

CHAPTER

11**Sampling Distributions and Estimation****Chapter Outline**

- 11.1 SURVEYS AND SAMPLING**
 - 11.2 SAMPLING DISTRIBUTION**
 - 11.3 CENTRAL LIMIT THEOREM**
 - 11.4 CONFIDENCE INTERVALS**
-

11.1 Surveys and Sampling

Learning Objectives

- Differentiate between a census and a survey or sample.
- Distinguish between sampling error and bias.
- Identify and name potential sources of bias from both real and hypothetical sampling situations.

Census vs. Sample

A **sample** is a representative subset of a population. If a statistician or other researcher wants to know some information about a population, the only way to be truly sure is to conduct a census. In a **census**, every unit in the population being studied is measured or surveyed. If we really wanted to know the true approval rating of the president, for example, we would have to ask every single American adult his or her opinion. There are some obvious reasons why a census is impractical in this case, and in most situations.

Why is this impractical? First, it would be extremely expensive for the polling organization. They would need an extremely large workforce to try and collect the opinions of every American adult. Also, it would take many workers and many hours to organize, interpret, and display this information. Even if it could be done in several months, by the time the results were published, it would be very probable that recent events had changed peoples' opinions and that the results would be obsolete.

In addition, a census has the potential to be destructive to the population being studied. For example, many manufacturing companies test their products for quality control. A padlock manufacturer might use a machine to see how much force it can apply to the lock before it breaks. If they did this with every lock, they would have none left to sell! Likewise, it would not be a good idea for a biologist to find the number of fish in a lake by draining the lake and counting them all!

Sampling and Its Risks

Due to all of the difficulties associated with a census, sampling is much more practical. However, it is important to understand that even the most carefully planned sample will be subject to random variation between the sample and the population. Recall that these **differences due to chance** are called **sampling error**. Opinion polls, like the *New York Times* poll mentioned in the introduction, tend to refer to this as **margin of error**. You will learn how to calculate sampling error, or the margin of error, associated with samples in the next section.

The second statement quoted from the *New York Times* article mentions another problem with sampling. That is, it is often difficult to obtain a sample that accurately reflects the total population. It is also possible to make mistakes in selecting the sample and collecting the information. These problems result in a **non-representative sample**, or one in which our conclusions differ from what they would have been if we had been able to conduct a census.

Bias in Samples and Surveys

The term most frequently applied to a non-representative sample is **bias**. Bias has many potential sources. It is important when selecting a sample or designing a survey that a statistician make every effort to eliminate potential sources of bias. In this section, we will discuss some of the most common types of bias. While these concepts are universal, the terms used to define them here may be different than those used in other sources.

Sampling Bias

In general, sampling bias refers to the methods used in selecting the sample. The **sampling frame** is the term we use to refer to the group or listing from which the sample is to be chosen. If you wanted to study the population of students in your school, you could obtain a list of all the students from the office and choose students from the list. This list would be the sampling frame.

Incorrect Sampling Frame

If the list from which you choose your sample does not accurately reflect the characteristics of the population, this is called **incorrect sampling frame**. A sampling frame error occurs when some group from the population does not have the opportunity to be represented in the sample. For example, surveys are often done over the telephone. You could use the telephone book as a sampling frame by choosing numbers from the telephone book. However, some phone numbers are not listed in the telephone book. In addition, younger adults in particular tend to only use their cell phones or computer-based phone services and may not even have traditional phone service. Even if you picked phone numbers randomly, the sampling frame could be incorrect, because there are also people, especially those who may be economically disadvantaged, who have no phone. There is absolutely no chance for these individuals to be represented in your sample. A term often used to describe the problems when a group of the population is not represented in a survey is **undercoverage**. Undercoverage can result from all of the different sampling biases.

You may have heard of one of the most famous examples of sampling frame error. It occurred during the 1936 U.S. presidential election. The *Literary Digest*, a popular magazine at the time, conducted a poll and predicted that Alf Landon would win the election. As it turned out, the election was won in a landslide by Franklin Delano Roosevelt. The magazine obtained a huge sample of ten million people, and from that pool, 2 million replied. With these numbers, you would typically expect very accurate results. However, the magazine used their subscription list as their sampling frame. During the depression, these individuals would have been only the wealthiest Americans, who tended to vote Republican, and left the majority of typical voters under-covered.

Convenience Sampling

Suppose your statistics teacher gave you an assignment to perform a survey of 20 individuals. You would most likely tend to ask your friends and family to participate, because it would be easy and quick. This is an example of **convenience sampling**, or **convenience bias**. While it is not always true, your friends are usually people who share common values, interests, and opinions. This could cause those opinions to be over-represented in relation to the true population. Also, have you ever been approached by someone conducting a survey on the street or in a mall? If such a person were just to ask the first 20 people they found, there is the potential that large groups representing various opinions would not be included, resulting in undercoverage.

Judgment Sampling

Judgment sampling occurs when an individual or organization that is usually considered an expert in the field being studied chooses the individuals or group of individuals to be used in the sample. Because it is based on a subjective choice, even by someone considered an expert, it is very susceptible to bias. In some sense, this is what those responsible for the *Literary Digest* poll did. They incorrectly chose groups they believed would represent the population. If a person wants to do a survey on middle-class Americans, how would this person decide who to include? It would be left to this person's own judgment to create the criteria for those considered middle-class. This individual's judgment might result in a different view of the middle class that might include wealthier individuals that others would not consider part of the population. Similar to judgment sampling, in **quota sampling**, an individual or organization attempts to include the proper proportions of individuals of different subgroups in their sample. While it might sound like a good idea, it is subject to an individual's prejudice and is, therefore, prone to bias.

Size Bias

If one particular subgroup in a population is likely to be over-represented or under-represented due to its size, this is sometimes called **size bias**. If we chose a state at random from a map by closing our eyes and pointing to a particular place, larger states would have a greater chance of being chosen than smaller ones. As another example, suppose that we wanted to do a survey to find out the typical size of a student's math class at a school. The chances are greater that we would choose someone from a larger class for our survey. To understand this, say that you went to a very small school where there are only four math classes, with one class having 35 students, and the other three classes having only 8 students. If you simply choose students at random, it is more likely you will select students for your sample who will say the typical size of a math class is 35, since there are more students in the larger class.

Response Bias

The term **response bias** refers to problems that result from the ways in which the survey or poll is actually presented to the individuals in the sample.

Voluntary Response Bias

Television and radio stations often ask viewers/listeners to call in with opinions about a particular issue they are covering. The websites for these and other organizations also usually include some sort of online poll question of the day. Reality television shows and fan balloting in professional sports to choose all-star players make use of these types of polls as well. All of these polls usually come with a disclaimer stating that, "This is not a scientific poll." While perhaps entertaining, these types of polls are very susceptible to **voluntary response, or self-selection, bias**. The people who respond to these types of surveys tend to feel very strongly one way or another about the issue in question, and the results might not reflect the overall population. Those who still have an opinion, but may not feel quite so passionately about the issue, may not be motivated to respond to the poll.

Non-Response Bias

One of the biggest problems in polling is that most people just don't want to be bothered taking the time to respond to a poll of any kind. They hang up on a telephone survey, put a mail-in survey in the recycling bin, or walk quickly past an interviewer on the street. We just don't know how much these individuals' beliefs and opinions reflect those of the general population, and, therefore, almost all surveys could be prone to **non-response bias**.

Questionnaire Bias

Questionnaire bias occurs when the way in which the question is asked influences the response given by the individual. It is possible to ask the same question in two different ways that would lead individuals with the same basic opinions to respond differently. Consider the following two questions about gun control.

"Do you believe that it is reasonable for the government to impose some limits on purchases of certain types of weapons in an effort to reduce gun violence in urban areas?"

"Do you believe that it is reasonable for the government to infringe on an individual's constitutional right to bear arms?"

A gun rights activist might feel very strongly that the government should never be in the position of limiting guns in any way and would answer no to both questions. Someone who is very strongly against gun ownership, on the other hand, would probably answer yes to both questions. However, individuals with a more tempered, middle position on the issue might believe in an individual's right to own a gun under some circumstances, while still feeling that there is a need for regulation. These individuals would most likely answer these two questions differently.

You can see how easy it would be to manipulate the wording of a question to obtain a certain response to a poll question. Questionnaire bias is not necessarily always a deliberate action. If a question is poorly worded, confusing, or just plain hard to understand, it could lead to non-representative results. When you ask people to choose between two options, it is even possible that the order in which you list the choices may influence their response!

Incorrect Response Bias

A major problem with surveys is that you can never be sure that the person is actually responding truthfully. When an individual intentionally responds to a survey with an untruthful answer, this is called **incorrect response bias**. This can occur when asking questions about extremely sensitive or personal issues. For example, a survey conducted about illegal drinking among teens might be prone to this type of bias. Even if guaranteed their responses are confidential, some teenagers may not want to admit to engaging in such behavior at all. Others may want to appear more rebellious than they really are, but in either case, we cannot be sure of the truthfulness of the responses.

Identifying Sources of Bias

Example A

You are assisting with a study attempting to determine the satisfaction of school communication with students who speak a second language at home. The plan is to send home a questionnaire to the parents of the students, asking them about their opinion.

What kind(s) of bias is this survey method particularly prone to? How might they be addressed?

Solution

This method of sampling is liable to result in both non-response and undercoverage bias. Non-response bias is an issue any time a sample population is expected to submit a questionnaire, as your results are going to include more input from the type of person who is willing and able to complete and submit your survey. In this case, undercoverage is a particular problem, since the population most affected by the study is also unusually liable to misinterpret the questions or the reason for them due to the language barrier.

One possible solution might be to conduct a phone survey conducted by a native speaker in the target language(s).

Example B

What type(s) of bias do the experiments below suggest?

- An experiment to determine the danger of mixing household chemicals is conducted by collecting samples of chemicals found under the experimenter's sink.
- Mall shoppers are asked to fill out and return a form rating their shopping experiences at each of the 26 stores to identify the most popular stores in each of 4 categories.
- A study of the average grades of mathematics students polls 16 Algebra I students, 14 Geometry students, 7 Calculus students, and 19 Statistics students.

Solution

- Undercoverage bias –This experiment is a prime example of the problems associated with **convenience sampling**, since the only chemicals used were the ones conveniently found in one location, the results could not be assumed to be the same as with chemicals found under other sinks.

- b. Non-response bias –Since the results are dependent on the shoppers turning in a response form on their own, the results will be biased toward a specific type of personality, and will not reflect a true cross-section of shoppers' experiences.
- c. Undercoverage –The study only includes approximately $\frac{1}{2}$ as many Calculus students as the other subjects.

Reducing Bias

Randomization

The best technique for reducing bias in sampling is **randomization**. When a **simple random sample** of size n (commonly referred to as an SRS) is taken from a population, all possible samples of size n in the population have an equal probability of being selected for the sample. For example, if your statistics teacher wants to choose a student at random for a special prize, he or she could simply place the names of all the students in the class in a hat, mix them up, and choose one. More scientifically, your teacher could assign each student in the class a number from 1 to 25 (assuming there are 25 students in the class) and then use a computer or calculator to generate a random number to choose one student. This would be a simple random sample of size 1.

A Note about Randomness

Technology Note: Generating Random Numbers

It is important that you understand that there is no such thing as true randomness, especially on a calculator or computer. When you choose the 'rand' function, the calculator has been programmed to return a ten digit decimal that, using a very complicated mathematical formula, simulates randomness. Each digit, in theory, is equally likely to occur in any of the individual decimal places. What this means in practice is that if you had the patience (and the time!) to generate a million of these on your calculator and keep track of the frequencies in a table, you would find there would be an approximately equal number of each digit. However, two brand-new calculators will give the exact same sequences of random numbers! This is because the function that simulates randomness has to start at some number, called a **seed value**. All the calculators are programmed from the factory (or when the memory is reset) to use a seed value of zero. If you want to be sure that your sequence of random digits is different from everyone else's, you need to seed your random number function using a number different from theirs.

Systematic Sampling

There are other types of samples that are not simple random samples, and one of these is a **systematic sample**. In **systematic sampling**, after **choosing a starting point at random**, subjects are selected using a jump number. If you have ever chosen teams or groups in gym class by **counting off by threes or fours**, you were engaged in systematic sampling. The jump number is determined by dividing the population size by the desired sample size to insure that the sample combs through the entire population. **If we had a list of everyone in your class of 25 students in alphabetical order, and we wanted to choose 5 of them, we would choose every 5th student.** Let's try choosing a starting point at random by generating a random number from 1 to 25. Assume we get the number 14 as our seed value.

In this case, we would start with student number 14 and then select every 5th student until we had 5 in all. When we came to the end of the list, we would continue the count at number 1. Thus, our chosen students would be: 14, 19, 24, 4, and 9. It is important to note that this is not a simple random sample, as not every possible sample of 5 students has an equal chance of being chosen. For example, it is impossible to have a sample consisting of students 5, 6, 7, 8, and 9.

Cluster Sampling

Cluster sampling is when a naturally occurring group is selected at random, and then either all of that group, or randomly selected individuals from that group, are used for the sample. If we select at random from out of that group, or cluster into smaller subgroups, this is referred to as **multi-stage sampling**. For example, to survey student opinions or study their performance, we could choose 5 schools at random from your state and then use an SRS (simple random sample) from each school. If we wanted a national survey of urban schools, we might first choose 5 major urban areas from around the country at random, and then select 5 schools at random from each of these cities. This would be both cluster and multi-stage sampling. Cluster sampling is often done by selecting a particular block or street at random from within a town or city. It is also used at large public gatherings or rallies. If officials take a picture of a small, representative area of the crowd and count the individuals in just that area, they can use that count to estimate the total crowd in attendance.

Stratified Sampling

In **stratified sampling**, the population is divided into groups, called **strata** (the singular term is 'stratum'), that have some meaningful relationship. Very often, groups in a population that are similar may respond differently to a survey. In order to help reflect the population, we stratify to insure that each opinion is represented in the sample. For example, we often stratify by gender or race in order to make sure that the often divergent views of these different groups are represented. In a survey of high school students, we might choose to stratify by school to be sure that the opinions of different communities are included. If each school has an approximately equal number of students, then we could simply choose to take an SRS of size 25 from each school. If the numbers in each stratum are different, then it would be more appropriate to choose a fixed sample (100 students, for example) from each school and take a number from each school proportionate to the total school size.

Lesson Summary

If you collect information from every unit in a population, it is called a census. Because a census is so difficult to do, we instead take a representative subset of the population, called a sample, to try and make conclusions about the entire population. The downside to sampling is that we can never be completely sure that we have captured the truth about the entire population, due to random variation in our sample that is called sampling error. The list of the population from which the sample is chosen is called the sampling frame. Poor technique in surveying or choosing a sample can also lead to incorrect conclusions about the population that are generally referred to as bias. Selection bias refers to choosing a sample that results in a subgroup that is not representative of the population. Incorrect sampling frame occurs when the group from which you choose your sample does not include everyone in the population, or at least units that reflect the full diversity of the population. Incorrect sampling frame errors result in undercoverage. This is where a segment of the population containing an important characteristic did not have an opportunity to be chosen for the sample and will be marginalized, or even left out altogether.






Points to Consider

- How is the **margin of error** for a survey calculated?
- What are the effects of sample size on sampling error?

Review Questions

1. Brandy wanted to know which brand of soccer shoe high school soccer players prefer. She decided to ask the girls on her team which brand they liked.
 - a. What is the population in this example?



- b. What are the units? 
 - c. If she asked all high school soccer players this question, what is the statistical term we would use to describe the situation? 
 - d. Which group(s) from the population is/are going to be under-represented?
 - e. What type of bias best describes the error in her sample? Why?
 - f. Brandy got a list of all the soccer players in the Colonial conference from her athletic director, Mr. Sprain. This list is called the what? 
 - g. If she grouped the list by boys and girls, and chose 40 boys at random and 40 girls at random, what type of sampling best describes her method? 
2. Your doorbell rings, and you open the door to find a 6-foot-tall boa constrictor wearing a trench coat and holding a pen and a clip board. He says to you, "I am conducting a survey for a local clothing store. Do you own any boots, purses, or other items made from snake skin?" After recovering from the initial shock of a talking snake being at the door, you quickly and nervously answer, "Of course not," as the wallet you bought on vacation last summer at Reptile World weighs heavily in your pocket. What type of bias best describes this ridiculous situation? Explain why. 

In each of the next two examples, identify the type of sampling that is most evident and explain why you think it applies.

3. In order to estimate the population of moose in a wilderness area, a biologist familiar with that area selects a particular marsh area and spends the month of September, during mating season, cataloging sightings of moose. What two types of sampling are evident in this example?
4. The local sporting goods store has a promotion where every 1000th customer gets a \$10 gift card.

For questions 5-9, an amusement park wants to know if its new ride, The Pukeinator, is too scary. Explain the type(s) of bias most evident in each sampling technique and/or what sampling method is most evident. Be sure to justify your choice.

5. The first 30 riders on a particular day are asked their opinions of the ride.
6. The name of a color is selected at random, and only riders wearing that particular color are asked their opinion of the ride.
7. A flier is passed out inviting interested riders to complete a survey about the ride at 5 pm that evening.
8. Every 12th teenager exiting the ride is asked in front of his friends: "You didn't think that ride was scary, did you?"
9. Five riders are selected at random during each hour of the day, from 9 AM until closing at 5 PM.
10. There are 35 students taking statistics in your school, and you want to choose 10 of them for a survey about their impressions of the course. Use your calculator to select a SRS of 10 students. (Seed your random number generator with the number 10 before starting.) Assuming the students are assigned numbers from 1 to 35, which students are chosen for the sample?

11.2 Sampling Distribution

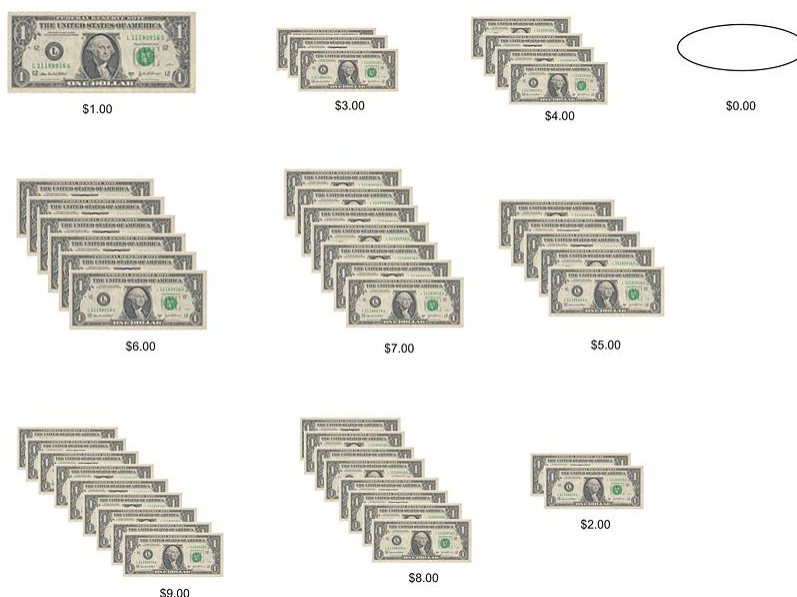
Learning Objectives

- 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 we can learn what is true in a population when it would be impossible to contact everyone? Statistics allows us to make use of the tool of probability to estimate what is true from just a sample of the subjects we are interested in.

Suppose, for example, that we want to know how much cash people carry around in their pockets, on average. To make this simple, we are going to work with a very small population of people: 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.

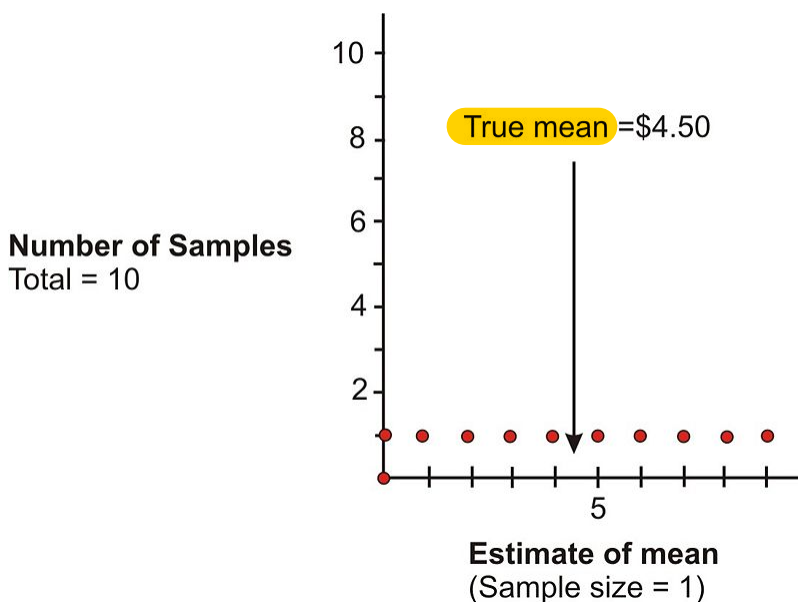


Our Scenario

In this scenario, we have a population of size ten. One person has no money, another has \$1.00, another has \$2.00, and so on, until we reach the person who has \$9.00. Our goal is to determine the average amount of money per person in this population. What is that true mean? 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. Of course, for the purpose of this exercise, we don't know this!

Suppose you couldn't count the money of all ten people at once. Let's say instead you had 10 different individuals all taking samples of the population. To start, suppose each of the ten researchers were to randomly select a sample of only **one** person from the ten. That makes 10 samples of 1 person each. In this example, we would say that $n = 1$

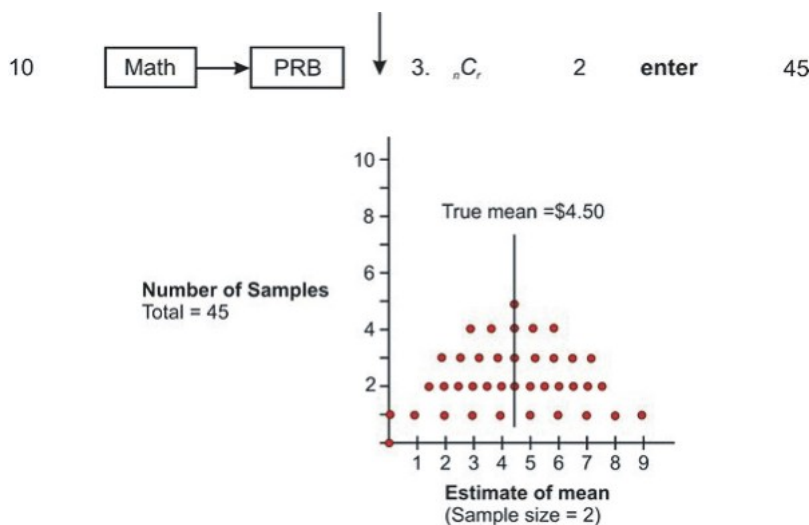
(or sample size is one). The graph below shows the mean of each possible sample of $n=1$. (Since there is only one person in each sample, the mean is the number of dollars in their pocket). Each of the 10 individuals was selected and constitutes their own sample of size $n=1$.



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 range of sample means is anywhere from \$0.00 to \$9.00. The true mean of \$4.50 could be missed by quite a bit with any one given sample.

Sample Size of $n=2$

What happens if we take samples of two people at a time? 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. We select all possible samples of size two from the population and graph the means of those samples. The **sampling distribution** of the sample means is as follows:

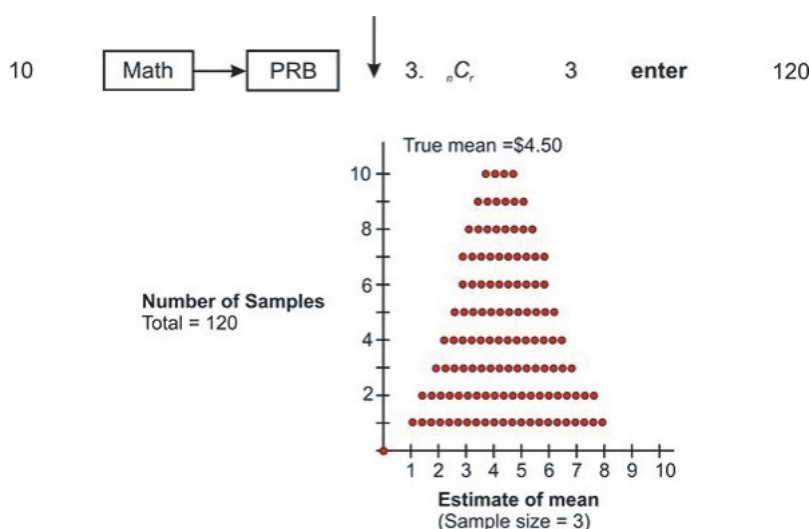


Increasing the sample size has improved your estimates. There are now 45 possible samples, such as (\$0, \$1), (\$0, \$2), (\$7, \$8), (\$8, \$9), and so on, and some of these samples produce the same means. For example, (\$0, \$6), (\$1,

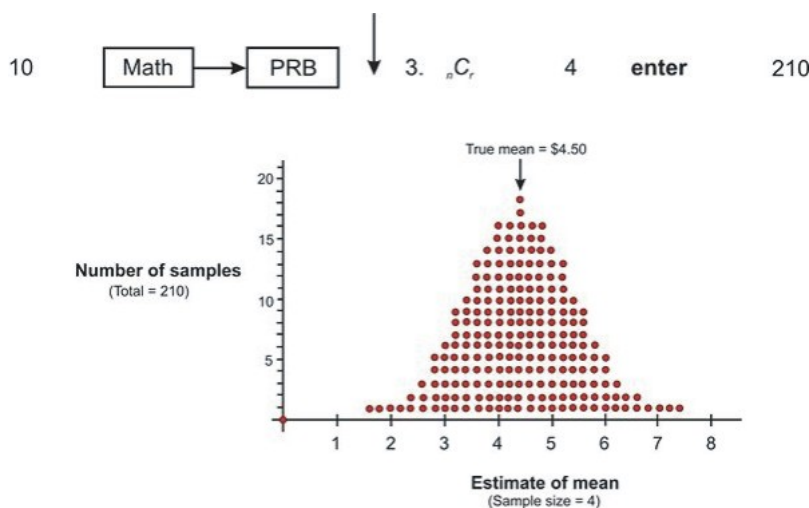
\$5), and (\$2, \$4) all produce means of \$3. The three dots above the mean of 3 represent these three samples. In addition, the 45 means are not evenly distributed, as they were when the sample size was one. Instead, they are more clustered around the true mean of \$4.50. (\$0, \$1) and (\$8, \$9) 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.

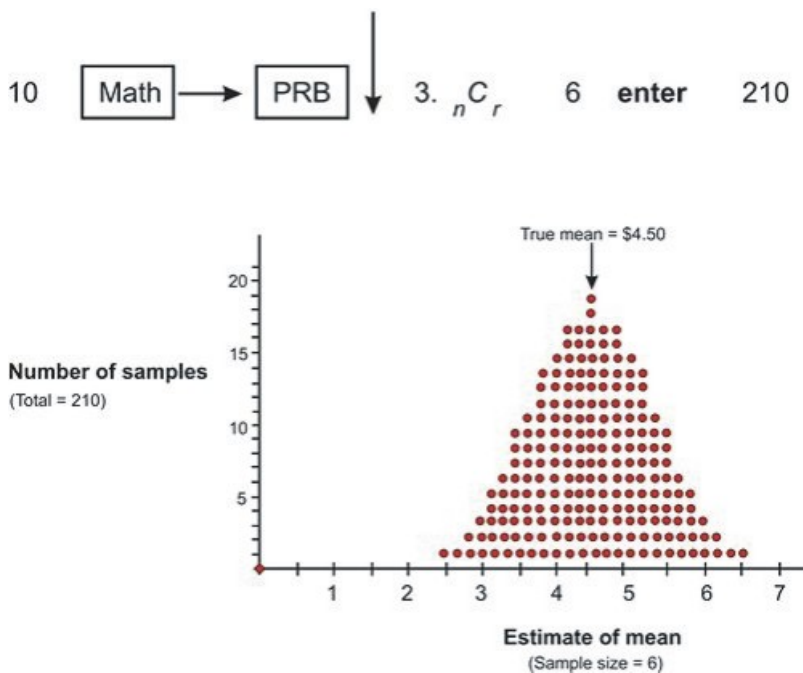
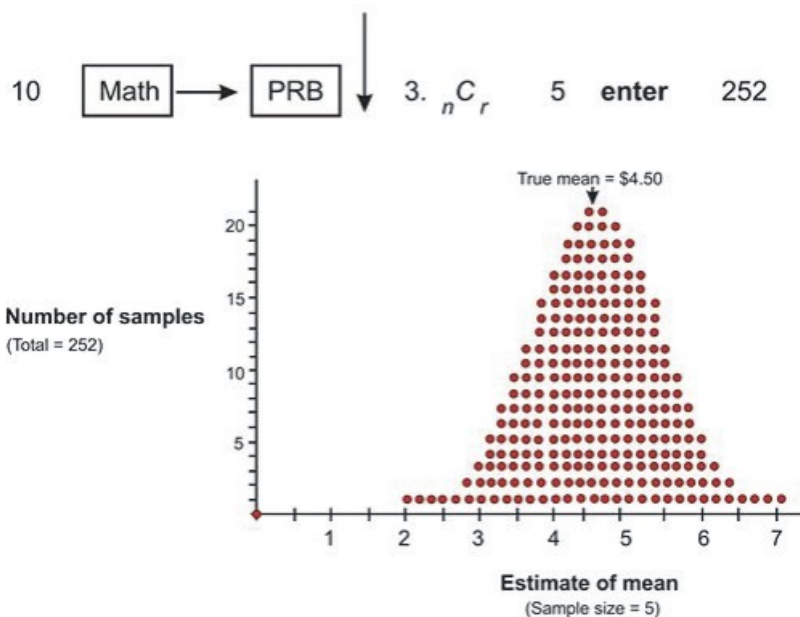
Sample Size of $n=3$

If three people are randomly selected 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:



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





Impact of Sample Size

From the graphs above, it is obvious that increasing the sample size resulted in a sampling distribution of means that were more closely clustered around the true mean. Also, the sampling distribution of the sample means is approximately normal, as can be seen by the bell shape in each of the graphs. In case you are wondering, if a sample of size $n=10$ were selected, there would be only one possible sample, and it would yield the true mean of \$4.50.

Important Lessons

There are two important pieces to take away from this lesson. First, notice that the sample means become more and more normally distributed around the true mean (the population parameter) as we increase our sample size. Second, notice that the variability of the sample means decreases as sample size increases. The sample means are more tightly

clustered around the true mean. This variability of sample means is called the **standard error**, s . You can think of it as the standard deviation of a sampling distribution.

And here is one last piece of information to take away. The sampling distribution, as it becomes more normal in shape, also adheres to the Empirical Rule. This means that certain proportions of the sample means will fall within defined increments. In this case, each increment would be one standard error from the population parameter. According to this rule, 34% of the sample means will fall within one standard error above the population parameter, and another 34% will fall within one standard error below the population parameter. 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.

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?

11.3 Central Limit Theorem

Learning Objectives

- Understand one of the more remarkable theorems in all of mathematics, the **Central Limit Theorem**.
- Recognize the **relationship between the Normal distribution and the Central Limit Theorem**.

Introduction

In the previous lesson, you learned that sampling is an important tool for determining the characteristic of a population. When we constructed a distribution of sample means, we saw that the sample means clustered around the true mean. As the sample size increased, the shape of that distribution became more and more Normal. Although the true mean of the population was unknown, random sampling yielded a reliable estimate. It is now time to learn how one of the most remarkable theorems in statistics will allow us to estimate what is true in a population *without* having to repeatedly sample!

Central Limit Theorem

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

Why is that idea so important? The reason is simple. Here is what this theorem allows us to do: **If we can select a single sample of a known size from our population and calculate its mean, we can use the Central Limit Theorem to predict what that true population mean must be within a defined degree of confidence.** And, more importantly, this holds true no matter what the shape of the original distribution. That's pretty amazing.

Before going any further, you should become familiar with (or reacquaint yourself with) the symbols that are commonly used when dealing with population values, sample statistics, and statistics of a sampling distribution of means. These symbols are shown in the table below. Note that the notation \bar{x} (x -bar) is used to represent each value in a sampling distribution (rather than the random variable x) to indicate that each value is a sample mean.

TABLE 11.1:

	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	

Formally, the CLT says:

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.

In “normal English”:

If you collect many samples from an ordinary random variable, and calculate the mean of each sample, then the means will be distributed in an approximate bell-curve, and the “mean of means” will be the same as the mean of the population. The larger the size of the samples you collect, the more closely the distribution of their means will

approximate a normal distribution.

Using the Central Limit Theorem to Construct the Sampling Distribution

So how can we use the Central Limit Theorem to help us construct a sampling distribution without repeatedly sampling? We use what we know about the **population** and our proposed **sample size** to sketch the theoretical sampling distribution. Remember, to sketch a distribution we need to know its shape, center and spread.

Notes to remember:

- Shape of the sampling distribution: As long as your sample size is **30 or greater**, you may assume the distribution of the sample means to be approximately normal. This is true regardless of the original distribution of the random variable.
- The mean of the distribution: The mean of a sampling distribution, as you saw in the last lesson, is the mean of the population. Formally: $\mu_{\bar{x}} = \mu$
- The standard error of the distribution: The standard deviation of the sample means can be estimated by dividing the standard deviation of the population by the square root of the sample size. Formally: $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$

So there you have it. You can use the three bullets points above to construct the sampling distribution. And then, since the distribution will be normal with a sample size of 30 or more, you can use what we've learned about the area under a Normal curve to calculate the probability of observing a particular sample mean!

Example A

The time it takes a student to complete the mid-term for Algebra II is a bi-modal distribution with $\mu = 1 \text{ hr}$ and $\sigma = 1 \text{ hr}$. During the month of June, Professor Spence administers the test 64 times. What is the probability that the average mid-term completion time for students during the month of June exceeds 48 minutes?

Solution

Important facts:

- There are more than 30 samples, so the Central Limit Theorem applies.
- The mean of the sample should approximate the mean of the population, in other words $\mu_{\bar{x}} = \mu$
- The standard deviation of Professor Spence's sample, $\sigma_{\bar{x}}$, can be calculated as $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$, where $n = 64$ (the number of tests/samples)
- 48 minutes is the same as $\frac{48}{60} = 0.8 \text{ hrs}$, so the range we are interested in is $x > 0.8 \text{ hrs}$

First calculate the standard deviation of the sample, using $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$:

$$\begin{aligned}\sigma_{\bar{x}} &= \frac{1}{\sqrt{64}} \\ \sigma_{\bar{x}} &= 0.125\end{aligned}$$

Since the sample is normally distributed, according to the CLT, we can use the standard deviation of the sample to calculate the z -score of the minimum value in the relevant range, 0.80 hrs:

$$Z = \frac{0.80 - 1}{0.125} = -1.60$$

Finally, we use the z -score probability reference above to correlate the z -score of -1.60 to the probability of a value greater than that

$$P(Z \geq -1.6) = .9452 \text{ or } 94.52\%$$

Example B

Evan price-checked 123 online auction sellers to record their average asking price for his favorite game. According to a major nation price-checking site, the national average online auction cost for the game is \$35.00 with a standard deviation of \$3.00. Evan found the prices less than \$34.86 on average. How likely is this result?

Solution

Since there are more than 30 samples ($123 > 30$), we can apply the CLT theorem and treat the sample as a normal distribution.

The standard deviation of the sample is: $\sigma_{\bar{x}} = \frac{3}{\sqrt{123}} = \frac{3}{11.09} = .27$

The z -score for Evan's price point of \$34.86 is:

$$Z = \frac{34.86 - 35}{.27} = \frac{-.14}{.27} = -0.518$$

Consulting the z -score probability table, we learn that the area under the normal curve less than 0.52 is .3015 or 30.15%

The likelihood of 123 samples having a mean of \$34.86 is approximately 30.15%

Example C

Mack asked 42 fellow high-school students how much they spent for lunch, on average. According to his research online, the amount spent for lunch by high school students nation wide has $\mu = \$15$, with $\sigma = \$9$. We would assume that Mark's random sample should fall within this sampling distribution. What is the probability that Mack's random sample will have a value that is within \$0.01 of the national average?

Solution

There are a few important facts to note here:

- Mack's sample is 42 students, since $42 \geq 30$, he can safely assume that the sampling distribution of the sample mean will be approximately normal, according to the Central Limit Theorem.
- The range we are considering is \$14.99 to \$15.01, since that represents \$0.01 above and below the mean.
- The mean of the sample should approximate the mean of the population, in other words $\mu_{\bar{x}} = \mu$
- The standard deviation of Mack's sample, $\sigma_{\bar{x}}$, can be calculated as $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$, where $n = 42$

Let's start by finding the standard deviation of the sample, $\sigma_{\bar{x}}$:

$$\begin{aligned}\sigma_{\bar{x}} &= \frac{9}{\sqrt{42}} \\ &= \frac{9}{6.48} \\ \sigma_{\bar{x}} &= 1.389\end{aligned}$$

Since Mack's sample of 42 samples can be assumed to be normally distributed, and since we now know the standard deviation of the sample, 1.389, we can calculate the z -scores for the score at each end of the range using $Z = \frac{\bar{x} - \mu_{\bar{x}}}{\sigma_{\bar{x}}}$:

$$\begin{aligned}Z_1 &= \frac{15.01 - 15.00}{1.389} = +0.01 \\ Z_2 &= \frac{14.99 - 15.00}{1.389} = -0.01\end{aligned}$$

Finally, we look up Z_1 and Z_2 on the Z -score probability table and we calculate the probability associated with the range from $z = -0.01$ to $z = 0.01$. That value is $50.4\% - 49.6\% = \mathbf{0.80\%}$

The probability that Mack's sample will have a mean within \$0.01 of the population mean of \$15.00 is a little less than 1%.

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 .

Vocabulary

The **Central Limit Theorem** states that if samples are drawn at random from any population with a finite mean and standard deviation, then the sampling distribution of the sample means approximates a normal distribution as the sample size increases beyond 30.

The **sampling distribution of the sample means** is a distribution of the means of multiple samples. It is commonly assumed to be a normal distribution, though technically it is normal only if the sample size is greater than 30.

Point to Consider

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

Guided Practice

1. The time it takes to drive from Cheyenne WY to Denver CO has a μ of 1 *hr* and σ of 15 *mins*. Over the course of a month, a highway patrolman makes the trip 55 times. What is the probability that his average travel time exceeds 60 minutes?
2. Abbi polls 95 high school students for their G.P.A.. According to the school, the average G.P.A. of high school students has a mean of 3.0, and standard deviation of .5. What is the probability that Abbi's random sample will have a mean within 0.01 of the population?

3. A recipe website has calculated that the time it takes to cook Sunday dinner has a μ of 1 *hr* with σ of 25 *mins*. Over the course of a month, 172 users report their time spent cooking Saturday dinner, what is the probability that the average user reports spending less than 45 mins cooking dinner?

Solutions

1. The sample mean, $\mu_{\bar{x}}$ is the same as the population mean: 1 *hr* = 60 *mins*.

The sample standard deviation is $\frac{15 \text{ mins}}{\sqrt{55}} = \frac{15}{7.42} = 2.02 \text{ min}$

The 55 trips made by the patrolman exceed the minimum sample size of 30 required to apply the CLT, so we may assume the sample means to be normally distributed.

The z -score of the patrolman's average time is: $\frac{60-60}{2.02} = \frac{0}{2.02} = 0$

According to the z -score percentage reference, a z -score of 0 corresponds to .50 or 50%

There is a 50% probability that the patrolman's mean travel time is greater than 60 mins.

2. The sample mean of the 95 polled G.P.A. scores is the same as the population mean: **3.0**

The sample standard deviation is $\frac{.5}{\sqrt{95}} = \frac{.5}{9.75} = .05$

The 95 sampled G.P.A.'s exceed the minimum sample size of 30, so we may apply the CLT.

The z -scores of the minimum and maximum values in the range of interest, 2.99 to 3.01 is:

$$Z_1 = \frac{2.99 - 3.00}{.05} = \frac{-.01}{.05} = -0.2$$

$$Z_2 = \frac{3.01 - 3.00}{.05} = \frac{.01}{.05} = +0.2$$

Referring to the z -score reference table, **the z -scores -0.2 and 0.2 cover a range of apx. 15.86%**

3. The sample mean, $\mu_{\bar{x}}$ is the same as the population mean: 1 *hr* = 60 *mins*.

The sample standard deviation is $\frac{25 \text{ mins}}{\sqrt{172}} = \frac{25}{13.11} = 1.91 \text{ min}$

The 172 users reporting cooking times exceed the minimum sample size of 30 required to apply the CLT, so we may assume the sample means to be normally distributed.

The z -score of the average reported cooking time is: $\frac{45-60}{1.91} = \frac{-15}{1.91} = -7.85$

According to the z -score percentage reference, a z -score of -7.85 corresponds to 0%.

There is essentially zero probability that 172 users would average only 45 mins.

Review Questions

- 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.
 - What is the expected shape of the resulting distribution?
 - 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.

11.4 Confidence Intervals

Learning Objectives

- Calculate the mean of a sample as a point estimate of the population mean.
- Construct and interpret a confidence interval for a population mean.
- Understand the logic of confidence intervals, as well as the meaning of confidence level and confidence intervals.

Introduction

This lesson introduces the branch of statistics called **inferential statistics**. Earlier, we used *descriptive statistics* to organize and describe our data, or to explore relationships between quantitative and categorical variables. **The goal of inferential statistics is to use *sample* data to increase our knowledge about the *entire population*.** The remainder of this text deals with the different kinds of inferential statistical methods that you can use to test ideas about what is true in a population. In this section, we will focus on **estimation**. Estimation is the inferential technique used to estimate the true value of a population parameter, typically a mean, from a sample.

Confidence Intervals

A sample mean can be referred to as a **point estimate** of a population mean. We call a sample mean 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 any estimate - the true population mean may be larger or smaller than the sample mean.

But not many of us would feel particularly confident in a point estimate. For example, let's say you wanted to know what the average SAT was for students at a particular college. You asked a few students while visiting, and the sample mean was 1280. Would you feel comfortable saying, "The average SAT at this school is 1280." You probably would realize that there is some sampling error involved. The true average SAT may be somewhat higher or somewhat less.

An alternative to reporting a point estimate is identifying a range of possible values the parameter might take. 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.

The general concept of confidence intervals is pretty intuitive: It is easier to predict that an unknown value will lie *somewhere within a wide range* than to predict it will occur *within a narrow range (a single value!)*.

Calculating Confidence Intervals

A **confidence interval** is always centered around the mean of your sample. To construct the interval, you add a margin of error. **The *margin of error* is found by multiplying the standard error of the mean by the *z*-score of the percent confidence level:**

$$\text{confidence interval} = \bar{x} \pm \text{margin of error}$$

$$\text{margin of error} = Z \times \frac{\sigma}{\sqrt{n}}$$

The end result looks something like this: we are 95% confident that the true average SAT for this college falls between 1210 and 1330. Sometimes, the confidence interval is expressed like this: (1210, 1330).

What do we mean by **confidence level**? Common choices for the confidence level are 90%, 95%, and 99%. The selection of a confidence level determines the probability that the confidence interval produced will *contain* the true parameter value. So a confidence level of 99% is higher than a confidence level of 95%. The interval constructed with 99% confidence will have a higher chance of containing the true mean than an interval constructed with 95% confidence.

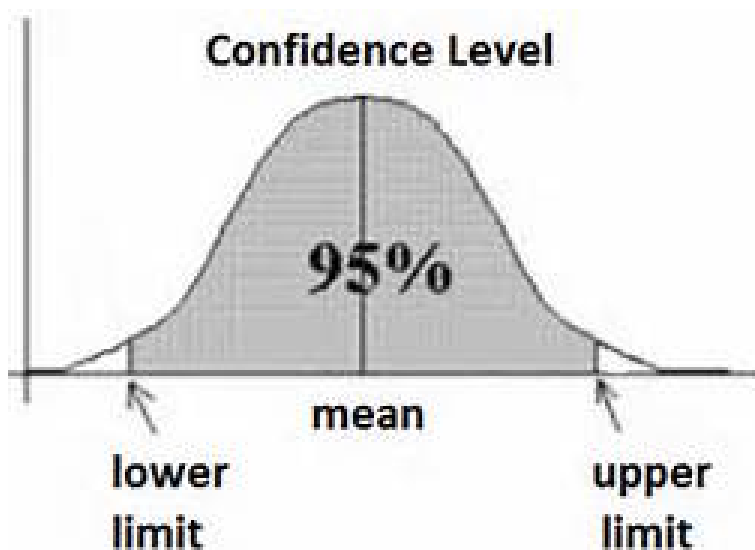


FIGURE 11.1

The confidence level derives from our understanding of the Central Limit Theorem. Think about the sampling distribution for a population mean. If the sample size is at least thirty, the sampling distribution of the mean will be nearly normal. Therefore, any sample mean that we draw has a 95% chance of falling within ± 1.96 standard errors of the true population mean. (Remember your z scores).

So, if we want to construct an interval estimate for our population mean with 95% confidence, we want to add a margin of error that corresponds to ± 1.96 standard errors.

So wouldn't we always want to construct an interval with the highest confidence level possible? Maybe not – and here is why. **The more confidence in the interval, the wider it becomes.** This means that you've lost precision and this could be a problem. Let's read on to see why that is.

Example A

Suppose the average height of a sample of 100 women is 5'5", in other words, $\bar{X} = 5'5''$. Within what range of heights can we expect the population mean to be, with 95% confidence? Assume a standard deviation for the population of 1.5".

Solution

Here is what we know:

- Our sample mean is 5'5"
- The standard deviation of the population is 1'5"
- Our sample size is 100.

We are asked for 95% confidence, so we want to use $z = \pm 1.96$.

$$5.5 \pm 1.96 * 1.5 / \sqrt{100} = 5.5 \pm .294 = (5.21, 5.79)$$

We can conclude, with 95% confidence, that the true average height for women is between 5'2" and 5'8".

If we instead wanted a 99% confidence interval, what would that look like? Now we need $z = \pm 2.58$.

We would conclude, with 99% confidence, that the true average height for women is most likely between 5'1" and 5'9". Notice that we have greater confidence, but we also have a less precise interval. Selecting the level of confidence is a tradeoff between how certain you want to be and how precise an interval you want.

Example B

Suppose you had 40 samples of bags of candy, each of which contains some number of pieces. The number of pieces in each bag is said to be normally distributed. The mean number of candies in your sample is 38 pieces. The standard deviation for bags in the population is 2 pieces. What is the average number of candies in each bag in the population? To answer this question, you need to create a confidence interval. Let's assume that you have been asked to report a confidence interval with 99% certainty.

Solution

Since the population is normally distributed, we can state that the mean of the sample follows the Empirical Rule.

The standard error of the mean is calculated as $\frac{\sigma}{\sqrt{n}}$, so $SEM = \frac{2}{\sqrt{40}} = \frac{2}{6.32} = .316$

Saying that you "expect 99 out of each 100 samples contain the population mean", is the same as saying that the interval has a 99% confidence level.

The interval is called the **confidence interval**, and it is calculated as:

$$38 \pm z_{0.005} \times .316$$

$$38 \pm 2.58 \times .316$$

$$38 \pm 0.81528$$

Therefore, the confidence interval is approximately 37.18 to 38.82. **We are 99% confident that the average number of pieces in each bag of this candy in the population falls between 37 and 39 pieces (with rounding).**

Well, not exactly

You might have noticed in all the examples of confidence intervals above that we used the *population* standard deviation (σ) to calculate the standard error. In reality, this would never happen. In real life, you would only know the *sample* standard deviation (s). And what's the difference, you might ask?. It's very simple: if you only know *sample* statistics, you can't use z in the confidence interval formula. You have to use a different statistic called t . You haven't been introduced to t yet, so we stuck with z . That meant we had to make the crazy assumption that you actually knew the standard deviation of the population! Think about how odd that would be — you are trying to estimate the unknown mean of a population, but somehow you know the standard deviation.

You will be introduced to the t statistic in the next chapter. The logic of the confidence interval will be the same but the formula will change slightly:

$$\text{confidence interval} = \bar{x} \pm \text{margin of error}$$

$$\text{margin of error} = t \times \frac{s}{\sqrt{n}}$$

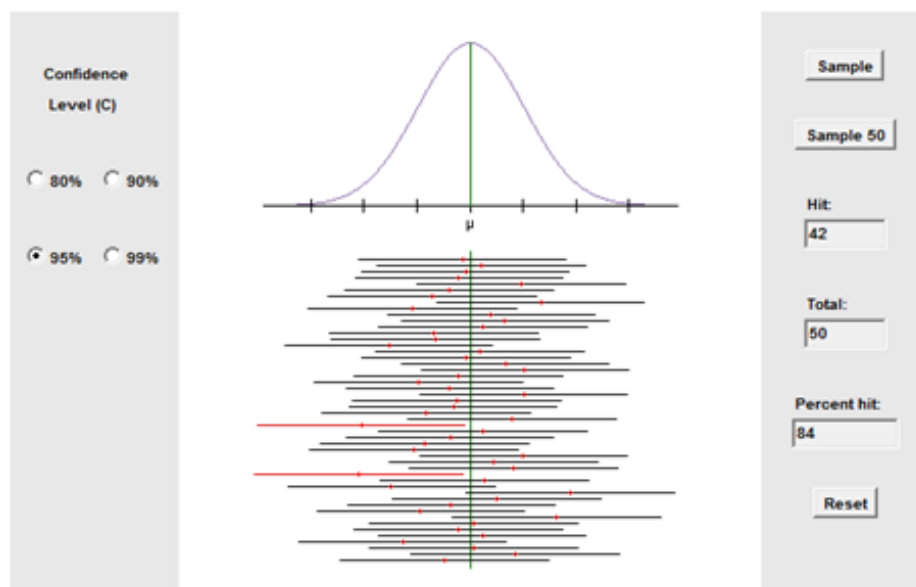
We wanted you to see this formula now, as it is the "real" formula for confidence intervals.

Interpretation of a Confidence Interval

The most common mistake made by persons interpreting a confidence interval is claiming that a confidence level indicates the probability that the mean of the population will occur within your interval! This is not true. Your interval either does - or does not - contain the true population mean.

What a 95% confidence interval means is that if you took 100 samples, all of the same size, and formed 100 confidence intervals, 95 of these intervals would capture the population mean. The probability is attributed to the method, not to any particular confidence interval. This means if you repeated this sampling procedure 100 times, 95 of the intervals produced would contain the population mean. **The confidence level indicates the number of times out of 100 that the mean of the population will be within the given interval of the sample mean.**

Suppose you plot the mean of each of 50 height samples on a graph, and drawing a line each way of the mean of each sample to represent 2 standard deviations. If you were to do this for all 50 of the samples, you might end up with an image like the one below.

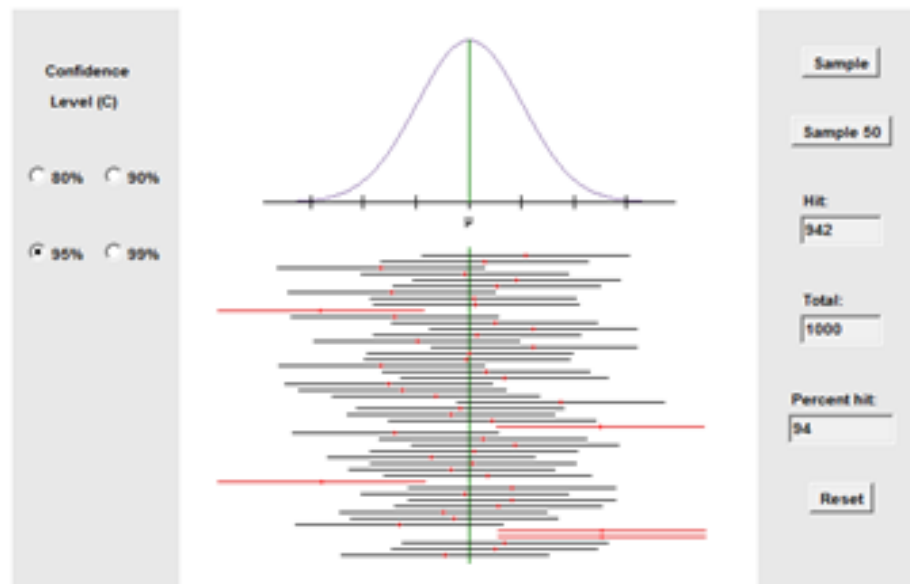


At the top of the image is a normal curve. Each of the lines below the curve has a length that represents a 95% confidence interval, centered on the mean (in red) of the sample.

- What is indicated by the lines that are all red in color?
- What value is indicated by the vertical red center line on each interval?
- What does the "percent hit" number mean? How would it change if you were to continue taking more and more samples of 60 each?

Solutions

- The lines that are colored entirely red have a mean that is greater than 2 standard deviations away from the population mean. In other words, the mean of those two samples was not within the stated *confidence interval* (95%).
- The vertical red center line represents the mean of each sample.
- The “percent hit” number indicates the percentage of times that the population mean was included in the confidence interval of sample means. If you were to continue plotting sample means and confidence intervals, the percent hit would approach 95%. In fact, here is the same graph after 1000 sample runs:



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.

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?

Vocabulary

A **confidence interval** is the interval within which you expect to capture a specific value. The confidence interval width is dependent on the confidence level.

A **confidence level** is the probability value associated with a confidence interval.

More Practice

1. What is a confidence interval?
2. What is the formula for calculating the confidence interval?
3. What is the difference between a confidence interval and a confidence level?
4. What is a margin of error?
5. How is the margin of error calculated?
6. What common misconception about confidence level is corrected by stating that a 99% confidence level means that 99 out of 100 samples are expected to contain the population mean?
7. If a population is known to have an approximately normal distribution, but the standard deviation is unknown, how can the population standard deviation be approximated?
8. If the sample mean is unknown, is it safe to use the population mean as the sample mean?
9. What Z-score corresponds to a 98% confidence interval?
10. What confidence interval is associated with a Z-score of 2.576, assuming a two-tailed test?
11. Which confidence level would describe a wider confidence interval, 80% or 85%?
12. A factory produces bags of marbles for a toy store. The factory has previously calculated that the $\sigma = 1$ marble per bag. If you were to sample 35 bags and calculate $\bar{\mu} = 40$, within what range could you predict μ , with 98% confidence?
13. Interpret your results from question 12, in context.
14. The manager of a clothing store is attempting to estimate the mean number of customers that pass through her store each day. If the data from past estimates and other franchises suggests that $\sigma = 78$, and the manager has collected the customer counts in the table below from a SRS (Simple Random Sample), what can the manager predict the range of customers to be, with 50% confidence?

TABLE 11.2:

148	298	210	213	315	129	145	148	131	281	317
-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----

15. Interpret your answer from problem 14, in context.

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%