

CHAPTER 7

Samples, Sampling Distributions and Confidence Intervals

Summary _____

The fundamental concept discussed in this chapter is the *sampling distribution*, which is a *theoretical* distribution. The standard deviation of any sampling distribution is called the *standard error*; the mean is its *expected value*. A sampling distribution is always of a particular statistic, such as the mean, variance, or correlation coefficient. The sampling distribution of the mean, for example, is a frequency distribution of all possible means (that have the same sample size) from a population. This frequency distribution shows the probability of obtaining any of those sample means with chance draws from the population.

For illustration purposes, you could construct an empirical sampling distribution of the mean. First, choose a population of numbers (you can get a population from the table of random numbers from the back of the book). Second, draw many random samples of the same size. Third, calculate the mean for each sample, and fourth, arrange all the sample means into a frequency polygon. You will find that the polygon looks very much like a normal distribution.

The graph of a sampling distribution of a statistic (either empirical or theoretical) is a picture of the effects that chance has when many random samples of the same size (each with its own statistic) are drawn from a population and for each sample the statistic is calculated. The graph (or table) of these statistics can be used to determine the probability that a particular sample and its statistic came from a particular population.

Sampling distributions come in many shapes (as you will see in later chapters). However, the *sampling distribution of the mean* is a normal distribution if the sample size is large enough.

There are very few phrases or sentences in statistics that are worth memorizing; the Central Limit Theorem (CLT) is probably one of them. The CLT

Chapter 7

says that as sample size approaches ∞ , the form of the sampling distribution of the mean approaches a normal curve that has a mean, μ , and a standard deviation (standard error), $\frac{\sigma}{\sqrt{N}}$. This theorem is true regardless of the form of the population from which the samples are drawn. The CLT only applies to distributions created with sample means.

As can be determined by examining the formula for the standard error of the mean, the sampling distribution of the mean becomes more compact, or narrow, as N is increased.

If the Central Limit Theorem applies, you can use the normal distribution to find the *probability of a sample mean*. When a sample comes from a population with an unknown mean, you can use z scores to determine the probability that the sample mean came from a population with a specific (known) mean. Of course, if the probability is very low, this is reason to believe that the population's unknown mean is different from the specific population mean you used in the z score.

You are justified in using the normal curve to determine probabilities when you know σ and you have an adequate sample. If you cannot meet those requirements (a very common problem in the social sciences), you should use the t distribution. William S. Gosset, known as "Student", developed the t distribution, which is the sampling distribution of \bar{X} when σ must be estimated with s . Each sample size (designated by its particular degrees of freedom) has its own t distribution.

The principal statistic described in this chapter is the *confidence interval*. A confidence interval is an inferential statistical technique that allows you to state, with a specified degree of confidence (such as 95 percent), that an interval of scores (defined by a lower and upper limit) contains the value of an unknown parameter.

For this chapter, the parameter of interest is the mean. Use the t distribution to establish the degree of confidence you select. Thus, the formulas in the chapter give you a *confidence interval about a sample mean*, which tells you about the mean of the population the sample was drawn from.

Chapter 7

Chapter 7 is about how to select samples, and how to use the information you obtain from the sample to understanding something about the population. Using a table of random numbers and the procedures described in the text to select a random sample from the population is the best way to ensure that the sample mirrors the population. In research situations, random samples are usually not feasible, so other methods, such as random assignment and replication are used to establish the validity of the results. Biased samples, in which certain samples of the population are systematically excluded, are to be avoided.

Multiple-Choice Questions

1. According to your text, if you draw a random sample, you are assured that
 - (1) the sample will always mirror the population;
 - (2) you will be somewhat uncertain about the population;
 - (3) the conclusions you draw will be correct;
 - (4) none of the above.

2. Suppose you had a rectangular distribution (like that of the playing cards, pictured in Chapter 6). Suppose you drew many, many random samples of 25 scores and found the mean. If these means were arranged into a frequency distribution, you would expect the distribution to be
 - (1) rectangular;
 - (2) bimodal;
 - (3) either (1) or (2);
 - (4) neither (1) or (2).

3. The Central Limit Theorem (CLT) states that a sampling distribution of the mean approaches the normal curve if
 - (1) the population is normally distributed;
 - (2) the sample size is large;
 - (3) the standard deviation is large;
 - (4) any of the above is sufficient.

4. A standard error is a measure of
 - (1) central tendency;
 - (2) variability;
 - (3) correlation;
 - (4) none of the above.

Chapter 7

5. Confidence intervals and hypothesis testing are parts of
 - (1) descriptive statistics;
 - (2) inferential statistics;
 - (3) both (1) and (2);
 - (4) experimental design.

6. A 95 percent confidence interval means that
 - (1) \bar{X} has a 95 percent probability of being in that interval;
 - (2) the interval has a 95 percent probability of containing μ ;
 - (3) either (1) or (2);
 - (4) neither (1) or (2).

7. A 95 percent confidence interval of 14 to 17 means that
 - (1) 95 percent of the time \bar{X} will be between 14 and 17;
 - (2) 95 of the μ 's will be between 14 and 17;
 - (3) 95 percent of the confidence intervals calculated like this on will contain μ ;
 - (4) all of the above are correct.

8. Under which of the following conditions is the t distribution a normal curve?
 - (1) When $df= 1$;
 - (2) When the population from which the sample is drawn is normal;
 - (3) Both (1) and (2);
 - (4) Neither (1) nor (2).

9. A biased sample is one that
 - (1) is too small;
 - (2) will always lead to a wrong conclusion;
 - (3) has certain groups from the population overrepresented or underrepresented;
 - (4) is always nonrepresentative.

Chapter 7

10. The names of the mean and standard deviation of a sampling distribution are
- (1) mean, standard deviation;
 - (2) mean, standard error;
 - (3) expected value, standard deviation;
 - (4) expected value, standard error.
11. As N becomes larger, $\sigma_{\bar{x}}$
- (1) becomes smaller;
 - (2) becomes larger;
 - (3) gets closer in value to the mean;
 - (4) gets farther in value from the mean.
12. Uncertainty regarding conclusions about a population can be eliminated by
- (1) drawing a sample;
 - (2) drawing a large sample;
 - (3) drawing a large, random sample;
 - (4) none of the above.
13. The word or phrase closest in meaning to the statistical meaning of the word error is
- (1) arithmetic mistake;
 - (2) conceptual mistake;
 - (3) deviation;
 - (4) statistic.
14. Prof. Gus LaPlace, the mad statistician, was fiddling around in his statistical laboratory one stormy night in 1801. He had a large pile of papers in front of him, each with a measurement written on it. "What would I get," he mused, "if I counted the number of papers I have, took the square root, and then divided that into the standard deviation of all the measurements? Hmmmmmmmmmmmmmmmmmm... Well, maybe I'll do it tomorrow," he said. If Prof. LaPlace had carried out his plan, he would have discovered (invented)
- (1) the standard error of the standard deviation;
 - (2) the standard error of the mean;
 - (3) the standard error of the median;
 - (4) none of the above.

Chapter 7

15. When the participants who arrive for a study are assigned to a group on the basis of chance
- (1) random assignment has occurred;
 - (2) random sampling has occurred;
 - (3) the statistical conclusions will be exact;
 - (4) all of the above.
16. The t distribution was invented to handle
- (1) confidence intervals with more than 99 percent confidence;
 - (2) very large samples;
 - (3) not knowing the values of σ ;
 - (4) common arithmetic errors.
17. The difference between a random sample and a biased sample is that
- (1) biased samples are larger;
 - (2) random samples more accurately reflect the population;
 - (3) biased samples more accurately reflect the population;
 - (4) random samples are larger.
18. As N increases, the mean of the sample is
- (1) less representative of the population mean;
 - (2) likely to increase in size;
 - (3) likely to decrease in size;
 - (4) more representative of the population mean.
19. A distribution is most likely to be normal when
- (1) N is large;
 - (2) the population is normal;
 - (3) both (1) and (2);
 - (4) neither (1) nor (2).

Chapter 7

20. A study examining the IQ scores of men at a small private university found that the average IQ for a sample of 5 participants is 140. Which of the following conclusions can be made from this sample?
- (1) Men from this university are smart;
 - (2) Men from this university are below average IQ;
 - (3) Men from this university are smarter than men from other universities;
 - (4) Not enough information to answer this question.

Short-Answer Questions _____

1. In a sentence, describe the Central Limit Theorem (CLT).
2. Distinguish between the concept of a sampling distribution and the sampling distribution of the mean.
3. Suppose you wanted to know whether the weight of vegetarians was less than that of the general population in the United States. Suppose also that you were fortunate enough to have the weights of a representative sample of 49 male vegetarians who were college age. Now, suppose we know that the average weight of 20-to 29-year-old male Americans is 166 pounds. Suppose further that we also know that weight is *not* normally distributed but is positively skewed. Can you use the techniques described in Chapter 7 to determine the probability that the mean weight of college-age male vegetarians came from a population with a mean weight of 166? Write your conclusion and your reasoning.
4. What improvements could be made to the study described in Question 20 of the multiple choice section to allow for greater confidence concerning the results?
5. What does a confidence interval tell us about a sample mean?

Problems

1. Most companies that manufacture light bulbs advertise their 100-watt bulbs as having a life of 750 hours (Actually, this is an “it depends” statistic. In this case, it depends partly on the number of times the light is turned on and partly on whether or not the light is ventilated). Let’s assume a standard deviation of 100 hours. A consumer organization bought 50 bulbs and burned them until they failed. For the 50 bulbs, the mean number of hours was 725. What is the probability of obtaining a mean this low or lower if the population mean for this brand of bulbs is 750 hours? Write a sentence about advertised claims.
2. Refer to short-answer question 3. Suppose you had the weights of the 49 vegetarians in kilograms. The sum of these weights was 3616 kg. Calculate the mean. What is the probability of obtaining such a mean weight (or one smaller) from a population with a mean of 75.4 kg (166 pounds) and a standard deviation of 10 kg (22 pounds)?
3. Stanley Milgram published a study that became famous because he concluded that average Americans follow orders that lead to apparent injury to others. In Milgram’s experiment, a cross-section of Americans increased the apparent shock they were administering to a fellow participant to an average of 285 volts. Suppose a psychologist at a small, New England college with humanitarian ideals was sure that students at her college would not be so cruel. She set up the same apparatus and procedures and gathered data on a random sample of 36 students from her school. She found

$$\Sigma X = 11,340$$

$$\Sigma X^2 = 3,698,100$$

Construct a 99 percent confidence interval about the mean. Interpret this confidence interval by describing students at the small college.

4. A teacher was interested in the mathematical ability of graduating high school seniors in her state. She gave a 32-item test to a random sample of 75 seniors with the following results: $\Sigma X = 1275$, $\Sigma X^2 = 23,525$. Establish a 95% confidence interval about the sample mean, and write a sentence that explain the interval you found.

Chapter 7

5. The instructor of a popular course on health used a number of films in her course. After a film, each student filled out a questionnaire. High scores indicated that the film was valuable. Over the previous five years, the scores were negatively skewed with a mean of 33 and a standard deviation of 2.40. This year the instructor showed students a new film on smoking, and got the following statistics: $\sum X = 1224$, $\sum X^2 = 41818$. Calculate a 95% confidence interval about the sample mean and tell what the students thought of the film.

6. In the field of testing it is common to design a test so that the population mean is 50. The California Personality Inventory (CPI), which is designed to access the general population, has these characteristics. One of the scales of the CPI assesses socialization. Suppose that a researcher thought that fraternity and sorority members are more sociable than the general population. She gathered data on 49 Greeks, finding a mean of 52 and a standard deviation of 10. Find the 99% confidence interval about the mean. Do the data indicate that Greeks are more sociable than average?

7. Draw a random sample with $N = 6$ from the following scores. Write down each step in your procedure.

21 31 17 13 02 09 57 26 72 140

8. From that sample in Question 7, what is the standard error of the mean? Write a sentence explaining what standard error of the mean reveals about the population mean.