduction

The objective of inferential statistics is to use sample data to increase knowledge about the entire population. In this lesson, we will examine how to use samples to make estimates about the populations from which they came. We will also see how to determine how wide these estimates should be and how confident we should be about them.

Confidence Intervals

Sampling distributions are the connecting link between the collection of data by unbiased random sampling and the process of drawing conclusions from the collected data. Results obtained from a survey can be reported as a *point estimate*. For example, a single sample mean is a point estimate, because this single number is used as a plausible value of the population mean. Keep in mind that some error is associated with this estimate-the true population mean may be larger or smaller than the sample mean. An alternative to reporting a point estimate is identifying a range of possible values the parameter might take, controlling the probability that the parameter is not lower than the lowest value in this range and not higher than the largest value. This range of possible values is known as a *confidence interval*. Associated with each confidence interval is a *confidence level*. This level indicates the level of assurance you have that the resulting confidence interval encloses the unknown population mean.

In a normal distribution, we know that 95% of the data will fall within two standard deviations of the mean. Another way of stating this is to say that we are confident that in 95% of samples taken, the sample statistics are within plus or minus two standard errors of the population parameter. As the confidence interval for a given statistic increases in length, the confidence level increases.

The selection of a confidence level for an interval determines the probability that the confidence interval produced will contain the true parameter value. Common choices for the confidence level are 90%, 95%, and 99%. These levels correspond to percentages of the area under the normal density curve. For example, a 95% confidence interval covers 95% of the normal curve, so the probability of observing a value outside of this area is less than 5%. Because the normal curve is symmetric, half of the 5% is in the left tail of the curve, and the other half is in the right tail of the curve. This means that 2.5% is in each tail.

The graph shown above was made using a TI-83 graphing calculator and shows a normal distribution curve for a set of data for which $\mu = 50$ and $\sigma = 12$. A 95% confidence interval for the standard normal distribution, then, is the interval $(-1.96, 1.96)$, since 95% of the area under the curve falls within this interval. The ± 1.96 are the z-scores that enclose the given area under the curve. For a normal distribution, the *margin of error* is the amount that is added to and subtracted from the mean to construct the confidence interval. For a 95% confidence interval, the margin of error is 1.96σ. (Note that previously we said that 95% of the data in a normal distribution falls within ± 2 standard deviations of the mean. This was just an estimate, and for the remainder of this textbook, we'll assume that 95% of the data actually falls within ± 1.96 standard deviations of the mean.)

The following is the derivation of the confidence interval for the population mean, μ. In it, $z_{\frac{\alpha}{2}}$ refers to the positive z-score for a particular confidence interval. The Central Limit Theorem tells us that the distribution of $\bar{x}$ is normal, with a mean of μ and a standard deviation of $\frac{\sigma}{\sqrt{n}}$. Consider the following:

$$-z_{\frac{\alpha}{2}} < \frac{\bar{x}-\mu}{\frac{\sigma}{\sqrt{n}}} < z_{\frac{\alpha}{2}}$$

All values are known except for μ. Solving for this parameter, we have:

$$\begin{aligned}
-\bar{x} - z_{\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}} &< -\mu < z_{\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}} - \bar{x}\\
\bar{x} + z_{\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}} &> \mu > -z_{\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}} + \bar{x}\\
\bar{x} + z_{\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}} &> \mu > \bar{x} - z_{\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}\\
\bar{x} - z_{\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}} &< \mu < \bar{x} + z_{\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}
\end{aligned}$$

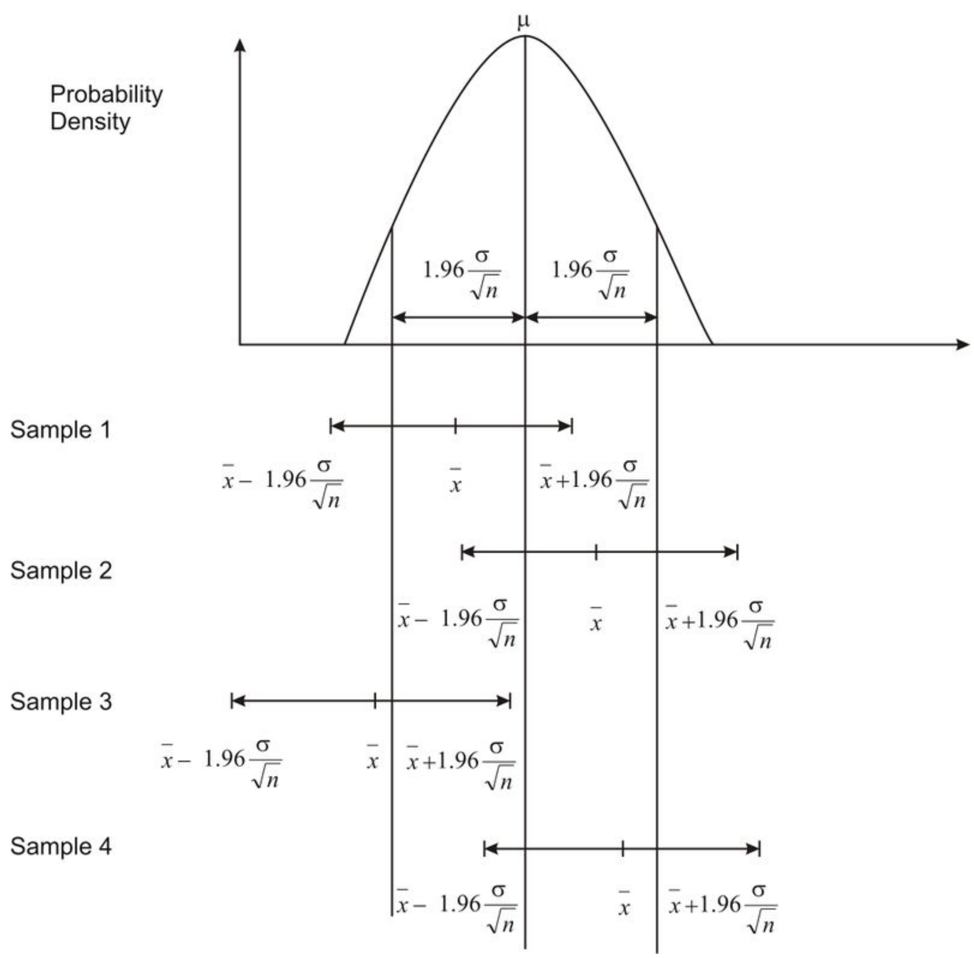
Another way to express this is: $\bar{x} \pm z_{\frac{\alpha}{2}}\left(\frac{\sigma}{\sqrt{n}}\right)$.

On the Web

http://tinyurl.com/27syj3x This simulates confidence intervals for the mean of the population.

Example: Jenny randomly selected 60 muffins of a particular brand and had those muffins analyzed for the number of grams of fat that they each contained. Rather than reporting the sample mean (point estimate), she reported the confidence interval. Jenny reported that the number of grams of fat in each muffin is between 10.3 grams and 12.1 grams with 95% confidence.

In this example, the population mean is unknown. This number is fixed, not variable, and the sample means are variable, because the samples are random. If this is the case, does the confidence interval enclose this unknown true mean? Random samples lead to the formation of confidence intervals, some of which contain the fixed population mean and some of which do not. The most common mistake made by persons interpreting a confidence interval is claiming that once the interval has been constructed, there is a 95% probability that the population mean is found within the confidence interval. Even though the population mean is unknown, once the confidence interval is constructed, either the mean is within the confidence interval, or it is not. Hence, any probability statement about this particular confidence interval is inappropriate. In the above example, the confidence interval is from 10.3 to 12.1, and Jenny is using a 95% confidence level. The appropriate statement should refer to the method used to produce the confidence interval. Jenny should have stated that the method that produced the interval from 10.3 to 12.1 has a 0.95 probability of enclosing the population mean. This means if she did this procedure 100 times, 95 of the intervals produced would contain the population mean. The probability is attributed to the method, not to any particular confidence interval. The following diagram demonstrates how the confidence interval provides a range of plausible values for the population mean and that this interval may or may not capture the true population mean. If you formed 100 intervals in this manner, 95 of them would contain the population mean.



Example: The following questions are to be answered with reference to the above diagram.

a) Were all four sample means within $1.96\frac{\sigma}{\sqrt{n}}$, or $1.96\sigma_{\bar{x}}$, of the population mean? Explain.

b) Did all four confidence intervals capture the population mean? Explain.

c) In general, what percentage of $\bar{x}'s$ should be within $1.96\frac{\sigma}{\sqrt{n}}$ of the population mean?

d) In general, what percentage of the confidence intervals should contain the population mean?

a) The sample mean, $\bar{x}$, for Sample 3 was not within $1.96\frac{\sigma}{\sqrt{n}}$ of the population mean. It did not fall within the vertical lines to the left and right of the population mean.

b) The confidence interval for Sample 3 did not enclose the population mean. This interval was just to the left of the population mean, which is denoted with the vertical line found in the middle of the sampling distribution of the sample means.

c) 95%

d) 95%

When the sample size is large ($n > 30$), the confidence interval for the population mean is calculated as shown below:

$\bar{x} \pm z_{\frac{\alpha}{2}}\left(\frac{\sigma}{\sqrt{n}}\right)$, where $z_{\frac{\alpha}{2}}$ is 1.96 for a 95% confidence interval, 1.645 for a 90% confidence interval, and 2.58 for a 99% confidence interval.

Example: Julianne collects four samples of size 60 from a known population with a population standard deviation of 19 and a population mean of 110. Using the four samples, she calculates the four sample means to be:

107 112 109 115

a) For each sample, determine the 90% confidence interval.

b) Do all four confidence intervals enclose the population mean? Explain.

a)

$$\bar{x} \pm z\frac{\sigma}{\sqrt{n}}$$
$$107 \pm (1.645)(\frac{19}{\sqrt{60}})$$
$$107 \pm 4.04$$
from 102.96 to 111.04

$$\bar{x} \pm z\frac{\sigma}{\sqrt{n}}$$
$$112 \pm (1.645)(\frac{19}{\sqrt{60}})$$
$$112 \pm 4.04$$
from 107.96 to 116.04

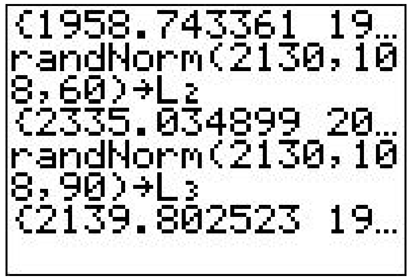
$$\bar{x} \pm z\frac{\sigma}{\sqrt{n}}$$
$$109 \pm (1.645)(\frac{19}{\sqrt{60}})$$
$$109 \pm 4.04$$
from 104.96 to 113.04

$$\bar{x} \pm z\frac{\sigma}{\sqrt{n}}$$
$$115 \pm (1.645)(\frac{19}{\sqrt{60}})$$
$$115 \pm 4.04$$
from 110.96 to 119.04

b) Three of the confidence intervals enclose the population mean. The interval from 110.96 to 119.04 does not enclose the population mean.

Technology Note: Simulation of Random Samples and Formation of Confidence Intervals on the TI-83/84 Calculator

Now it is time to use a graphing calculator to simulate the collection of three samples of sizes 30, 60, and 90, respectively. The three sample means will be calculated, as well as the three 95% confidence intervals. The samples will be collected from a population that displays a normal distribution, with a population standard deviation of 108 and a population mean of 2130. We will use the **randNorm(** function found in **[MATH]**, under the **PRB** menu. First, store the three samples in **L1**, **L2**, and **L3**, respectively, as shown below:



Store 'randNorm(μ,σ,n)' in **L1**. The sample size is $n = 30$.

Store 'randNorm(μ,σ,n)' in **L2**. The sample size is $n = 60$.

Store 'randNorm(μ,σ,n)' in **L3**. The sample size is $n = 90$.

The lists of numbers can be viewed by pressing **[STAT][ENTER]**. The next step is to calculate the mean of each of these samples.

To do this, first press **[2ND][LIST]** and go to the **MATH** menu. Next, select the 'mean(' command and press **[2ND][L1][ENTER]**. Repeat this process for **L2** and **L3**.

Note that your confidence intervals will be different than the ones calculated below, because the random numbers generated by your calculator will be different, and thus, your means will be different. For us, the means of **L1**, **L2**, and **L3** were 2139.1, 2119.2, and 2137.1, respectively, so the confidence intervals are as follows:

$$\bar{x} \pm z\frac{\sigma}{\sqrt{n}}$$
$$2139.1 \pm (1.96)(\frac{108}{\sqrt{30}})$$
$$2139.1 \pm 38.65$$
from 2100.45 to 2177.65

$$\bar{x} \pm z\frac{\sigma}{\sqrt{n}}$$
$$2119.2 \pm (1.96)(\frac{108}{\sqrt{60}})$$
$$2119.2 \pm 27.33$$
from 2091.87 to 2146.53

$$\bar{x} \pm z\frac{\sigma}{\sqrt{n}}$$
$$2137.1 \pm (1.96)(\frac{108}{\sqrt{90}})$$
$$2137.1 \pm 22.31$$
from 2114.79 to 2159.41

As was expected, the value of $\bar{x}$ varied from one sample to the next. The other fact that was evident was that as the sample size increased, the length of the confidence interval became smaller, or decreased. This is because with the increase in sample size, you have more information, and thus, your estimate is more accurate, which leads to a narrower confidence interval.

In all of the examples shown above, you calculated the confidence intervals for the population mean using the formula $\bar{x} \pm z_{\frac{\alpha}{2}}\left(\frac{\sigma}{\sqrt{n}}\right)$. However, to use this formula, the population standard deviation σ had to be known. If this value is unknown, and if the sample size is large $(n > 30)$, the population standard deviation can be replaced with the sample standard deviation. Thus, the formula $\bar{x} \pm z_{\frac{\alpha}{2}}\left(\frac{s_x}{\sqrt{n}}\right)$ can be used as an interval estimator, or confidence interval. This formula is valid only for simple random samples. Since $z_{\frac{\alpha}{2}}\left(\frac{s_x}{\sqrt{n}}\right)$ is the margin of error, a confidence interval can be thought of simply as: $\bar{x}\pm$ the margin of error.

Example: A committee set up to field-test questions from a provincial exam randomly selected grade 12 students to answer the test questions. The answers were graded, and the sample mean and sample standard deviation were calculated. Based on the results, the committee predicted that on the same exam, 9 times out of 10, grade 12 students would have an average score of within 3% of 65%.

a) Are you dealing with a 90%, 95%, or 99% confidence level?

b) What is the margin of error?

c) Calculate the confidence interval.

d) Explain the meaning of the confidence interval.

a) You are dealing with a 90% confidence level. This is indicated by 9 times out of 10.

b) The margin of error is 3%.

c) The confidence interval is $\bar{x}\pm$ the margin of error, or 62% to 68%.

d) There is a 0.90 probability that the method used to produce this interval from 62% to 68% results in a confidence interval that encloses the population mean (the true score for this provincial exam).

Confidence Intervals for Hypotheses about Population Proportions

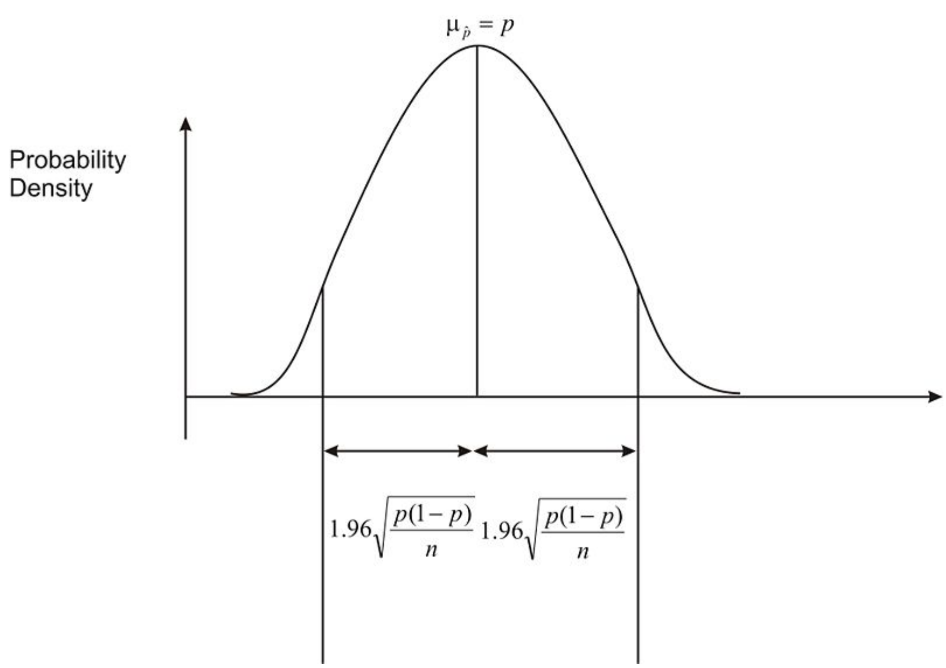
Often statisticians are interested in making inferences about a population proportion. For example, when we look at election results we often look at the proportion of people that vote and who this proportion of voters choose. Typically, we call these proportions percentages and we would say something like "Approximately 68 percent of the population voted in this election and 48 percent of these voters voted for Barack Obama."

In estimating a parameter, we can use a point estimate or an interval estimate. The point estimate for the population proportion, p, is $\hat{p}$. We can also find interval estimates for this parameter. These intervals are based on the sampling distributions of $\hat{p}$.

If we are interested in finding an interval estimate for the population proportion, the following two conditions must be satisfied:

1. We must have a random sample.
2. The sample size must be large enough ($n\hat{p} > 10$ and $n(1-\hat{p}) > 10$) that we can use the normal distribution as an approximation to the binomial distribution.

$\sqrt{\frac{p(1-p)}{n}}$ is the standard deviation of the distribution of sample proportions. The distribution of sample proportions is as follows:



Since we do not know the value of p, we must replace it with $\hat{p}$. We then have the standard error of the sample proportions, $\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$. If we are interested in a 95% confidence interval, using the Empirical Rule, we are saying that we want the difference between the sample proportion and the population proportion to be within 1.96 standard deviations.

That is, we want the following:

$$\begin{aligned}
&-1.96 \text{ standard errors} < \hat{p} - p < 1.96 \text{ standard errors}\\
&-\hat{p} - 1.96\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} < -p < -\hat{p} + 1.96\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\\
&\hat{p} + 1.96\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} > p > \hat{p} - 1.96\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\\
&\hat{p} - 1.96\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} < p < \hat{p} + 1.96\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}
\end{aligned}$$

This is a 95% confidence interval for the population proportion. If we generalize for any confidence level, the confidence interval is as follows:

$$\hat{p} - z_{\frac{\alpha}{2}}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} < p < \hat{p} + z_{\frac{\alpha}{2}}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

In other words, the confidence interval is $\hat{p} \pm z_{\frac{\alpha}{2}}\left(\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\right)$. Remember that $z_{\frac{\alpha}{2}}$ refers to the positive z-score for a particular confidence interval. Also, $\hat{p}$ is the sample proportion, and n is the sample size. As before, the margin of error is $z_{\frac{\alpha}{2}}\left(\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\right)$, and the confidence interval is $\hat{p}\pm$ the margin of error.

Example: A congressman is trying to decide whether to vote for a bill that would remove all speed limits on interstate highways. He will decide to vote for the bill only if 70 percent of his constituents favor the bill. In a survey of 300 randomly selected voters, 224 (74.6%) indicated they would favor the bill. The congressman decides that he wants an estimate of the proportion of voters in the population who are likely to favor the bill. Construct a confidence interval for this population proportion.

Our sample proportion is 0.746, and our standard error of the proportion is 0.0251. We will construct a 95% confidence interval for the population proportion. Under the normal curve, 95% of the area is between $z = -1.96$ and $z = 1.96$. Thus, the confidence interval for this proportion would be:

$$\begin{aligned}
&0.746 \pm (1.96)(0.0251)\\
&0.697 < p < 0.795
\end{aligned}$$

With respect to the population proportion, we are 95% confident that the interval from 0.697 to 0.795 contains the population proportion. The population proportion is either in this interval, or it is not. When we say that this is a 95% confidence interval, we mean that if we took 100 samples, all of size n, and constructed 95% confidence intervals for each of these samples, 95 out of the 100 confidence intervals we constructed would capture the population proportion, p.

Example: A large grocery store has been recording data regarding the number of shoppers that use savings coupons at its outlet. Last year, it was reported that 77% of all shoppers used coupons, and 19 times out of 20, these results were considered to be accurate within 2.9%.

a) Are you dealing with a 90%, 95%, or 99% confidence level?

b) What is the margin of error?

c) Calculate the confidence interval.

d) Explain the meaning of the confidence interval.

a) The statement 19 times out of 20 indicates that you are dealing with a 95% confidence interval.

b) The results were accurate within 2.9%, so the margin of error is 0.029.

c) The confidence interval is simply $\hat{p}\pm$ the margin of error.

$$77\% - 2.9\% = 74.1\% \qquad 77\% + 2.9\% = 79.9\%$$

Thus, the confidence interval is from 0.741 to 0.799.

d) The 95% confidence interval from 0.741 to 0.799 for the population proportion is an interval calculated from a sample by a method that has a 0.95 probability of capturing the population proportion.

On the Web

http://tinyurl.com/27syj3x This simulates confidence intervals for the population proportion.

http://tinyurl.com/28z97lr Explore how changing the confidence level and/or the sample size affects the length of the confidence interval.

Lesson Summary

In this lesson, you learned that a sample mean is known as a point estimate, because this single number is used as a plausible value of the population mean. In addition to reporting a point estimate, you discovered how to calculate an interval of reasonable values based on the sample data. This interval estimator of the population mean is called the confidence interval. You can calculate this interval for the population mean by using the formula $\bar{x} \pm z_{\frac{\alpha}{2}} \left(\frac{\sigma}{\sqrt{n}}\right)$. The value of $z_{\frac{\alpha}{2}}$ is different for each confidence interval of 90%, 95%, and 99%. You also learned that the probability is attributed to the method used to calculate the confidence interval.

In addition, you learned that you calculate the confidence interval for a population proportion by using the formula $\hat{p} \pm z_{\frac{\alpha}{2}} \left(\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\right)$.

Points to Consider

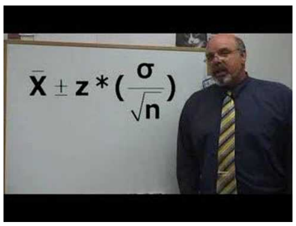
- Does replacing σ with s change your chance of capturing the unknown population mean?
- Is there a way to increase the chance of capturing the unknown population mean?

Multimedia Links

For an explanation of the concept of confidence intervals **(17.0)**, see kbower50, What are Confidence Intervals? (3:24).

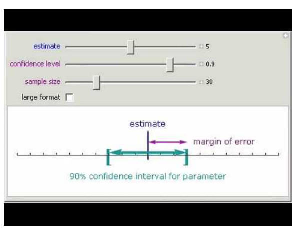
MEDIA
Click image to the left for more content.

For a description of the formula used to find confidence intervals for the mean **(17.0)**, see mathguyzero,Statistics Confidence Interval Definition and Formula (1:26).



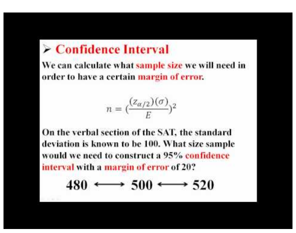
MEDIA
Click image to the left for more content.

For an interactive demonstration of the relationship between margin of error, sample size, and confidence intervals **(17.0)**, see wolframmathematica, Confidence Intervals: ConfidenceLevel, Sample Size,andMargin of Error (0:16).



MEDIA
Click image to the left for more content.

For an explanation on finding the sample size for a particular margin of error **(17.0)**, see statslectures, Calculating R equired Sample Size to Estimate Population Mean (2:18).



MEDIA
Click image to the left for more content.

Review Questions

1. In a local teaching district, a technology grant is available to teachers in order to install a cluster of four computers in their classrooms. From the 6,250 teachers in the district, 250 were randomly selected and asked if they felt that computers were an essential teaching tool for their classroom. Of those selected, 142 teachers felt that computers were an essential teaching tool.
   a. Calculate a 99% confidence interval for the proportion of teachers who felt that computers are an essential teaching tool.

b. How could the survey be changed to narrow the confidence interval but to maintain the 99% confidence interval?

2. Josie followed the guidelines presented to her and conducted a binomial experiment. She did 300 trials and reported a sample proportion of 0.61.

   a. Calculate the 90%, 95%, and 99% confidence intervals for this sample.
   b. What did you notice about the confidence intervals as the confidence level increased? Offer an explanation for your findings?
   c. If the population proportion were 0.58, would all three confidence intervals enclose it? Explain.

Keywords

Central Limit Theorem

Confidence interval

Confidence level

Margin of error

Parameter

Point estimate

Sample means

Sample proportion

Sampling distributions

Standard error