UNIT 4

Inference from Data: Principles

TOPIC 16

Confidence Intervals: Proportions

The generation of children being raised in the 2000s has been dubbed Generation M, because of the extent to which various forms of media (hence the *M*) permeate their lives. The Kaiser Family Foundation commissioned an extensive survey to investigate this phenomenon. In this topic, you will learn more about how to use sample results to estimate population values. For example, what proportion of all American teens have a television in their room? Is this proportion higher for boys or for girls? Is the proportion higher or lower for other forms of media, such as CD players and video game players and computers? How can we answer these questions without asking all American teens?

Overview

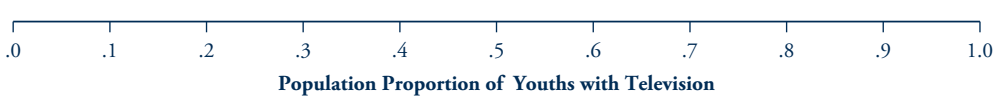
In the last unit, you explored how sample statistics vary from sample to sample. You studied this phenomenon empirically through simulations and theoretically with the Central Limit Theorem. You learned that this variation has a predictable long-term pattern. This pattern enables you to make probability statements about sample statistics, provided you know the value of the population parameter. These probability statements allow you to turn the tables and address the much more common goals of estimating and making decisions about an unknown population parameter based on an observed sample statistic. These are the goals of statistical inference.

There are two major techniques for classical statistical inference: confidence intervals and tests of significance. Confidence intervals seek to estimate a population parameter with an interval of values calculated from an observed sample statistic. Tests of significance assess the extent to which sample data refute a particular hypothesis concerning the population parameter. This topic extends your study of the concept of statistical confidence, begun in Topic 13, by introducing you to confidence intervals for estimating a population proportion.

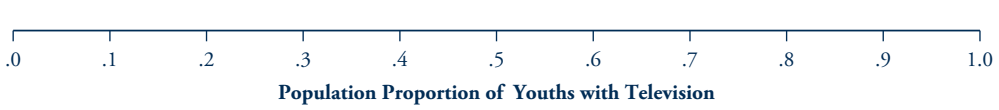
Preliminaries

1. Guess the proportion of American youth from ages 8–18 who have a television in their bedroom. (Activity 16-1)

2. Mark on the following number line an interval that you believe with 80% confidence to contain the actual proportion of American youth from ages 8–18 who have a television in their bedroom. (In other words, if you were to create a large number of these intervals, 80% of them should succeed in *capturing*—containing within the interval—the actual population proportion.) (Activity 16-3)



3. Mark on the following number line an interval that you believe with 99% confidence to contain this population proportion. (Activity 16-3)



4. Which of these two intervals is wider, the 80% interval or the 99% interval? Explain why this makes sense. (Activity 16-3)

In-Class Activities

Activity 16-1: Generation M 3-8, 4-14, 13-6, 16-1, 16-3, 16-4, 16-7, 16-25, 16-26, 21-11, 21-12

Do you have a television in your room?

Recall from Exercise 3-8 the Kaiser Family Foundation commissioned an extensive survey in 2004 that investigated the degree to which American youth from ages 8–18 have access to various forms of media. They distributed written questionnaires to a random sample of 2032 youths in this age range. One of the many questions asked youths whether or not they have a television in their bedroom, to which 68% answered yes.

a. What are the observational units in this study?

each american youth 8-18

b. Identify the variable, and classify it as quantitative or categorical.

television? cat bin

Because the variable of interest is categorical and binary, the relevant parameter of interest is a *proportion*.

c. Is .68 a parameter or a statistic value? What symbol would you use to represent it?

d. Describe in words the relevant population parameter of interest in this study. What symbol would you use to represent it?

proportion of all american youths 8-18 that have a television in their room

e. Does the Kaiser survey allow the researchers to determine the exact value of the parameter? Explain.

f. Is the Central Limit Theorem for a sample proportion valid in this context? If so, what does it predict for the shape, center, and spread of the sampling distribution of the sample proportion? Include a sketch of this distribution, indicating with symbols the mean and standard deviation.

probably(we don't know parameter tho), center is parameter, shape normal, spread is sqrt (parameter*(1-para)/2032)

g. Based on the distribution in part f, in general, how close do you expect the sample proportion to be to the population proportion? [*Hint:* Apply the empirical rule.]

95% of the time it deviates 2 SD

h. Based on your answer to g, how close do you expect the population proportion to be to the sample proportion?

Sane as abvoe
sanae

These questions deal with the issue of statistical confidence as was briefly introduced in Topic 15. Although a sample statistic from a simple random sample provides a reasonable estimate of a population parameter, you certainly do not expect the sample statistic to equal the (unknown) population parameter exactly (sampling variability). It is likely, however, that the unknown parameter value is in the ballpark of the sample statistic. The purpose of confidence intervals is to use the sample statistic to construct an interval of values that you can be reasonably confident contains the actual, though unknown, parameter.

Applying the CLT, we see that extending two standard deviations on either side of the sample proportion $\hat{p}$ produces an *interval* (a set of plausible values) that succeeds in capturing the value of the population proportion, π, for about 95% of all samples, that is, within plus or minus $2 \times \sqrt{\pi(1-\pi)/n}$.

However, you cannot use this formula in practice because you do not know the value of π. (Indeed, the whole point of establishing the interval is to estimate the value of π based on the sample proportion $\hat{p}$.)

i. What value (in general) seems like a reasonable replacement to use as an estimate for π in this standard deviation expression?

The estimated standard deviation of the sample statistic $\hat{p}$

$$SE_{\hat{p}} = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

is called the **standard error** of $\hat{p}$.

16

j. Calculate the standard error of $\hat{p}$ for the Kaiser survey about the proportion of American children who have a television in their bedroom.

.01

k. Now, go two standard errors on either side of $\hat{p}$ by doubling this standard error, subtracting the result from $\hat{p}$, and adding it to $\hat{p}$. This technique forms a reasonable interval estimate (set of plausible values) of π, the population proportion of all American 8–18-year-olds who have a television in their bedrooms.

.66-.70

l. Do you know for sure whether or not the actual value of π is contained in this interval?

we are 95% sure

All you can say for certain is that if you were to construct many intervals this way using different random samples from the same population, then approximately 95% of the resulting intervals would contain π. Note that it is not technically appropriate to say that this interval (.66, .70) has a .95 probability of containing π (see Activity 16-5 for more discussion on this point). This has led statisticians to coin a new term, **confidence,** to describe the level of certainty in the interval estimate, and so your conclusion is that you are 95% confident that the interval you calculated contains π.

there is a minor distinction, consider this
before we flip a coin, probability of heads is 1/2.

once i flip the coin, and don't show results, we aren't sure what result it is, so we say we are 1/2 confident in it being heads

it could be argued that before an observation even if the event is occurred, we can say they are the same, but it is more of a philosophical interpretation.

The other complication is that so far we have limited you to the value 2 for the number of standard errors and the resulting approximate 95% confidence interval. Returning to the empirical rule, you know that going just one standard error (standard deviation) on either side of $\hat{p}$ corresponds to 68%, whereas going three standard errors corresponds to 99.7%. By using a different multiple of the standard error to vary the distance that you go from $\hat{p}$, you can vary the **confidence level,** which is the measure of how confident you are that the interval does, in fact, contain the true parameter value. You find this multiplier, called the **critical value,** using the normal distribution, and you will denote it by z^* [see Activity 16-2]. You specify the confidence level by deciding the level of confidence necessary for the given situation; common values are 90%, 95%, and 99%. The confidence level, in turn, determines the critical value.

Confidence interval for a population proportion π:

$$\hat{p} \pm z^* \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

where $\hat{p}$ denotes the sample proportion, n is the sample size, and z^* represents the critical value from the standard normal distribution for the confidence level desired.

The "±" in this expression means that you subtract the term following it from the term preceding it to get the lower endpoint of the interval, and then you add the terms to get the upper endpoint of the interval. This expression is often denoted as

$$\left(\hat{p} - z^* \sqrt{\frac{\hat{p}\,(1-\hat{p})}{n}},\ \hat{p} + z^* \sqrt{\frac{\hat{p}\,(1-\hat{p})}{n}}\right)$$

For this confidence interval procedure to be valid, the sample size needs to be large relative to the value of the population proportion so that the Central Limit Theorem for a sample proportion applies. This assures you that the sampling distribution of $\hat{p}$ is approximately normal. To check this condition, you will see whether $n\hat{p} \geq 10$ and whether $n(1 - \hat{p}) \geq 10$, which says there must be at least 10 successes and 10 failures in the sample. If this condition is not met, then the stated confidence level may be inaccurate, and the interval could be very misleading.

Furthermore, if you do not have a random sample from the population of interest, then the confidence interval may not be estimating the desired parameter. Without a random sample, you might still be able to argue that the sample is representative of the population of interest, but you should only do so with caution.

Therefore, before interpreting your confidence interval, make sure these technical conditions are met.

Technical conditions:

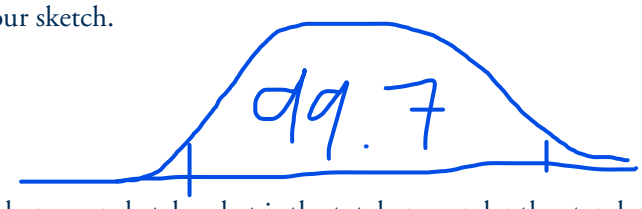
- The sample is a simple random sample from the population of interest. Check the data collection protocol.
- The sample size is large relative to the value of the population proportion. Check whether or not $n\hat{p} \geq 10$ and $n(1 - \hat{p}) \geq 10$.

- Independent - if sample size < 10% population size (ie 10n < N)

Activity 16-2: Critical Values 12-18, 16-2, 16-20, 19-13

Before calculating a confidence interval, you need to be able to determine the critical value, z^*. As an example, suppose you want to find the value of z^* for a 98% confidence interval. In other words, find the value z^* such that the area under the standard normal curve between $-z^*$ and z^* is equal to .98 by following these steps:

a. Draw a sketch of the standard normal curve and shade the area corresponding to the middle 98% of the distribution. Also indicate roughly where $-z^*$ and z^* fall in your sketch.



b. Based on your sketch, what is the total area under the standard normal curve to the left of the value z^*?

1%

c. Use the Standard Normal Probabilities Table (Table II) or technology to find the value (z^*) that has this area (your answer to part b) to its left under the standard normal curve.

z* = 2.326347877 SD

This value z^*, which you should have found to be approximately 2.33, is called the **upper .010 critical value** of the standard normal distribution because the area to its right under the standard normal curve is .010. This critical value is used for 98% confidence intervals.

d. Repeat parts a–c to find the critical value corresponding to a 95% confidence interval.

2.5% 1.959963

[*Note:* 1.96 is the more exact value that we previously approximated as being 2.]

Critical values for any confidence level can be found in this manner. Some commonly used confidence levels and critical values are listed here:

Confidence Level	80%	90%	95%	99%	99.9%
Critical Value z^*	1.282	1.645	1.960	2.576	3.291

Activity 16-3: Generation M

3-8, 4-14, 13-6, 16-1, **16-3,** 16-4, 16-7, 16-25, 16-26, 21-11, 21-12

a. Recall that the survey of 2032 children in Activity 16-1 found that 68% had a television in their bedrooms. Use this sample to construct a 95% confidence interval for π, the proportion of all American 8–18-year-olds who have a television in their bedrooms. [*Hint:* Now use the more exact z^* value instead of 2 for the multiplier.]

0.6597 to .7003

b. Write a sentence interpreting what this interval reveals.

we are 95% confident that the population proportion(parameter) of youths 8-18 that have a television in their bedroom lie within this interval

c. Can you be *certain* that this interval contains the actual value of π?

no

d. Calculate the *width* of this interval. [*Hint:* The width is the difference between the upper and lower endpoints of the interval.]

.04

e. Determine the *half-width* of this interval. [*Hint:* The half-width is the width divided by 2.]

.02

The half-width of a confidence interval is often called the survey's **margin-of-error.**

f. Explain how you could have determined the half-width of the interval without first calculating the width. [*Hint:* What part of the confidence interval formula corresponds to this margin-of-error?]

upper - phat

g. Determine the midpoint of this interval. [*Hint:* You can find the interval's midpoint by calculating the average of its endpoints.]

phat

h. Does this midpoint value look familiar? Explain why this value makes sense, based on how the confidence interval is constructed.

it appears familar

i. What technical conditions underlie the validity of this procedure? Check whether or not the conditions seem to be satisfied here.

r, i yes, n is yes because 10 and 10 win/l

j. How would you expect a 99% confidence interval to differ from the 95% confidence interval? Provide an intuitive explanation for your answer without doing the calculation and not based on the formula.

iwider

k. Determine a 99% confidence interval for the population proportion of American 8–18-year-olds who have a television in their bedrooms.

0.6533 .7067

l. How does the 99% interval compare with the 95% interval? Compare the midpoints as well as the margin-of-error.

no

As the confidence level increases, the margin-of-error and width of the confidence interval also increase. The midpoint of the interval remains at the sample proportion ($\hat{p}$) regardless of the confidence level.

16

m. Based on these intervals, does it seem plausible that .5 is the proportion of all American youths who have a television set in their bedrooms? Does .75 seem plausible? What about two-thirds? Explain.

.50: not possible .75: not possible Two-thirds: possible

Explanation: we are less thant .5% sure for the first two, but 99% sure of the third

The survey also provided sample results broken down separately for boys and girls. There were 1036 girls and 996 boys in the sample. Among the boys, 72% had a television in their bedroom, as compared to 64% of the girls.

n. Determine a 95% confidence interval for the proportion of American boys in this age group who have a television in their bedrooms. Then do the same for the analogous proportion of American girls.

Boys: .6921 to .7479

Girls: .6107 - .6692

o. How do these intervals compare? Do the intervals seem to indicate that the proportion of boys who have a television in their bedrooms is different from that proportion for girls? Explain.

boys are larger range

p. Report the margin-of-error for these intervals. Are these greater than or less than (or the same as) the margin-of-error based on the entire sample (from part e)? Explain why this makes sense intuitively.

.0279 boys
.02925 girls

greater tan the one found in part e

A larger sample size produces a narrower confidence interval whenever other factors remain the same. In this case, we generally say our estimate of the population parameter is more *precise*.

Watch Out

A confidence interval is just that—an *interval*—so it includes all values between its endpoints. For example, the 95% confidence interval (.660, .700) means you are 95% confident that the population proportion of 8–18-year-olds who have a TV in their bedrooms is somewhere between .660 and .700. Do not mistakenly think that only the endpoints matter or that only the margin-of-error matters. The midpoint and actual values within the interval matter.

Activity 16-4: Generation M 3-8, 4-14, 13-6, 16-1, 16-3, 16-4, 16-7, 16-25, 16-26, 21-11, 21-12

Reconsider the Kaiser Foundation survey that found 68% of a sample of 2032 American youth had a television in their rooms. You found this allowed you to estimate the population proportion with 95% confidence with a $z^*\sqrt{\hat{p}(1-\hat{p})/n} = 1.96\sqrt{.68(.32)/2032} = .0203$ margin-of-error.

a. Suppose you intend to do a follow-up study, and you want the margin-of-error to be no larger than .03. That is, you want to estimate the population proportion with a TV in their room to within $\pm .03$. Keeping the confidence level at 95%, do you think this will require a larger or a smaller sample size or the same sample size as the original study? Explain your reasoning.

smaller sample size = smaller margin of area since less variability

b. Set the expression for the margin-of-error in the 95% confidence interval (the half-width of the interval) equal to .03. Use .68 as a ballpark guess for the value of $\hat{p}$ you will find, based on the current results. Now the only unknown in the equation should be the sample size n. Rearrange the terms in this equation algebraically to solve for n.

n=928.8

.03 = 1.96* sqrt(.68 * .32 / n)

Because you cannot sample a fraction of a person, the convention is to round up to the next integer when performing these sample size calculations for the margin-of-error to be less than or equal to the desired value.

c. How does this sample size compare to that of the original study? Was your prediction in part a correct? Also, explain in your own words what this sample size calculation has revealed.

sample size is smaller

d. Now suppose you wanted to do a follow-up study with at most a .0203 margin-of-error as with the original study, but with a 99% confidence level instead of 95%. Do you think this will require a larger or a smaller sample size than the original study? Explain your reasoning.

larger

e. Solve for the sample size necessary for this follow-up study, and compare your results to part b.

n=3500

.0203 = .25758 * sqrt(.58 * .32/n)

f. If we did not even have the original study from which to obtain an initial estimate for $\hat{p}$ in this equation, we could use .5. This is a conservative approach because the margin of error is largest when $\hat{p}$ equals .5. Repeat parts b and e using .5 instead of .68, and compare the results.

n=1068

.0203 = 1.96* sqrt(.25/n)

4025.68

This activity reveals that you can determine the necessary sample size before conducting a study, in order to obtain a desired margin-of-error with a certain confidence level.

Watch Out

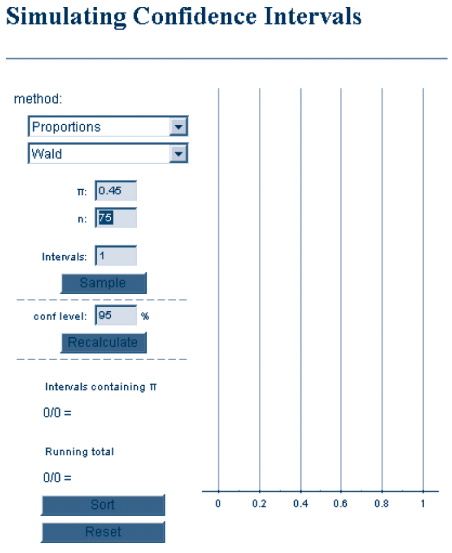
Many people believe you would need a much larger sample size than 1000 to describe the population of (millions and millions of) American teens. But if the sample is randomly selected, you showed a sample size of roughly 1000 is sufficient to obtain a sample percentage within 3 percentage points of the population percentage (with 95% confidence).

Activity 16-5: Candy Colors 1-13, 2-19, 13-1, 13-2, 13-15, 13-16, 13-20, 16-5, 16-22, 24-15, 24-16

To help us understand properties of confidence intervals, let's examine a hypothetical situation where we can control the population parameter.

Assume for the moment that 45% of all Reese's Pieces are orange, that is, that the population proportion of orange candies is $\pi = .45$. To examine more closely what the interpretation of "confidence" means in this statistical context, you will use an applet to simulate confidence intervals for 200 random samples of 75 candies each.

a. Open the `Simulating Confidence Intervals` applet and make sure that it is set for **Proportions** with the **Wald** method (the name for the method you just learned). Set the population proportion value to $\pi = .45$, the sample size to $n = 75$, the number of intervals to **1**, and the confidence level to **95%**. Click **Sample.** The horizontal line that appears represents the confidence interval resulting from the sample.



.302 .525

Click on the line, and report the endpoints of the interval. Does this interval succeed in capturing the true value of the population proportion (which you have set to be $\pi = .45$)?

b. Click **Sample** four more times. Does the interval change each time? Do all of the intervals succeed in capturing the true value of π?

yes, yes

c. Now click **Reset** and then change the number of intervals to **200** and click **Sample**. How many and what proportion of these 200 intervals succeed in capturing the true value of π? (Notice that the intervals that succeed are colored green and those that fail to capture π are colored red.)

188/200=94%

d. Click **Sort**. What do you notice about the intervals that fail to capture the true value of π? [*Hint:* What is true about those intervals' sample proportions, or midpoints?]

extreme intervals, midpt is more than 2SD

e. If you had taken a *single* sample of 75 candies (in a real situation, not in this applet environment), would you have any *definitive* way of knowing whether or not your 95% confidence interval contained the actual value of π? Explain.

no

f. Keep clicking **Sample** until you have generated 1000 samples and intervals. Report the Running Total of intervals that succeed in capturing the true value of π and the percentage of intervals this represents.

.95

g. Is this percentage close to 95%? Should it be? Explain what this simulation reveals about the phrase "95% confident."

yes, in the long run 95% of similarly constructed intervals will succeed in capturing the population proportion/parameter.

This simulation illustrates the following points about the proper use and interpretation of confidence intervals:

- Interpret a 95% (for instance) confidence interval by saying that you are 95% *confident* that the interval contains the actual value of the population proportion.
- More specifically, interpret "95% confidence" to mean that if you repeatedly take simple random samples from the population and construct 95% confidence intervals for each sample, then in the long run approximately 95% of these confidence intervals will succeed in capturing the actual population proportion.

h. Predict how the resulting confidence intervals will change if the sample size were increased to 150. [*Hint:* Think about width and the percentage that will contain the parameter.]

smaller, since dividing by sample size

i. Change the sample size to **150** and generate 1000 intervals (200 intervals at a time). How many and what proportion of the intervals succeed in capturing the actual value of π? Is this reasonably close to the percentage from part f? What is different about these intervals compared to those generated with a sample size of 75 candies?

94.6% they change m ore ig

j. Before you change the confidence level to 90%, predict *two* things that will change about the intervals. [*Hint:* One change concerns their length, and the other concerns their success rate.]

gap will be smaller, the success rate will lower

k. Change the confidence level to **90%** and click **Recalculate**. How many and what proportion of the intervals succeed in capturing the actual value of π? Is this reasonably close to the percentages from parts f and i? What is different about these intervals?

91.4%

smaller ig, since intervals they sample are much smaler

Watch Out

- It is incorrect to say that π has a .95 probability of falling within the 95% confidence interval. It is also technically incorrect to say that the probability is .95 that a particular calculated 95% confidence interval contains the actual value of π. The technicality here is that π is not random; it is some fixed (but unknown) value. What is random, changing from sample to sample, is the sample proportion and thus the interval based on it. Thus the probability statement applies to what values an interval will take prior to the sample being collected (i.e., to the *method*), not whether or not a particular interval contains the fixed parameter value once it has been calculated. If you did have all intervals from all possible samples in a bag, the probability that you will randomly select an interval that contains π is .95.
- Confidence intervals estimate the value of a population *parameter;* they do not estimate the value of a sample statistic or of an individual observation.

the process itself has a 95% confidence interval, but if you select it the "proabbiltiy" of it not being the value itself is not 95%???? idk

Self-Check

Activity 16-6: Kissing Couples

15-1, **16-6,** 17-1, 17-2, 17-12, 18-1, 24-4, 24-14

Recall from Activity 15-1 the study in which a sample of 124 kissing couples were observed, with 80 of them leaning their heads to the right.

a. Identify the observational units and variable in this study.

b. Describe the sample data numerically and graphically.

c. Determine and interpret a 95% confidence interval for the population proportion of all kissing couples who lean to the right. Also provide an interpretation of what is meant by "95% confidence" in this context.

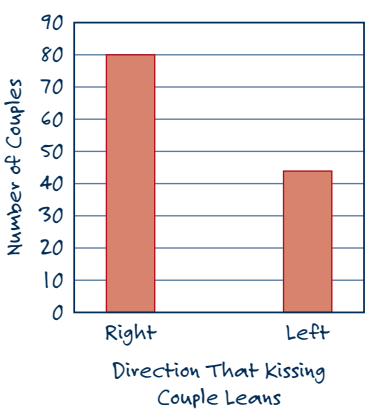
d. Also determine a 90% and a 99% confidence interval for this population proportion. Comment on how these intervals compare.

e. Based on these intervals, comment on whether or not it seems plausible to believe that 50% of all kissing couples lean to the right. Also comment on the plausibility of 2/3, and of 3/4, as the value of the population proportion of kissing couples who lean to the right.

f. Comment on whether or not the technical conditions necessary for these confidence intervals to be valid are satisfied here.

Solution

a. The observational units are the kissing couples. The variable is which direction they lean their heads while kissing.



b. The sample consists of the 124 kissing couples observed by the researchers in various public places. The statistic is the sample proportion of couples who lean to the right when kissing: $\hat{p} = 80/124 = .645$. A bar graph is shown here.

c. A 95% confidence interval for the population proportion of couples who lean to the right is

$$.645 \pm 1.96\sqrt{\frac{(.645)(1-.645)}{124}}$$

which is .645 ± .084, or (.561, .729). You are 95% confident that the population proportion of kissing couples who lean to the right is somewhere between .561 and .729. This "95% confidence" means that if you were to take many random samples and generate a 95% confidence interval (CI) from each, then in the long run, 95% of the resulting intervals would succeed in capturing the actual value of the population proportion, in this case the proportion of all kissing couples who lean their heads to the right.

d. A 90% CI is

$$.645 \pm 1.645\sqrt{\frac{(.645)(1-.645)}{124}}$$

which is .645 ± .071, or (.574, .716).
A 99% CI is

$$.645 \pm 2.576\sqrt{\frac{(.645)(1-.645)}{124}}$$

which is .645 ± .111, or (.534, .756).

The higher confidence level produces a wider confidence interval. All of these intervals have the same midpoint: the sample proportion .645.

e. Because none of these intervals includes the value .5, it does not appear to be plausible that 50% of all kissing couples lean to the right. In fact, all of the intervals lie entirely above .5, so the data suggest that more than half of all kissing couples lean to the right. The value 2/3 is quite plausible for this population proportion, because .667 falls within all three confidence intervals. The value 3/4 is not very plausible, because only the 99% CI includes the value .75; the 90% CI and 95% CI do not include .75 as a plausible value.

f. The sample size condition is clearly met, as $n\hat{p} = 80$ is greater than 10, and $n(1 - \hat{p}) = 44$ is also greater than 10. But the other condition is that the sample be randomly drawn from the population of all kissing couples. In this study, the couples selected for the sample were those who happened to be observed in public places while the researchers were watching. Technically, this is not a random sample, and so you should be cautious about generalizing the results of the confidence intervals to a larger population.

Watch Out

Don't forget about the necessity for random sampling. When you don't have a true random sample, which may often be the case, proceed cautiously and ask yourself what population

you might be willing to believe the sample results represent. For example, in this kissing study, you would probably generalize the results only to the population of couples who kiss in public places in the countries where the researchers gathered their data.

Wrap-Up

This topic provided your first formal exposure to statistical inference by introducing you to confidence intervals, a widely used technique. You learned how to construct a confidence interval for a population proportion, and you examined how to interpret both the resulting interval and also what the **confidence level** means. For example, using a 95% confidence level gives you "95% confidence" that the interval contains the actual value of the (unknown) population parameter. As you saw by exploring the `Simulating Confidence Intervals` applet, saying you are 95% confident means that 95% of all intervals generated by the procedure in the long run will succeed in capturing the unknown population value.

You also investigated the effects of sample size and confidence level on the interval and its **margin-of-error.** The ideal confidence interval is very narrow with a high confidence level. But there's a trade-off: using a higher confidence level produces a wider interval if all else remains the same. One solution is to use a larger sample because a larger sample produces a narrower interval (for the same level of confidence). This sample size issue explains why the margin-of-error was greater for estimating the proportion of 8–18-year-old boys with a TV than for estimating the proportion of 8–18-year-olds (boys and girls combined) with a TV. You also learned how to plan ahead by determining, prior to collecting data, the sample size needed to achieve a certain margin-of-error for a given confidence level.

In Brief

Some useful definitions to remember and habits to develop from this topic are

- The purpose of a **confidence interval** is to estimate the value of a population parameter with an interval of values that you believe with high confidence to include the actual parameter value. The confidence level indicates how confident you are in the interval.
- The general form of all confidence intervals in this course is *estimate* $\pm$ *margin-of-error,* where *margin-of-error* = (*critical value*) $\times$ (*standard error of estimate*). The estimate is a sample statistic, calculated from sample data. The **standard error** is an estimate of that statistic's standard deviation (how much the statistic varies from sample to sample), also calculated from sample data.
- A specific confidence interval procedure for estimating a population proportion is
$$\hat{p} \pm z^* \sqrt{\frac{\hat{p}\,(1-\hat{p})}{n}}.$$
- The margin-of-error is affected by several factors, primarily
  - A higher confidence level produces a greater margin-of-error (a wider interval).
  - A larger sample size produces a smaller margin-of-error (a narrower interval).
- Common confidence levels are 90%, 95%, and 99%. The phrase "95% confidence" means that if you were to take a large number of random

16

samples and use the same confidence interval procedure on each sample, then in the long run 95% of those intervals would succeed in capturing the actual parameter value. Note that this is not the same as saying there is a 95% probability that the parameter is inside the calculated interval.

- Always check the technical conditions before applying this procedure. The sample is considered large enough for this procedure to be valid as long as $n\hat{p} \geq 10$ and $n(1 - \hat{p}) \geq 10$. If this condition is not met, then the normal approximation of the sampling distribution is not valid and the reported confidence level may not be accurate.
- Always consider how the sample was selected to determine the population to which the interval applies. If the sample was randomly selected from the population of interest, then the interval applies to that population.
- Remember that you can plan ahead to determine the sample size necessary to achieve a desired margin-of-error for a given confidence level.

You should be able to

- Calculate and interpret a confidence interval for a population proportion. (Activities 16-1, 16-3)
- Determine critical values from the normal distribution to be used in calculating confidence intervals. (Activity 16-2)
- Explain the effects of sample size and confidence level on a confidence interval. (Activities 16-3, 16-5)
- Determine the sample size necessary to achieve a certain margin-of-error with a given confidence level. (Activity 16-4)
- Explain what confidence level means in terms of repeated random sampling from a population. (Activity 16-5)

In the next topic, you will continue to study inference procedures for a population proportion. You will investigate the second major type of statistical inference procedure: a test of significance.

Exercises

Exercise 16-7: Generation M

3-8, 4-14, 13-6, 16-1, 16-3, 16-4, 16-7, 16-25, 16-26, 21-11, 21-12

Recall the study from Activity 16-1 about media in the lives of American youths aged 8–18. Among the 2032 youths who were sampled, 49% reported having a video game player and 31% reported having a computer in their bedroom.

a. Determine and interpret a 95% confidence interval for the population proportion of American youths who have a video game player in their bedrooms.
b. Repeat part a for the proportion who have a computer in their bedrooms.
c. Determine the margin-of-error (half-width) for each of these confidence intervals.
d. Does a survey's margin-of-error depend on more than its sample size and the confidence level? Explain.

Exercise 16-8: Cursive Writing

13-14, 16-8

Recall from Exercise 13-14 that an article in the October 11, 2006, issue of the *Washington Post* claimed that 15% of high school students used cursive writing on the essay portion of the SAT exam in the academic year 2005–2006.

Suppose you want to design a study to estimate this proportion for the following year.

a. How many essays would you need to sample to estimate this proportion to within ± .01 with 99% confidence? (Use the 2005–2006 proportion as a reasonable starting guess for how the following year's proportion will turn out.)
b. Describe the two changes you could make to this goal (estimating this proportion to within ± .01 with 99% confidence) if you could not afford to sample so many essays.

Exercise 16-9: Penny Activities

16-9, 16-10, 16-11, 17-9, 18-14, 18-15

a. Flip a penny in the air 50 times and keep track of how many heads and tails result. Use your sample results to construct a 95% confidence interval for the probability (long-term relative frequency) that a *flipped* penny lands heads.
b. Instead of flipping a penny in the air, consider spinning it on its side and seeing whether it lands on heads or tails. Do this 50 times, and use your sample results to construct a 95% confidence interval for the probability that a *spun* penny lands heads.
c. Now consider balancing a penny on its side and then striking the table nearby so the penny tilts to land heads or tails. Do this 50 times, and use your sample results to construct a 95% confidence interval for the probability that a *tilted* penny lands heads.
d. Based on these results, which activities (flipping, spinning, or tilting) could plausibly be 50/50 for landing heads or tails? Explain.
e. Which activity appears to result in the highest probability of landing heads?

Exercise 16-10: Penny Activities

16-9, **16-10,** 16-11, 17-9, 18-14, 18-15

Refer to the previous exercise.

a. Determine the sample size necessary to estimate the probability that a *flipped* penny lands heads to within a margin-of-error of ± .02 with confidence level 95%. [*Hint:* Use your estimate of this probability from part a in the previous exercise.]
b. How would you expect this sample size to change if you increased the confidence level to 99%? Explain.
c. Perform the sample size calculation requested in part b, and see whether your prediction was right.
d. How would you expect this sample size to change if you also adjusted the margin-of-error to ± .01 with 99% confidence? Explain.
e. Perform the sample size calculation requested in part d, and see whether your prediction was right.

Exercise 16-11: Penny Activities

16-9, 16-10, **16-11,** 17-9, 18-14, 18-15

Refer to the previous two exercises.

a. Determine the sample size necessary to estimate the probability that a *spun* penny lands heads to within a margin-of-error of ± .02 with a confidence level of 95%. Use your estimate from part b of Exercise 16-9.
b. Determine the sample size necessary to estimate the probability that a *tilted* penny lands heads to within a margin-of-error of ± .02 with a confidence level of 95%. Use your estimate from part c of Exercise 16-9.
c. With which penny activity (flipping, spinning, or tilting) is the necessary sample size largest? Why do you think this is the case?

Exercise 16-12: Credit Card Usage

1-9, **16-12,** 19-23, 20-10

Recall from Exercise 1-9 that the Nellie Mae organization conducts an extensive annual study of credit card usage by college students. For their 2004 study, they analyzed credit bureau data for a random sample of 1413 undergraduate students between the ages of 18 and 24. They found that 76% of the students sampled held a credit card.

a. Determine this result's margin-of-error with 95% confidence.
b. Determine and interpret a 95% confidence interval for the proportion of all undergraduate college students in the United States who held a credit card in 2004.
c. Comment on whether or not the technical conditions required for the validity of this procedure are satisfied here.

Exercise 16-13: Responding to Katrina

2-12, 4-10, **16-13**

On September 8–11, 2005, less than two weeks after the destruction caused by Hurricane Katrina,

a CNN/*USA Today*/Gallup poll asked, "Just your best guess, do you think one reason the federal government was slow in rescuing these people was because many of them were black, or was that not a reason?" Of the 848 non-Hispanic white adults who were interviewed, 12% said yes. Of the 262 black adults interviewed, 60% said yes.

a. Determine the margin-of-error and a 95% confidence interval for the population proportion of white adults who answered yes.
b. Repeat for the black adults.
c. Compare these intervals (not just their widths).
d. Which group has the greater margin-of-error? Explain why.
e. What additional information (not given above) would you like to know in order to determine if these intervals are valid?

Exercise 16-14: *West Wing* Debate

16-14, 18-6

On Sunday, November 6, 2005, the popular television drama *The West Wing* held a live debate between two fictional candidates for president. Immediately afterward, an MSNBC/Zogby poll found that 54% favored Democratic Congressman Matt Santos, played by Jimmy Smits, whereas 38% favored Republican Senator Arnold Vinick, played by Alan Alda. The poll was conducted online with a sample of 1208 respondents; the Zogby company screened the online respondents to try to ensure they were representative of the population of adult Americans.

a. Are the conditions met for the confidence interval procedure to be valid with these data? Explain.
b. Produce and interpret a 95% confidence interval for the population proportion who favored Santos.
c. Based only on this confidence interval, do the sample data suggest that more than half of the population favored Santos? Explain.

Exercise 16-15: Magazine Advertisements

16-15, 21-21

The September 13, 1999, issue of *Sports Illustrated* contained 116 pages, and 54 of those pages contained an advertisement. The September 14, 1999, issue of *Soap Opera Digest* consisted of 130 pages, including 28 pages with advertisements.

a. What are the observational units in this study?
b. For each magazine, treat this issue's pages as a sample from the population of all the magazine's pages over all issues, and construct a 95% confidence interval for the population proportion of pages that contain ads.
c. Write a sentence or two interpreting these intervals. (Be sure to relate your interpretation to the context.)
d. Clearly explain what is meant by the phrase "95% confidence" in this context.
e. Does each interval contain the sample proportion of pages with ads?
f. Explain why the previous question is silly and did not require you to even look at the intervals.
g. Find a recent issue of another magazine and repeat this analysis to produce a confidence interval for the proportion of its pages that contain ads.

Exercise 16-16: Phone Book Gender

4-17, 16-16, 18-11

Suppose you want to estimate the proportion of women among the residents of San Luis Obispo County, California. A random sample of columns from the phone book reveals 36 listings with both male and female names, 77 listings with a male name only, 14 listings with a female name only, 34 listings with initials only, and 5 listings with a pair of initials.

a. How many first names are supplied in these listings altogether? [*Hint:* Ignore the listings with initials, and remember to count both male and female names for the couples.]
b. How many and what proportion of those names are female?
c. Use these sample data to form a 90% confidence interval for the proportion of women in San Luis Obispo County.
d. Do you have any concerns about the sampling method that might render this interval invalid? Explain.
e. Suggest a more reasonable, but still practical, sampling method for estimating the proportion of women in the county.

Exercise 16-17: Random Babies

11-1, 11-2, 14-8, 16-17, 20-16, 20-22

Reconsider the simulation data collected by the class in Activity 11-1.

a. What proportion of the simulated repetitions resulted in no mothers getting the correct baby?
b. Use this simulated sample information to form a 95% confidence interval for the long-term proportion of times that no mother would get the right baby.
c. Write a sentence or two interpreting this interval. (Be sure to relate your interpretation to the context.)

This is a rare instance in which you can actually calculate the population parameter and check whether or not the confidence interval succeeds in capturing it. The parameter here is the theoretical probability of zero matches, which you calculated in Activity 11-2.

d. Report the value of this population parameter.
e. Does the 95% confidence interval succeed in capturing the population parameter?
f. If 1000 different statistics classes carried out this simulation and calculated a 95% confidence interval as you did in part b, approximately how many of their intervals would you expect to succeed in capturing the parameter value? Explain.
g. Use the simulated sample data to form an 80% confidence interval for the parameter. Does this interval succeed in capturing its value? Re-answer f for this confidence level.

Exercise 16-18: Charitable Contributions

16-18, 18-7

The 2004 General Social Survey found that 78.9% of the adult Americans interviewed claimed to have given a financial contribution to charity in the previous 12 months.

a. Describe the parameter value of interest in this situation.
b. If the survey had involved 250 households, determine a 99% confidence interval for this parameter.
c. Repeat part b, supposing the survey had involved 500 households.
d. Repeat part b, supposing the survey had involved 1000 households.
e. Repeat part b, supposing the survey had involved 2000 households.
f. Compare the margin-of-error for these four intervals. Describe how they are related. (Be as specific as possible.)
g. Does doubling the sample size cut the margin-of-error in half? If not, by what factor does the sample size have to be increased in order to cut the margin-of-error in half?
h. The actual survey involved 1334 households. Determine and interpret a 99% confidence interval for the parameter.

Exercise 16-19: Marriage Ages

8-17, 9-6, 16-19, 16-24, 17-22, 23-1, 23-12, 23-13, 26-4, 29-17, 29-18

Reconsider Activity 9-6, in which you analyzed the ages of couples who applied for marriage licenses. Now consider the issue of estimating the proportion of all marriages in this county for which the bride is younger than the groom. These 24 couples are actually a subsample from a larger sample of 100 couples who applied for marriage licenses in Cumberland County, Pennsylvania, in 1993. The bride was younger than the groom for 67 couples, and the groom was younger than the bride for 27 couples. Both people listed the same age on the marriage license for the remaining 6 couples, so disregard them and consider 94 as the new sample size.

a. Are the technical conditions for a one-proportion *z*-interval satisfied?
b. Determine a 99% confidence interval for the proportion of all marriages in this county for which the *bride* is younger than the groom.
c. Determine a 99% confidence interval for the proportion of all marriages in this county for which the *groom* is younger than the bride.
d. Comment on how the interval in part c compares to the interval in part b.
e. Based only on the 99% confidence interval, comment on whether the sample data suggest that the bride is younger than the groom for more than half of the marriages in this county.

Exercise 16-20: Critical Values

12-18, 16-2, **16-20,** 19-13

a. Determine the critical value z^* corresponding to
   i. 85% confidence.
   ii. 97.5% confidence.
   iii. 51.6% confidence.

b. Which of these three z^* values is largest? Which is smallest? Explain why this makes sense.

Exercise 16-21: Wrongful Conclusions

8-5, **16-21,** 17-8, 28-25

Suppose that Andrew and Becky both study a random sample that has a sample proportion of $\hat{p} = .4$. Each uses the sample data to produce a confidence interval for the population proportion, with Andrew obtaining the interval (.346, .474) and Becky obtaining the interval (.286, .514).

a. One of these intervals has to be incorrect. Identify which one, and explain why.

Suppose that Andrew and Becky decide to gather new random samples, with one using a sample size of 100 and the other using a sample size of 200. Andrew's confidence interval turns out to be (.558, .682), and Becky's is (.611, .779).

b. Report the sample proportion $\hat{p}$ obtained by each researcher. (Assume that both intervals were calculated correctly.)

c. Report the margin-of-error for each interval.

d. Explain why the information provided does not enable you to tell which sample size was used by which researcher.

e. Suppose that Andrew had the sample size of 100 and Becky 200. Determine the confidence level used by each. [*Hint:* First determine the critical value z^* used by each.]

Suppose that Andrew and Becky decide to study another issue with new samples, one with a sample size of 250 and one with a sample size of 1000. Both decide to form 90% confidence intervals from their samples. Andrew obtains the interval (.533, .635), and Becky gets (.550, .602).

f. Report the sample proportion and the margin-of-error for each interval.

g. Which sample size goes with which researcher? Explain.

h. Which researcher is more likely to obtain an interval that succeeds in capturing the population proportion? Explain.

Exercise 16-22: Candy Colors

1-13, 2-19, 13-1, 13-2, 13-15, 13-16, 13-20, 16-5, **16-22,** 24-15, 24-16

Assume for now that 15% of all Reese's Pieces are orange, that is, that the population proportion of orange candies is $\pi = .15$.

a. Open the `Simulating Confidence Intervals` applet and make sure it is set for **Proportions** with the **Wald** method. Set the population proportion value to $\pi =$ **.15**, the sample size to $n =$ **10**, the number of intervals to **200**, and the confidence level to **95%**. Click **Sample**, and keep clicking until you have generated 1000 intervals. What percentage of these intervals succeed in capturing the true value of the population proportion (which you have set to be $\pi = .15$)? Is this percentage close to 95%?

b. Explain why it is not surprising that the success rate was not very close to 95% in this case. [*Hint:* Check the technical conditions for the validity of the confidence interval procedure.]

Exercise 16-23: Penny Thoughts

2-1, 3-10, **16-23**

Recall from Exercise 3-10 that a Harris Poll in 2003 asked a national sample of 2316 adults whether they favored or opposed abolishing the penny, with 59% saying that they oppose it.

a. Clearly define the population parameter of interest.

b. Determine and interpret a 95% confidence interval for the population parameter of interest.

c. Comment on whether the technical conditions are satisfied.

Exercise 16-24: Marriage Ages

8-17, 9-6, 16-19, **16-24,** 17-22, 23-1, 23-12, 23-13, 26-4, 29-17, 29-18

Reconsider the Exercise 16-19. In the original sample, the bride was younger than the groom for 16 couples, and the groom was younger than the bride for 6 couples.

a. Based on this sample of 22 couples, show that the sample size condition required for a confidence interval is not satisfied.

An alternative confidence interval procedure for a population proportion π is to use:

$$\tilde{p} \pm z^* \sqrt{\frac{\tilde{p}(1-\tilde{p})}{n+4}}$$

where $\tilde{p}$ is the sample proportion obtained after adding four fictional observations to the sample, with two in each category. (Exercise 16-27 asks you to investigate the beneficial properties of this alternative procedure.)

b. Calculate the value of $\tilde{p}$ for this marriage study with 22 couples. [*Hint:* Use $22 + 4 = 26$ as the sample size, with $16 + 2 = 18$ couples where the bride was younger and $6 + 2 = 8$ couples where the groom was younger.]

c. Use the alternative confidence interval procedure to determine a 95% confidence interval for the population proportion of couples in this county for which the bride is younger than the groom.

d. How do the midpoint and length of this interval compare to those in the Exercise 16-19.

Exercise 16-25: Generation M

3-8, 4-14, 13-6, 16-1, 16-3, 16-4, 16-7, 16-25, 16-26, 21-11, 21-12

A rough approximation of the margin-of-error for a 95% confidence interval can be found with the expression $1/\sqrt{n}$, where n represents the sample size.

a. Apply this approximation to the Exercise 16-7, which involved a sample of 2032 American youths.

b. Comment on how close this approximate margin-of-error comes to the two margin-of-error that you reported in part c of the Exercise 16-7. In particular, for which statistic (.49 or .31) does this approximation come closer to the calculated margin-of-error?

Exercise 16-26: Generation M

3-8, 4-14, 13-6, 16-1, 16-3, 16-4, 16-7, 16-25, 16-26, 21-11, 21-12

Consider the previous exercise.

a. Apply the $1/\sqrt{n}$ approximation to a survey with a sample size of 2032 and then with a sample size of 508, which is one-fourth as large as the first sample size.

b. Calculate the *ratio* of these two (approximate) margin-of-error. What does this indicate about how much the sample size must increase in order to cut the margin-of-error in half?

Exercise 16-27: Alternative Procedure 16-24, 16-27

Recall from Exercise 16-24 that an alternative confidence interval procedure for a population proportion π is to use:

$$\tilde{p} \pm z^* \sqrt{\frac{\tilde{p}(1-\tilde{p})}{n+4}}$$

where $\tilde{p}$ is the sample proportion obtained after adding four fictional observations to the sample, with two in each category. (In this exercise you will investigate beneficial properties of this alternative procedure.)

a. Go to the `Simulating Confidence Intervals` applet that you used in Activity 16-5. (Keep the "method" set to Proportions/Wald.) Set the sample size to $n =$ **15** and the population proportion to $\pi =$ **.2**. Set the confidence level to **95%** and the number of intervals to **200**. Click on **Sample** and keep clicking on **Sample** until you have generated a total of 2000 intervals. Examine the running total. What percentage of these 2000 intervals succeed in capturing the actual value of the population proportion ($\pi = .2$)?

b. If the (usual) confidence interval procedure were working well, what would the percentage in part a be very close to? Explain.

c. Judging from your answers to parts a and b, is the (usual) confidence interval procedure working well? Explain.

d. Now use the second pull-down menu to change from **Wald** (which is the name of the usual procedure) to **adjusted Wald** (a name for this alternative procedure). Now click on **Sample** until you have produced a total of 2000 intervals. What percentage of *these* 2000 intervals succeed in capturing the actual value of the population proportion ($\pi = .2$)?

e. Does your answer to part d suggest that the alternative procedure is working better than the usual procedure in this case? Explain.

Exercise 16-28: Feeling Rushed?

6-1, 6-30, 6-31, **16-28,** 17-11, 24-26, 25-30, 25-31

Recall from Activity 6-1 that the 2004 General Social Survey asked a random sample of 977 adult Americans how often they feel rushed, with the result that 304 said "always."

a. Check whether the technical conditions for calculating a confidence interval for a population proportion are satisfied.
b. Determine the midpoint of a 95% confidence interval for the population proportion in 2004 who would have said that they always feel rushed.
c. Determine the margin-of-error for the confidence interval mentioned in part b.
d. Determine the endpoints of this confidence interval.
e. Write a sentence interpreting what this interval says.

Exercise 16-29: Celebrating Mothers

3-35, **16-29**

Recall from Exercise 3-35 that the National Retail Federation sponsored a survey of 7859 consumers on April 4–11, 2007, with 84.5% saying that they were planning to celebrate Mother's Day.

a. Describe in words the relevant parameter for this survey. Also indicate the appropriate symbol for this parameter.
b. Comment on whether the technical conditions for a confidence interval for a population proportion are satisfied.
c. Estimate this parameter from part a with a 99% confidence interval.

Exercise 16-30: Hand Washing

2-2, **16-30,** 21-16, 25-16

Recall from Activity 2-2 the study in which researchers observed people using public restrooms in various public venues around the United States. They found that 2393 of 3206 men washed their hands and that 2802 of 3130 women washed their hands.

a. For each gender, construct a 99% confidence interval for the population proportion who wash their hands in public restrooms.
b Do these confidence intervals overlap? What does this suggest?
c. Describe the populations you are willing to generalize these results to.

Exercise 16-31: CPR on Pets

5-33, **16-31,** 21-35

A national survey of 1166 pet owners on October 1–5, 2009 found that 63% of dog owners and 53% of cat owners said that they would perform CPR on their pet in the event of a medical emergency. The news release reports that the margin-of-error is ± .029 among *all* pet owners.

a. Based on this information, assuming the margin-of-error reported was for 95% confidence, calculate the endpoints of the 95% confidence interval. Interpret this interval, making sure that you clearly define the parameter.
b. Even if we assume that the sampling method was random, we do not have enough information to determine a 95% confidence interval for the population proportion of *dog* owners who say that they would perform CPR. Explain what additional information is needed.
c. Explain why the margin-of-error for the confidence interval would be larger than .029.

Exercise 16-32: Reading *Harry Potter*

3-32, **16-32**

The 2008 *Kids and Family Reading Report* describes a survey of 501 children aged 5–17 from 25 major cities around the United States. For now, regard this sample as representative of the population of all American children aged 5–17. The survey found that 58% of these kids have read at least one Harry Potter book.

a. Confirm that the sample size technical condition is satisfied here.
b. Determine and interpret a 90% confidence interval for the population proportion of interest.
c. Describe how a 95% confidence interval would compare to the 90% confidence interval. Comment on the midpoint as well as margin-of-error.
d. Comment on whether you believe this sample of children from 25 major cities is likely to be representative of all American children on this issue.

Exercise 16-33: Friday Classes

3-30, **16-33,** 16-34

Suppose that you want to estimate the proportion of Cal Poly students who have at least one class on Friday during the current term. You want to determine how many students need to be sampled in order to estimate this proportion to within ± .04 with 95% confidence. Lacking any other information, you start by guessing that this proportion might be close to .5.

a. Determine how many students need to be sampled to achieve the desired margin-of-error and confidence level.

If you cannot afford to sample that many students, you could make do with fewer students by changing the desired margin-of-error or the confidence level.

b. In what direction would you have to change the desired margin-of-error, if you want to stick with 95% confidence but use fewer students?

c. In what direction would you have to change the confidence level, if you want to stick with ± .04 margin-of-error but use fewer students?

Exercise 16-34: Friday Classes

3-30, 16-33, **16-34**

Reconsider the previous exercise. Now suppose you have reason to believe that about three-fourths of the population have class on Fridays.

a. Use this estimate to determine how many students need to be sampled in order to estimate the population proportion to within ± .04 with 95% confidence.

b. Comment on how the required number of students changed, based on the change in the estimate of the population proportion.

Exercise 16-35: Hanging Toilet Paper

Suppose that you plan to take a random sample of hanging toilet paper rolls in your town, with a goal of estimating the proportion of rolls that are hung so that the paper comes out over the top, as opposed to coming out underneath. You want to determine how many rolls to sample for a 90% confidence interval to have margin-of-error no larger than ± .06. As an initial guess for this proportion, you find an Internet poll result that 75% of rolls are hung so the paper comes out over the top. Determine the necessary sample size.

16

TOPIC 17

17

Tests of Significance: Proportions

How often does the winning team in a baseball game score more runs in one inning than the losing team scores in the entire game? Does this happen three-quarters of the time, as someone once claimed in a letter to the "Ask Marilyn" columnist? In this topic, you will learn to assess the evidence that sample data provide about such a claim. Another example that you will investigate concerns whether or not people who need to make up an answer to a question on the spot tend to select certain answers more than others.

Overview

In the previous topic, you took what you learned about the concept of statistical confidence in Unit 3 and studied a procedure known as confidence intervals. In this topic, you will augment what you learned about the concept of statistical significance by studying a procedure known as a test of significance. Such a procedure assesses the degree to which sample data provide evidence against a particular conjecture about the value of the population parameter of interest. Your understanding of the reasoning process behind these tests will deepen, and you will study the formal structure of a test of significance, while learning a specific procedure for conducting a test concerning a population proportion.

Preliminaries

1. In what percentage of major-league baseball games would you predict that the winning team scores more runs in one inning than the losing team scores in the entire game? (Activity 17-5)

2. Consider a cola taste test in which a subject is presented with three cups, two that contain the same brand of cola and one that contains a different brand. The subject is to identify which one of three cups contains the different brand. In a test consisting of 60 trials, how convinced would you be that the subject did better than simply guessing if he or she was correct 21 times? (Exercise 17-24)

3. How many would a person have to get right out of 60 trials for you to be reasonably convinced that he or she does better than simply guessing at which one is the different brand? (Exercise 17-24)

4. If a subject identifies the different brand correctly in 40% of a sample of trials, would you be more impressed if it were a sample of 200 trials or a sample of 20 trials? (Exercise 17-24)

Collect Data

1. A campus legend tells the story of two friends who lied to their professor by blaming a flat tire for their having missed an exam. The professor sent them to separate rooms to take a make-up exam. The first question (worth 5 points) was easy, but the second question, worth 95 points, asked, "Which tire was it?" If you were caught in a situation where you suddenly had to choose a hypothetically flat tire, which of the four tires would you say went flat (left front, right front, left rear, or right rear)? (Activity 17-3)

2. Record the response count for yourself and your classmates:

right front approx .41
.25 is average

Left front: 5 Right front: 6

Left rear: 0 Right rear: 4

In-Class Activities

Activity 17-1: Kissing Couples 15-1, 16-6, 17-1, 17-2, 17-12, 18-1, 24-4, 24-14

Recall the kissing study from Topic 15. In Activity 15-1, you used the Central Limit Theorem to determine the probability of obtaining a sample proportion as extreme as .645 (the actual sample proportion obtained in the research study), assuming the population proportion was .5. You found that such a sample result (of 124 couples) would be highly surprising from this population, so you had very strong evidence against the original supposition that the population proportion of kissing couples that lean right is .5. We are now going to cast the calculations and decisions involved in this reasoning process into a more formal, step-by-step process called a **test of significance**. We will outline the six main steps in such a test.

Step one: Give a description of the parameter, being sure to identify the type of number (e.g., a mean or a proportion), the variable, and the population.

a. Write out a definition of the parameter of interest in the kissing study and indicate what symbol you will use to represent it.

let pi be the population f proportion of kissing couples who woudl turn right while kissing

17

Step two: State competing claims about the parameter of interest.
The **null hypothesis** states the parameter of interest is equal to a specific value:

H_0: parameter = hypothesized value

This is typically a statement of no effect or no difference.

b. In this kissing study, the null hypothesis is that there is no tendency to lean to the right more than to the left in the population of all kissing couples. Restate this using the symbol from part a and the appropriate hypothesized value, instead of words:

H_0: .5

The **alternative hypothesis** states what the researchers suspect or hope to be true about the parameter. It will take *one* of these three forms:

H_a: parameter < hypothesized value
H_a: parameter > hypothesized value
H_a: parameter ≠ hypothesized value

The specific form (i.e., the direction) of the alternative is determined by the research question, before the sample data are examined.

c. The researcher suspected more than half of all kissing couples lean right. Restate this conjecture (again with symbols and a number) as an alternative hypothesis.

H_a: pi > .5

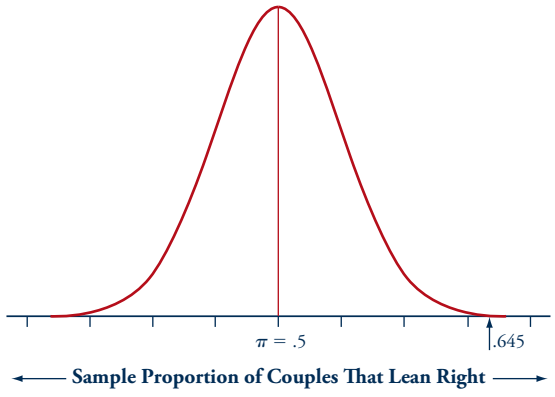
Note that the symbol and the hypothesized value don't change; you are just selecting between less than, greater than, or not equal to based on the research conjecture.

Step three: Specify the behavior of the sampling distribution under the null hypothesis. This typically involves checking some **technical conditions** that need to be met before applying, for example, the Central Limit Theorem. Typically there are two conditions to check: one involving randomness in the study design and one involving the sample size/normality of the sampling distribution.

d. What conditions needed to be met for the Central Limit Theorem of a Sample Proportion to be satisfied? How did you check them in Activity 15-1?

rni
random - not random, but most likely representative of all kissing couples(?)
normal - 124*whatever >10 so yes
independent - 10 * 124 probably less than pop of all kissing couples

Note that you will often check these conditions assuming the null hypothesis is true. At this point, it is very good practice to then draw a well-labeled sketch of the sampling distribution of the statistic.



Step four: Calculate a **test statistic**.
This is a measure of the discrepancy between our observed statistic and the hypothesized value of the parameter. If the discrepancy is large, we have evidence against the null hypothesis.

In Activity 15-1, you calculated the z-score of the observed sample proportion ($\hat{p} = .645$) under the assumption that the null hypothesis was true ($\pi = .5$), as follows:

$$z = \frac{.645 - .5}{\sqrt{.5(1 - .5) / 124}} = 3.23$$

which says that the observed sample proportion who lean to the right (.645) is more than 3 standard deviations above the hypothesized value of 0.5.

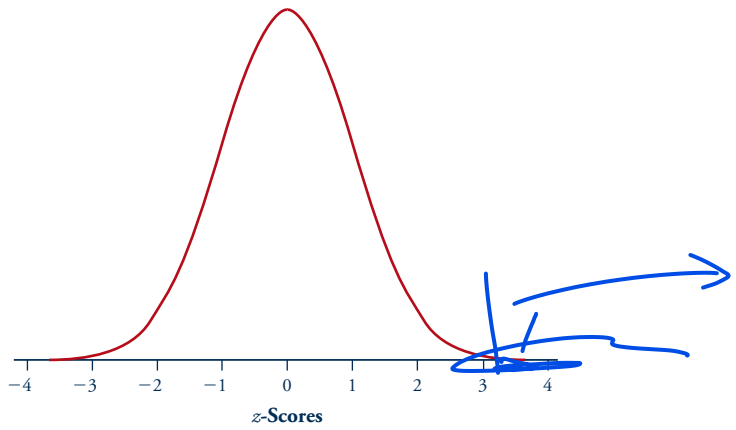
e. Rewrite the above z-score calculation as a general formula using $\hat{p}$, π_0, and n, where π_0 is the (null) hypothesized value of the population proportion.

z = phat - pi0 / sqrt(pi0 (1-pi0)/n)

Step five: Calculate the ***p*-value**, which is the probability, assuming the null hypothesis to be true, of obtaining a test statistic at least as extreme as the one actually observed. *Extreme* means "in the direction of the alternative hypothesis."

f. Below is a sketch of the standard normal distribution. Indicate where the test statistic ($z = 3.23$) falls on this graph. Because values above 3.23 would be even farther from the hypothesized value, shade to the right. The area you are shading is the p-value in this study.

17



Step six: Summarize your conclusion in context. First, either state a **test decision** or a comment evaluating the strength of evidence against the null hypothesis. Where a test decision needs to be made:

If the *p*-value is small, reject the null hypothesis.

If the *p*-value is not small, fail to reject the null hypothesis.

Second, respond to the research question, stating that you either have evidence for the alternative hypothesis (in context) or you do not. In other words, restate your final conclusions in the language of the research question.

Note that the smaller the *p*-value, the stronger the evidence against the null hypothesis and in favor of the alternative hypothesis. Typical evaluations are

- A *p*-value above .10 constitutes little or no evidence against the null hypothesis.
- A *p*-value below .10 but above .05 constitutes moderately strong evidence against the null hypothesis.
- A *p*-value below .05 but above .01 constitutes reasonably strong evidence against the null hypothesis.
- A *p*-value below .01 constitutes very strong evidence against the null hypothesis.

In some studies the researcher decides in advance how small the *p*-value must be to provide convincing evidence against the null hypothesis. This cutoff value is called a **significance level**, denoted by α (alpha). Common values are $\alpha = .10$, $\alpha = .05$, $\alpha = .01$. A smaller significance level indicates a stricter standard for deciding if the null hypothesis can be rejected. If the researcher specifies a level of significance in advance, you would then say you reject or fail to reject at that particular level. Another common expression is to say that the data are **statistically significant** if it is unlikely to have occurred by chance or sampling variability alone (assuming that the null hypothesis is true).

g. Does the *p*-value for the kissing study lead to rejecting or failing to reject the null hypothesis at the .05 level?

yes

h. Does this study provide convincing evidence that kissing couples turn to the right a majority of the time? Explain the reasoning process that leads to this conclusion.

yes

[*Hint*: To explain the reasoning process, you may want to refer back to the simulations conducted in Topics 13 and 15.]

Given [h_0 is true] there is no tendency for kissing couples to lean right while kissing, the chance/probability of observing a sample proportion at least .645 by chance alone is .06%. Because we actually foudn a phat of .645, we have strong evidence that the population parameter is greater than 0.5.

Watch Out

- It may take some time for you to fully understand the reasoning process of a test of significance. Some people find it to be a bit convoluted to start by assuming the null hypothesis to be true and ask how unlikely the observed sample data would be, given that hypothesis. If the answer is that the observed sample data would be *very unlikely* if the null hypothesis were true, then the sample data provide strong evidence against that null hypothesis.
- The *p*-value is *not* the probability that the null hypothesis is true. Rather, it is the probability of obtaining such an extreme sample result (or one even more extreme) if the null hypothesis is true.
- As always, be sure to relate your conclusions to the context. For example, do not think that "reject H_0" is a complete conclusion. In fact, it's even incomplete to say "reject H_0 and conclude that $\pi > .5$." You must instead express this conclusion in context by saying, "The sample data provide very strong evidence that more than half of all kissing couples lean to the right." ^with p value 0.0006
- Always include all steps of a significance test:

1. Identify and define the parameter.
2. State the null and alternative hypotheses, preferably in words as well as symbols.
3. Check the technical conditions.
4. Calculate the test statistic.
5. Report the *p*-value.
6. Summarize your conclusion in context, including a test decision if a significance level α is provided.

Activity 17-2: Kissing Couples

15-1, 16-6, 17-1, **17-2,** 17-12, 18-1, 24-4, 24-14

Reconsider the study concerning the direction that kissing couples lean their heads. The researchers noted that other right-sided tendencies appear in approximately two-thirds of the population. So they wanted to test this value for the proportion of all kissing couples who turn right. However, this time they did not have a prior suspicion of whether the population proportion would be greater or smaller than that number.

a. State the null hypothesis represented by this statement.

H_o: para = 2/3

b. State the alternative hypothesis to represent that the population proportion is not two-thirds.

H_a: para =/= 2/3

"two tailed alternative"

17

c. Using this new hypothesized value for π, is the sample size technical condition still met?

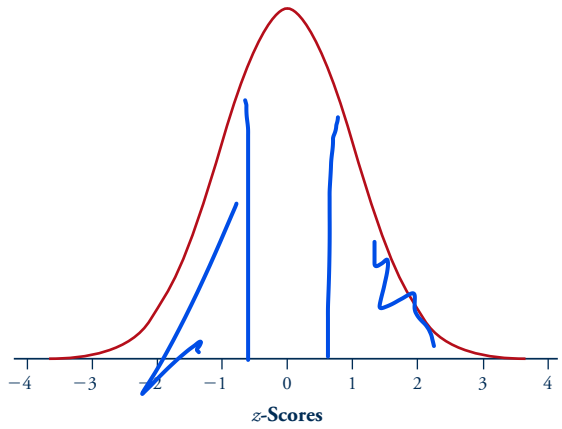
yea

d. Compute the test statistic for this hypothesized value.

-0.5118

To compute a *p*-value for a "two-sided" (not equal to) alternative, we look for results at least as extreme (farther from the hypothesized value) as the sample result *in both directions*.

e. For the following standard normal curve, shade the area to the left of the observed test statistic (in this problem negative), and the area to the right of the positive *z*-score.



f. How do these two areas compare?

same

When the alternative hypothesis is two-sided (not equal to), we find the *p*-value by computing $2 \times \Pr(Z \geq |z|)$.

g. Use this definition and Table II to determine the two-sided *p*-value for this test.

p=.61

h. What is your test decision at the .05 level? Do you reject or fail to reject the null hypothesis?

do not reject

i. Do these data provide convincing evidence that π differs from two-thirds? Explain your reasoning, including an interpretation of what this *p*-value measures (what is it the probability of?).

no

no, assumin the pop prop of kis couples =2/3, the probability of getting a |z|>.51 is .61

or phat that differs from at least |.645-.666...|=2.2% from 2/3, is .61

j. Determine the test statistic and *p*-value for testing whether the population proportion of all kissing couples differs from .60. State your test decision at the .05 level, and summarize your conclusion.

h0: pi=.6, ha: pi=/=.6

fail to reject sinc ep valu is larger than .05 level

Watch Out

Be careful never to *accept* a null hypothesis. Even when the *p*-value is not terribly small, the appropriate conclusion is that the sample data do not provide *enough* evidence to reject the null hypothesis in favor of the alternative. In other words, a test of significance assesses only the evidence *against* the null, not the evidence in favor of it. As this activity illustrates (as well as Activity 16-6), there are numerous plausible values for the population proportion, so we haven't *proven* π equals any one of these particular values.

Activity 17-3: Flat Tires 17-3, 17-4, 17-10, 17-16, 18-18, 24-17

Consider the flat-tire question in the Collect Data section. This campus legend has been reported in the news several times, including in an April 11, 2008 issue of the *Chronicle of Higher Education.* Use the six-step process to carry out a test of significance of whether your class data provide convincing evidence that students at your school tend to pick the front-right tire more than you would expect by random chance.

a. Define the parameter of interest in words.

the populatio parameter that all nash studets pick right frot tire.

b. State the null and alternative hypotheses about this parameter. [*Hint*: What is the value of the parameter if students are just guessing randomly among the four tires?]

H0 is pi=.25, ha is pi>.25

c. Based on the check of technical conditions (assuming the null hypothesis is true), what is the smallest sample size *n* for which the one-proportion *z*-test is valid?

n=40 z=phat-pi/sqrt(pi(1-pi)/n) z=3.5355, p=2e-4

d. Does the sample size for your class satisfy this condition? If so, then proceed to use your class sample data for the following questions. If not, then combine your

class data with results from some of the authors' recent classes, which found 58 of 145 students choosing the front-right tire.

e. Calculate the test statistic and *p*-value. Include a sketch of the sampling distribution of the test statistic with the *p*-value area shaded. Provide one-sentence interpretations of each of these calculated values.

o value 2*10^-4 = .0002
our result is 3.4 SD above ormal

Assumig there is no tedecy to pick frot right tire, the p of gettig at least 39/96 who chose RF is 2/10000

Which tire goes flat?

f. Confirm these calculations using technology, which could be the `Test of Significance` Calculator applet. Make sure the procedure is set to **One Proportion**. Enter the value of π_0 (the hypothesized value of the parameter) for the null hypothesis, and click on the inequality symbol (>) to specify the correct form for the alternative hypothesis. Then enter the sample size ***n*** and sample number of successes (**Count**), and press **Calculate**. The applet will automatically calculate the value of the sample proportion $\hat{p}$ (alternatively, you could have entered $\hat{p}$ and the applet would determine the count).

g. Evaluate the magnitude of the *p*-value you have determined, and summarize your conclusion in context.

the probablility sigificace depeds on N

reject sice p<.05

Activity 17-4: Flat Tires 17-3, 17-4, 17-10, 17-16, 18-18, 24-17

Reconsider the flat-tires situation from Activity 17-3. Now suppose that 30% of a random sample select the front-right tire.

a. Consider the question of whether or not this sample result constitutes strong evidence that the front-right tire would be chosen more than one-quarter of the time in the long run. Do you need more information to answer this question? Explain.

b. Suppose that this sample result (30% answering front-right) had come from a sample of $n = 50$ people. Use technology to conduct the appropriate test of significance. Record in the first row of the following table the value of the test statistic and the test's *p*-value. Also indicate (yes or no) whether or not the test is significant at each of the listed significance levels.

c. Repeat part b for the other sample sizes listed in the table.

Sample Size	# "Right Front"	$\hat{p}$	*z*-Statistic	*p*-value	$\alpha = .10$?	$\alpha = .05$?	$\alpha = .01$?	$\alpha = .001$?
50	15	.30	.816	.207				
100	30	.30	1.1547	.1241				
150	45	.30	1.4142	.0786				
250	75	.30	1.8257	.0339				
500	150	.30	2.572	.0049				
1000	300	.30	3.6515	1e-4				

d. Write a few sentences summarizing what your analysis reveals about whether or not a sample result of 30% is significantly greater than a hypothesized value of 25%.

Let pi=.71 let pi be th epopulatio proportio all americans over 25 in researcher state who are obese
ha is pi=/=.71
We dot have eough evidece with a p value of .031 given a statistical sigificance threshold of 1%, to say that the populatio proportion of all americans over 25 i the researcher's state differs from 0.71.

Watch Out

You cannot conduct a significance test without knowing the sample size involved. Sample size plays a key role in tests of significance. The statistical significance of a sample result depends largely on the sample size involved. With large sample sizes, even small differences (such as 30% in a sample compared to a hypothesized population percentage of 25%) can be statistically significant, because they are unlikely to occur by chance.

Self-Check
Activity 17-5: Baseball "Big Bang" 17-5, 17-17

A reader wrote in to the "Ask Marilyn" column in *Parade* magazine to say that his grandfather told him that in three-quarters of all baseball games, the winning team scores more runs in one inning than the losing team scores in the entire game. (This phenomenon is known as a "big bang.") Marilyn responded that this proportion seemed too high to be believable. Let π be the proportion of all major-league baseball games in which a big bang occurs.

a. Restate the grandfather's assertion as the null hypothesis, in symbols and in words.

b. Given Marilyn's conjecture, state the alternative hypothesis, in symbols and in words.

To investigate this claim, we randomly selected one week of the 2006 major-league baseball season, which turned out to be July 31–August 6, 2006. Then we examined the 95 games played that week to determine which had a big bang and which did not.

c. Sketch and label the sampling distribution for the sample proportion of games containing a big bang, according to the Central Limit Theorem, assuming that the grandfather's null hypothesis is true. Also check whether or not the conditions hold for the CLT to apply.

Of the 95 games in our sample, 47 contained a big bang.

d. Calculate the sample proportion of games in which a big bang occurred. Use an appropriate symbol to denote it.

e. Is this sample proportion less than three-fourths and therefore consistent with Marilyn's (alternative) hypothesis? Shade the area under your sampling distribution curve corresponding to this sample result in the direction conjectured by Marilyn.

f. Calculate the test statistic and use the Standard Normal Probabilities Table (Table II) or technology to find its *p*-value.

$z =$

p-value $=$

g. Based on this *p*-value, would you say that the sample data provide strong evidence to support Marilyn's contention that the proportion cited by the reader's grandfather is too high to be the actual value? Explain. Also indicate what test decision you would reach at the $\alpha = .01$ level.

In her response, Marilyn went on to conjecture that the actual proportion of big bang games is one-half.

h. Using a two-sided alternative, state the null and alternative hypotheses (in symbols and in words) for testing Marilyn's claim.

i. Use technology to determine the test statistic and *p*-value for this test.

j. What conclusion would you draw concerning Marilyn's conjecture?

k. Use the sample data to produce a 95% confidence interval to estimate the proportion of all major-league baseball games that contain a big bang. Interpret this interval, and comment on what it reveals about the grandfather's claim and Marilyn's response.

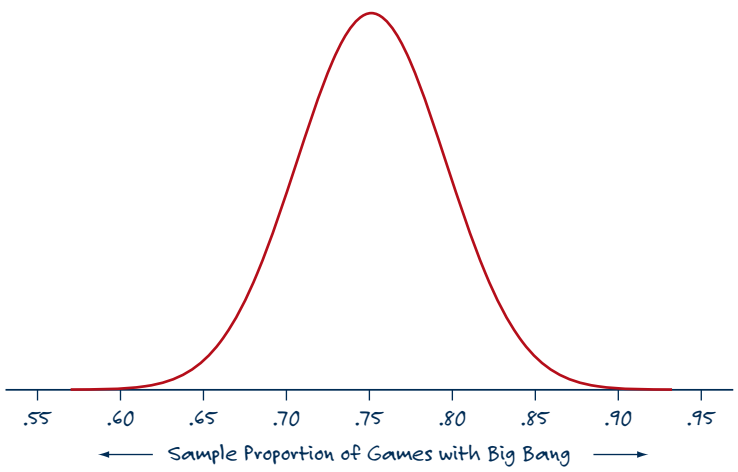
Solution

a. The null hypothesis is that the proportion of all major-league baseball games that contain a big bang is three-fourths. In symbols, the null hypothesis is $H_0: \pi = .75$.

b. The alternative hypothesis is that less than three-fourths of all major-league baseball games contain a big bang. In symbols, the alternative hypothesis is $H_a: \pi < .75$.

c. The CLT applies here because $95(.75) = 71.25$ is greater than 10 and $95(.25) = 23.75$ is also greater than 10. According to the CLT, the sample proportion would vary approximately normally with mean .75 and standard deviation

$$\sqrt{\frac{(.75)(.25)}{95}} \approx .0444$$



17

d. The sample proportion of games in which a big bang occurred is

$$\hat{p} = \frac{47}{95} \approx .495$$

e. Yes, this sample proportion is less than .75, as Marilyn conjectured.
f. The test statistic is

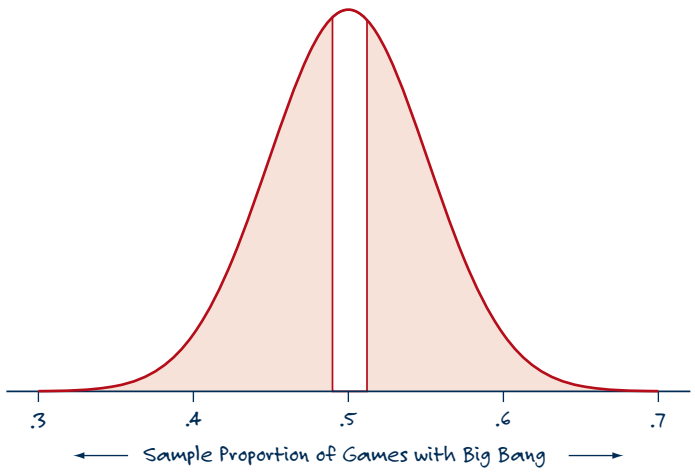
$$z = \frac{.495 - .75}{\sqrt{\frac{(.75)(.25)}{95}}} \approx \frac{.495 - .75}{.0444} \approx -5.74$$

This statistic says that the observed sample result is almost six standard deviations below what the grandfather conjectured. This z-score is way off the chart in Table II, indicating that the p-value is virtually zero.

g. Yes, this very small p-value indicates that the sample data provide extremely strong evidence against the grandfather's claim. There is extremely strong evidence that less than 75% of all major-league baseball games contain a big bang. The null (grandfather's) hypothesis would be rejected at the $\alpha = .01$ level.
h. The hypotheses for testing Marilyn's claim are H_0: $\pi = .5$ vs. H_a: $\pi \neq .5$.
i. The test statistic is

$$z = \frac{.495 - .5}{\sqrt{\frac{(.5)(.5)}{95}}} \approx \frac{.495 - .5}{.0513} \approx -0.10$$

The p-value is $2(.4602) = .9204$.



j. This p-value is not small at all, suggesting that the sample data are quite consistent with Marilyn's hypothesis that half of all games contain a big bang. The sample data provide no reason to doubt Marilyn's hypothesis.
k. A 95% confidence interval for π (the population proportion of games that contain a big bang) is given by

$$\hat{p} \pm z^* \sqrt{\frac{\hat{p}(1 - \hat{p})}{n}}$$

which is

$$.495 \pm 1.96\sqrt{\frac{(.495)(.505)}{95}}$$

which is .495 ± .101, which is the interval from .394 to .596. Therefore, you are 95% confident that between 39.4% and 59.6% of all major-league baseball games contain a big bang. The grandfather's claim (75%) is not within this interval or even close to it, which explains why it was so soundly rejected. Marilyn's conjecture (50%) is well within this interval of plausible values, which is consistent with the value .50 not being rejected.

Watch Out

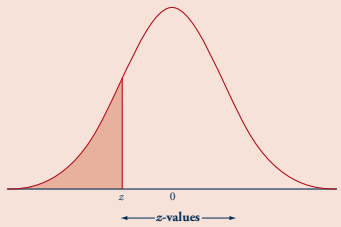
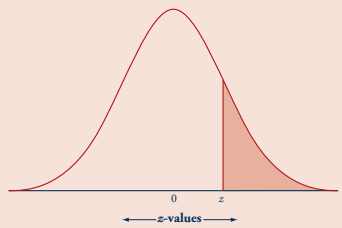
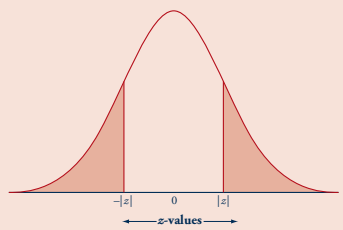
- Remember that the hypotheses are always statements about a parameter, not about a statistic. The whole point is to see what you can infer about an unknown parameter value based on a sample statistic.
- Remember to use the hypothesized value of the parameter (denoted by π_0) of the test statistic calculation under the square root in the denominator and also in checking the technical conditions. It is easy to mistakenly use the sample proportion $\hat{p}$ in those calculations.
- Try to carry many decimal places of accuracy in intermediate calculations. If you round too much in an early calculation, that error can get magnified in subsequent calculations.
- Again remember that you do not *accept* a null hypothesis, even one with a *p*-value as great as Marilyn's. The sample data are in very close agreement with Marilyn's hypothesis, but you still should not conclude that exactly 50% of all games contain a big bang.
- Even with an extremely small *p*-value, stop short of saying that the data *prove* that the null hypothesis is false. Even though you have overwhelming evidence against the grandfather's claim, you have not technically *proven* that his claim is wrong.

Wrap-Up

This topic has introduced you to the formal structure of a **test of significance**. The reasoning process, as you saw in Topic 15, asks how often the observed sample result or one even more extreme would occur purely by chance, given the hypothesized value of the population parameter. If such a sample result turns out to be unlikely, then the sample data provide evidence against the hypothesized parameter value.

The basic structure of a test of significance follows six steps. The following table applies these steps to the setting of examining one sample on a binary response variable (a *z*-test for a population proportion). In later topics, you will see that the structure and reasoning are the same for many other situations.

1. Identify and define the population parameter of interest.	π
2. State the **null** and **alternative hypotheses** based on the study question.	$H_0: \pi = \pi_0$ $H_a: \pi < \pi_0$ or $H_a: \pi > \pi_0$ or $H_a: \pi \neq \pi_0$
3. Check whether or not the **technical conditions** required for the procedure to be valid are satisfied.	• Simple random sample from population of interest • $n\pi_0 \geq 10$ and $n(1 - \pi_0) \geq 10$

4. Calculate the **test statistic**, which standardizes the distance between the observed sample statistic and the hypothesized value of the parameter.	$z = \dfrac{\hat{p} - \pi_0}{\sqrt{\pi_0(1-\pi_0)/n}}$
5. Calculate the ***p*-value**, which is the probability of obtaining such an extreme test statistic value when the null hypothesis is true. It is often very helpful to include a sketch of the sampling distribution, shading in the appropriate *p*-value area.	*p*-value = $\Pr(Z \le z)$  or $\Pr(Z \ge z)$  or $2 \times P(Z \ge \lvert z \rvert)$  depending on the form of the alternative hypothesis

You also learned that the smaller the *p*-value, the stronger the evidence against the null hypothesis. In other words, a small *p*-value indicates that the observed result would be surprising if the null hypothesis were true, and so provides evidence that the null hypothesis is not true. For example, the sample baseball data provided very strong evidence against the grandfather's big bang claim (that $\pi = .75$) because the tiny *p*-value revealed that the sample data observed ($\hat{p} = .495$) would have been extremely unlikely to occur if the grandfather's claim had been true. Instead of thinking that we happened to observe an incredibly unlucky sample, we will conclude that the data provide overwhelming evidence against the grandfather's claim. On the other hand, the *p*-value associated with Marilyn's claim (that $\pi = .5$) was not small, and so the sample data were not inconsistent with her claim (are consistent with sampling variability), providing no reason to doubt her hypothesis.

You also discovered that the sample size in a study plays a large role in calculating the *p*-value and, therefore, in determining whether or not the sample result is statistically significant (i.e., unlikely to occur due to sampling variability alone). For example, simply learning that 30% of a sample chooses the right-front tire does not enable you to say whether or not this is (statistically) significantly greater than 25%. If 30% of a sample of 10 people answer front-right, then this result is not at all significant. But if the sample contains 1000 people, then 30% answering front-right is significantly greater than 25%, because such an extreme result would almost never happen by random chance alone.

In Brief

Some useful definitions to remember and habits to develop from this topic are

- State the null and alternative hypotheses based on the research question *before* examining the sample data.
- The alternative hypothesis can have one of three different forms, corresponding to $<$, $>$, and $\neq$. The first two of these are **one-sided** alternatives and the last is a **two-sided** alternative.
- Always check the technical conditions before applying the test procedure.
- The test statistic provides a measure of how far the observed sample result is from the hypothesized value. With this test for a population proportion, the test statistic is the *z*-score for the sample proportion under the assumption that the null hypothesis is true.
- The *p*-value reports the probability of obtaining such an extreme sample result or more extreme by chance alone when the null hypothesis is true.
- The direction of the alternative hypothesis indicates how to calculate the *p*-value from the standard normal table.
- The smaller the *p*-value, the stronger the evidence against the null hypothesis.
- When the *p*-value is less than (or equal to) a prespecified **significance level,** the **test decision** is to reject the null hypothesis. Otherwise, the test decision is to fail to reject the null hypothesis.
- The greater the sample size, the smaller the *p*-value (if all else remains the same), and, therefore, the stronger the evidence against the null hypothesis (as long as the observed sample result is in the direction of the alternative hypothesis).
- As always, remember to relate your conclusions to the study's context and research question!

You should be able to

- State appropriate null and alternative hypotheses for assessing the plausibility of a claim about a population parameter. (Activities 17-1, 17-2, 17-3)
- Check conditions for whether a one-proportion *z*-test is valid to apply. (Activities 17-1, 17-3)
- Calculate a test statistic and *p*-value for a test of significance concerning a population proportion. (Activities 17-1, 17-2, 17-3)
- Make a test decision and summarize a conclusion from a test significance concerning a population proportion. (Activities 17-1, 17-2, 17-3, 17-4)
- Explain the reasoning process of a test of significance as it applies to a particular research question and set of data. (Activities 17-1, 17-2, 17-4)
- Explain the impact of sample size on all components of a test of significance. (Activity 17-4)

Confidence intervals and significance tests are the most widely used techniques in statistical inference, so you will continue to study them in the next topic. You will discover how these two techniques are related to each other, and you will learn to avoid some common misinterpretations.

Exercises

17

Exercise 17-6: Properties of *p*-values

17-6, 17-7

Suppose you conduct a significance test and decide to reject the null hypothesis at the $\alpha = .05$ level.

a. If you were to use the $\alpha = .10$ level instead, would you reject the null hypothesis, fail to reject it, or do you not have enough information to know? Explain. [*Hint:* What must be true about the *p*-value, knowing that you reject the null hypothesis at the $\alpha = .05$ level?]
b. If you were to use the $\alpha = .01$ level instead, would you reject the null hypothesis, fail to reject it, or do you not have enough information to know? Explain.

Now suppose you conduct a different significance test on a different set of data, and your test decision is to fail to reject the null hypothesis at the $\alpha = .05$ level.

c. If you were to use the $\alpha = .03$ level instead, would you reject the null hypothesis, fail to reject it, or do you not have enough information to know? Explain.
d. If you were to use the $\alpha = .07$ level instead, would you reject the null hypothesis, fail to reject it, or do you not have enough information to know? Explain.

Exercise 17-7: Properties of *p*-values

17-6, **17-7**

a. Is it possible for a *p*-value to be greater than .5? If so, explain the circumstances under which this could happen. If not, explain why not.
b. Is it possible for a *p*-value to be greater than 1? If so, explain the circumstances under which this could happen. If not, explain why not.

Exercise 17-8: Wrongful Conclusions

8-5, 16-21, **17-8,** 28-25

Describe what is incorrect about each of the following hypotheses.

a. $H_0: \hat{p} = .5$
b. $H_0: \pi = 1.2$
c. $H_0: \pi = .5, H_a: \pi \geq .5$
d. $H_0: \pi = .5, H_a: \pi \neq .6$
e. $H_0: \pi \neq .5, H_a: \pi = .5$

Exercise 17-9: Penny Activities

16-9, 16-10, 16-11, **17-9,** 18-14, 18-15

Suppose we claim to have a special penny that lands heads when flipped more than half the time. Suppose that we also tell you that we tossed this penny multiple times and obtained 75% heads. Would you be reasonably convinced that this was not a 50-50 process? If so, explain why. If not, describe what additional information you would ask for and explain why that information is necessary.

Exercise 17-10: Flat Tires

17-3, 17-4, **17-10,** 17-16, 18-18, 24-17

Conduct a significance test of whether or not *your* class data (collected in the Collect Data section) provides strong evidence that more than one-fourth of the students at your school would select the right-front tire. Report all components of the significance test. Summarize your conclusion, and explain the reasoning process that follows from your test.

Exercise 17-11: Feeling Rushed?

6-1, 6-30, 6-31, 16-28, **17-11,** 24-26, 25-30, 25-31

Recall from Activity 6-1 that the 2004 General Social Survey asked a random sample of 977 adult Americans how often they feel rushed, with three options to choose from: always, sometimes, and never.

a. State the appropriate null and alternative hypotheses for testing whether one-third of all adult Americans would answer "always."
b. Check and comment on whether the technical conditions for conducting this test are satisfied. The sample result was that 304 of the 977 people responded "always."
c. Calculate the value of the test statistic.
d. Calculate the *p*-value.
e. Interpret what this *p*-value means in this context. [*Hint*: The *p*-value is the probability of what, assuming what?]
f. State your test decision at the $\alpha = .05$ significance level.
g. Summarize your conclusion in the context of this study.

Exercise 17-12: Kissing Couples

15-1, 16-6, 17-1, 17-2, **17-12,** 18-1, 24-4, 24-14

Recall from Activities 15-1 and 16-6 the study in which a sample of 124 kissing couples were observed, with 80 of them leaning their heads to the right.

a. Use these sample data to test whether or not more than half of the population of kissing couples lean their heads to the right. Report the hypotheses, test statistic, and *p*-value. Summarize your conclusion at the $\alpha = .01$ significance level, and explain how your conclusion follows from your test.
b. Repeat part a, testing whether or not the population proportion of kissing couples who lean to the right is less than three-fourths.
c. Repeat part a, testing whether or not the population proportion of kissing couples who lean to the right differs from two-thirds.

Exercise 17-13: Political Viewpoints

17-13, 24-10

The 2004 General Social Survey asked a random sample of 1309 American adults to report their political viewpoint as liberal, moderate, or conservative. The number who classified themselves as moderate was 497.

a. Can you determine whether or not more than one-third of the sample consider themselves to be political moderates? If so, answer the question. If not, explain why not.
b. Suppose that one-third of the population consider themselves to be political moderates. What then would be the probability that 497 or more in a random sample of 1309 people would call themselves political moderates? [*Hint:* Answer this question by determining the *p*-value of the relevant test.]
c. How would you expect the *p*-value to change if the sample had contained 327 people, of whom 124 called themselves political moderates? Explain. [*Hint:* Note that the sample size and number of moderates have been reduced by a factor of four.]
d. Determine the *p*-value based on the sample data in part c. Was your answer to part c correct?

Exercise 17-14: Calling Heads or Tails

15-10, **17-14,** 17-15, 24-19

Refer to the data collected in Topic 15 and analyzed in Exercise 15-10 about whether you would call heads or tails if asked to predict the result of a coin flip.

a. What proportion of the responses were heads?
b. Is this proportion a parameter or a statistic? Explain.
c. Write a sentence identifying the parameter (and population) of interest in this situation.
d. Specify the null and alternative hypotheses, in words and in symbols, for testing whether or not the sample result differs significantly from .5.
e. Sketch and label the sampling distribution for the sample proportion specified by the null hypothesis. Shade the region corresponding to the *p*-value. Also provide a check of the technical conditions.
f. Calculate the test statistic and *p*-value for this test.
g. Is the sample result statistically significantly different from .5 at the .10 level? At the .05 level? At the .01 level?
h. Write a few sentences summarizing and explaining your conclusion.
i. Describe specifically what would have changed in this analysis if you had worked with the proportion of tails responses rather than heads responses.

Exercise 17-15: Calling Heads or Tails

15-10, 17-14, **17-15,** 24-19

Refer to the previous activity, where you examined data on the proportion of students who would respond heads if asked to predict a coin flip. In his book *Statistics You Can't Trust,* Steve Campbell claims that people call heads 70% of the time when asked to predict the result of a coin flip. Conduct a test of whether or not your sample data provide evidence against Campbell's hypothesis. Report the hypotheses, sketch the sampling distribution specified by the null hypothesis, check the technical conditions, and calculate the test statistic and *p*-value. Write a few sentences describing your conclusion.

17

Exercise 17-16: Flat Tires

17-3, 17-4, 17-10, 17-16, 18-18, 24-17

Reconsider the hypothetical results presented in Activity 17-4 for the flat-tires question. Determine the smallest sample size n for which a sample result of 30% answering front-right would be significant at the .10 level. [*Hint:* You may either use trial and error with technology or work analytically with the normal table and the formula for the test statistic.]

Exercise 17-17: Baseball "Big Bang"

17-5, 17-17

Consider again the big bang phenomenon described in Activity 17-5. Statistician Hal Stern examined all 968 baseball games played in the National League in 1986 and found that 419 of them contained a big bang.

a. Perform the appropriate test to see whether or not this sample proportion differs significantly from .5 at the $\alpha = .02$ level. Report your hypotheses in symbols and in words, your well-labeled sketch of the sampling distribution under the null hypothesis, your check of the technical conditions, the test statistic, and the p-value, in addition to stating and explaining your conclusion.

b. If you redefine big bang to mean that the winning team scores *at least* as many (instead of more) runs in one inning as the losing team scores in the entire game, then 651 of those 968 games contained a big bang. Does this sample proportion differ significantly from the "Ask Marilyn" reader's grandfather's assertion of .75 at the $\alpha = .08$ level? Again report the details of your analysis.

Exercise 17-18: Racquet Spinning

11-9, 13-11, 15-11, 17-18, 17-28, 18-3, 18-12, 18-13, 18-26, 18-27

Refer to Exercise 15-11, where you performed a test of whether or not a spun tennis racquet would *not* land up 50% of the time in the long run.

a. Report the p-value, and indicate whether or not you would reject the null hypothesis at the .05 level.

b. Does the test result indicate that the racquet would definitely land up 50% of the time in the long run? Explain.

c. What is the smallest significance level at which you would reject the null hypothesis? [*Hint:* Do not confine your consideration to common α levels.]

d. Explain precisely how your analysis would change depending on whether or not you work with the proportion landing up or the proportion landing down.

e. Use the sample data to find a 95% confidence interval for the long-run proportion of times that the racquet would land up.

f. Does this interval include the value .5?

g. Explain the consistency between your answers to parts a and f.

Exercise 17-19: Therapeutic Touch

5-21, 17-19

In the therapeutic touch experiment described in Exercise 5-21, subjects identified which of their hands the experimenter had placed her hand over.

a. Identify the null and alternative hypotheses, in symbols and in words, for testing whether or not the subjects could distinguish more often than not over which hand the experimenter's hand was held. Also clearly identify the parameter of interest in words.

b. Combining the results of the 21 subjects, there were a total of 123 correct identifications in 280 repetitions of the experiment. Use these sample data to conduct the test of the hypotheses specified in part a. Report the test statistic and p-value.

c. Explain why it makes sense that the p-value is greater than .5 in this situation.

d. Is it fair to conclude that you should accept the null hypothesis in this situation? Explain.

e. What conclusion do you draw from this study about the effectiveness of therapeutic touch?

Exercise 17-20: Smoking in the Military

In July of 2009, health experts urged the U.S. military to ban smoking, even though smoking has long been associated with military service. Recall from Activity 13-3 that the Centers for Disease Control and Prevention estimate that 20.9% of American adults smoke.

a. Define the relevant parameter for testing whether American soldiers are more likely to smoke than American adults in general.
b. State the appropriate null and alternative hypotheses, in symbols and in words.

An article in *USA Today* (Zoroya, 2009) reported that 37% of U.S. soldiers smoke, but the article did not mention a sample size on which this statistic was based. For now suppose that the sample size was 100.

c. Check whether the technical conditions for the *z*-test are satisfied.
d. Calculate the test statistic and *p*-value.
e. State the test decision at the .01 level, and summarize your conclusion about whether American soldiers are more likely to smoke than Americans in general.
f. If the sample actually involved more than 100 soldiers, would the sample data be statistically significant at the .01 level? Explain how you know.

Exercise 17-21: Hiring Discrimination **17-21,** 18-21

In the case of *Hazelwood School District vs. United States* (1977), the U.S. government sued the City of Hazelwood, a suburb of St. Louis, on the grounds that it discriminated against African-Americans in its hiring of school teachers (Finkelstein and Levin, 1990). The statistical evidence introduced noted that of the 405 teachers hired in 1972 and 1973 (the years following the passage of the Civil Rights Act), only 15 had been African-American. If you include the city of St. Louis itself, then 15.4% of the teachers in the county were African-American; if you do not include the city of St. Louis, then 5.7% of the teachers in the county were African-American. Consider the population of interest to be all potential hires in this school district.

a. Identify the parameter of interest here in words.
b. Conduct a significance test to assess whether or not the proportion of African-American teachers hired by the school district is statistically significantly less than .154 (the proportion of county teachers who were African-American). Use the .01 significance level. Along with your conclusion, report the null and alternative hypotheses, a sketch of the sampling distribution specified by the null hypothesis, a check of the technical conditions, the test statistic, and the *p*-value.
c. Conduct a significance test to assess whether or not the proportion of African-American teachers hired by the school district is statistically significantly less than .057 (the proportion of county teachers who were African-American if you exclude the city of St. Louis). Again use the .01 significance level, and report the null and alternative hypotheses, test statistic, and *p*-value along with your conclusion.
d. Write a few sentences comparing and contrasting the conclusions of these tests with regard to the issue of whether or not the Hazelwood School District was practicing discrimination.

Exercise 17-22: Marriage Ages 8-17, 9-6, 16-19, 16-24, **17-22,** 23-1, 23-12, 23-13, 26-4, 29-17, 29-18

Reconsider Exercise 16-19, in which you analyzed sample data and found the sample proportion of marriages in which the bride was younger than the groom. Conduct a test of significance to address whether or not the sample data support the theory that the bride is younger than the groom in more than half of all the marriages in that particular county. Report the details of the test, and write a short paragraph describing and explaining your findings.

Exercise 17-23: Veterans' Marital Problems **17-23,** 18-20

Researchers found that in a sample of 2101 Vietnam veterans, 777 had been divorced at least once (Gimbel and Booth, 1994). U.S. Census figures indicate that among all American men aged 30–44 when the study was conducted in 1985, 27% had been divorced at least once. Conduct a test of significance to assess whether or not the sample data from the study provide strong evidence that the divorce rate among all Vietnam veterans is greater than 27%.

a. Define the parameter of interest in words.
b. Report the null and alternative hypotheses in symbols and in words.
c. Report the test statistic and *p*-value.

d. Write a one-sentence conclusion, and explain the reasoning process by which the conclusion follows from the test results.

Exercise 17-24: Distinguishing Between Colas 13-13, 15-13, **17-24,** 18-9

Reconsider a cola discrimination taste test in which a subject is presented with three cups, two of which contain the same brand of cola and one of which contains a different brand. The subject is to identify which one of the three cups contains a different brand of cola than the other two. The parameter of interest is this subject's actual probability of correctly identifying the different brand.

a. State the relevant null and alternative hypotheses for testing whether or not the subject would correctly identify the different brand more than one-third of the time in the long run.

Suppose a test consists of 60 trials.

b. Determine the test statistic and *p*-value if the subject identifies the different brand correctly on 21 trials.
c. Determine the test statistic and *p*-value if the subject identifies the different brand correctly on 30 trials.
d. Determine the smallest number of correct identifications that would lead to rejecting the null hypothesis at the $\alpha = .05$ significance level. Along with your answer, report the value of the test statistic and *p*-value corresponding to that answer. [*Hint*: First draw a sketch of the sampling distribution of the sample proportion $\hat{p}$.]
e. Repeat part d for the $\alpha = .01$ significance level.
f. Which answer (part d or e) is greater? Explain why this makes sense.

Exercise 17-25: Monkeying Around

Can rhesus monkeys understand the gestures of humans? Researchers investigated this question with a sample of 40 monkeys (Hauser, Glynn, and Wood, 2007). In front of each monkey, a human separated two boxes and then jutted his head three times in the direction of one particular box. Researchers then waited and saw which box the monkey approached. They found that 30 of the 40 monkeys approached the box that the human had jutted his head toward. Conduct a significance test of whether this sample result provides strong evidence, at the $\alpha = .05$ level, that more than half of all rhesus monkeys would approach the targeted box. Include all six steps of the significance test.

Exercise 17-26: Employee Sick Days

17-26, 18-22

Suppose that the human resources manager for a large company suspects that employees might be abusing the company's sick day policy by taking an inordinate number of sick days on Mondays and Fridays in order to enjoy a long weekend. She decides to investigate this issue by taking a random sample of sick days and determining the proportion of the sick days that were taken on a Monday or Friday.

a. Identify the observational units and variable in this study.
b. Identify (in words) the relevant parameter of interest in this study.
c. State the appropriate null and alternative hypotheses, in terms of the parameter in part b.

Exercise 17-27: Stating Hypotheses

State the appropriate null and alternative hypotheses, in symbols and in words, for the following research questions:

a. A quality control engineer needs to test whether the proportion of defective items in a production process is less than .005.
b. An amateur bowler used to bowl a strike 20% of the time, but she believes that she has now increased that probability.
c. A professor claims that 60% of students at his college have at least one class on Friday and his colleague believes that the proportion is not .6.

Exercise 17-28: Racquet Spinning

11-9, 13-11, 15-11, 17-18, **17-28,** 18-3, 18-12, 18-13, 18-26, 18-27

Tennis players often spin a racquet as a random mechanism for deciding who serves first. Is a spun tennis racquet equally likely to land with the label

up or down? To investigate this question, a tennis racquet was spun 100 times, with the result that it landed up 46 times.

a. Is .46 a parameter or a statistic? Explain.
b. Clearly identify (in words) the parameter of interest in this situation.
c. Conduct a significance test of whether the sample data provide strong evidence against the hypothesis that the racquet is equally likely to land up or down. Report the hypotheses, test statistic, and *p*-value, as well as checking the technical conditions.
d. Interpret the *p*-value. [*Hint*: This is the probability of what, assuming what?]
e. What test decision would you reach at the $\alpha = .10$ significance level?
f. Explain what is wrong with a conclusion that says: "The sample data provide strong evidence that this tennis racquet would land up 50% of the time in the long run."

Exercise 17-29: Matching Pets to Owners 17-29, 17-30

Cultural lore holds that people tend to look like their pets, or vice versa. To investigate this claim, a professor who is also a cat owner showed pictures of three cats to her class of 34 introductory statistics students. One photo was of her cat Tigger, and the other two photos were of different cats that she had never met. Students were asked to guess which cat was actually the professor's.

a. State the appropriate null and alternative hypotheses for testing whether the sample data suggest that students are able to do better than random guessing when trying to pick out the professor's cat. [*Hint*: First think about what the students' probability of success would be if they were randomly guessing.]
b. Check whether the sample size conditions for the one-proportion *z*-test are satisfied.

It turned out that 15 of the 34 students correctly picked out the professor's cat.

c. Calculate the test statistic and *p*-value.
d. Provide a one-sentence summary of this *p*-value: It's the probability of what, assuming what?
e. State the test decision at the $\alpha = .10$ significance level. Also state the test decision at the $\alpha = .01$ significance level.
f. Summarize your conclusion in the context of this study.

Exercise 17-30: Matching Pets to Owners 17-29, 17-30

Reconsider the previous exercise. The professor also gathered data on a larger sample of 43 students, finding that 19 of them correctly picked out her cat.

a. Calculate the sample proportion of successes in this study and in the previous one. How do these proportions compare?
b. Conduct the significance test based on this larger sample size. Report the test statistic and *p*-value. Also state the test decision at the $\alpha = .10$ and $\alpha = .01$ significance levels.
c. Compare the results of this significance test with the ones from the previous exercise. Comment on what this comparison reveals about the effect of sample size on a significance test.

TOPIC 18

More Inference Considerations

18

Do one-third of all American households own a pet cat? If not, is the actual proportion close to one-third? If a baseball player improves his batting success rate substantially, or if a new drug succeeds at alleviating pain more often than the standard drug, is a test of significance guaranteed to reveal the improvements? If not, what factors affect how likely the test is to show the improvements? Finally, if an alien landed on Earth and set out to estimate the proportion of human beings who are female, what might the alien do wrong in constructing its confidence interval? In this topic, you will examine such dissimilar, occasionally even silly, questions as you explore some of the finer points of confidence intervals and significance tests.

Overview

In the previous two topics, you explored and applied the two principal techniques of statistical inference: confidence intervals and tests of significance. This topic will give you more experience with applying these procedures, and you will also investigate their properties, including some fairly subtle ones, further. More specifically, you will consider the relationship between intervals and tests, learn to watch for ways in which these techniques are sometimes misapplied, and explore the important concept of power.

Preliminaries

1. Guess what proportion of American households includes a pet cat. (Activity 18-2)

2. If 31.6% of a random sample of American households includes a pet cat, would you be fairly convinced that the proportion of all American households who have a cat is different from one-third? (Activity 18-2)

3. If 31.6% of a random sample of American households includes a pet cat, would you be fairly convinced that the proportion of all American households who have a cat is much different from one-third? (Activity 18-2)

4. Suppose a baseball player who has always been a .250 career hitter (meaning that his hitting success probability has been 1/4) suddenly improves over one winter to the point where he now has a 1/3 success probability of getting a hit during an at-bat. Do you think he would be likely to convince the team manager of his improvement in a trial consisting of 30 at-bats? (Activity 18-5)

5. Do you think this baseball player would be more or less likely to convince the team manager of his improvement in a trial of 100 at-bats? (Activity 18-5)

In-Class Activities

Activity 18-1: Kissing Couples

15-1, 16-6, 17-1, 17-2, 17-12, **18-1,** 24-4, 24-14

Recall from Activities 15-1, 16-6, 17-1, and 17-2 the study of 124 kissing couples that found 80 leaned to the right. Earlier you found a 90% and a 99% confidence interval for the population proportion of couples who lean to the right (call this population proportion π):

90% CI for π: (.574, .716) 99% CI for π: (.534, .756)

a. Is the value .5 inside either interval? What about the value 2/3 (.667)?

.5: no .667: yes

You also conducted a two-sided test of whether the population proportion differs from two-thirds (.667):

$H_0: \pi = .667$ $H_a: \pi \neq .667$ $z \approx -0.52$ p-value $\approx .606$

b. Based on this p-value, would this test reject or fail to reject the value .667 at the .10 level? What about the .01 level?

.10 level: .01 level:

You conducted a one-sided test of whether the population proportion is greater than .5, but if you had conducted a two-sided test you should have found:

$H_0: \pi = .500$ $H_a: \pi \neq .500$ $z \approx 3.23$ p-value $\approx .001$

Note that the test statistic ($z \approx 3.23$) is the same but this p-value is actually twice the size of the one-sided p-value that you found.

c. Based on this two-sided p-value, would this test reject or fail to reject the value .500 at the .10 level? What about the .01 level?

.10 level: reject .01 level: reject

d. Summarize your answers to parts a–c in the first two lines of the following table:

Hypothesized Value	Contained in 90% CI?	Contained in 99% CI?	Test Statistic	*P*-value	Reject at .10 Level?	Reject at .01 Level?
.500	a.	a.	3.23	.001	c.	c.
.667	a.	a.	0.52	.606	b.	b.
.725	h.	h.	f.	f.	g.	g.

e. What do you notice about the relationship between whether a hypothesized value is in a confidence interval and whether it is rejected?

18

as long as a hypothesized value is good idk whaet th efu lma

f. Use technology to calculate the test statistic and *p*-value for a two-sided test of whether the population proportion differs from .725.

Test statistic: -1.99 *p*-value: .046

g. Would you reject the hypothesis that the population proportion who lean to the right is .725 at the .10 level? What about the .01 level?

.10 level: reject .01 level: fail to reject

h. Fill in the final row of the table in part d. Elaborate on what you answered in part e, about the relationship between whether a hypothesized value is in a confidence interval and whether it is rejected. In particular, how does your test decision for a particular level of α relate to the confidence level?

This activity reveals a **duality** between confidence intervals for estimating a population parameter and a two-sided test of significance regarding the value of that parameter. Roughly speaking, if a 99% confidence interval for a parameter does not include a particular value, then a two-sided test of whether or not the parameter equals that particular value will be statistically significant at the $\alpha = .01$ level. The same is true for a 90% confidence interval with the .10 significance level and for a 95% confidence interval with the .05 significance level and so on.

Confidence intervals and tests of significance are complementary procedures. Whereas tests of significance can establish strong evidence that a parameter differs from a hypothesized value, confidence intervals estimate the magnitude of that difference.

Watch Out

This duality does not always work exactly when the parameter is a population proportion, because the confidence interval (CI) uses the standard error of $\hat{p}$ and the test statistic uses the standard deviation of $\hat{p}$ based on the hypothesized value of π. But the principle still holds, and any departures from the rule will be very small except with small sample sizes.

Activity 18-2: Pet Ownership 13-9, 15-14, 15-15, 18-2, 20-21

Do you have a pet cat?

As you first saw in Exercise 13-9, a sample survey of 80,000 households in 2001, conducted by the American Veterinary Medical Association and published in the *Statistical Abstract of the United States 2006,* found that 31.6% of households sampled owned a pet cat.

a. Is this number a parameter or a statistic? Explain, and indicate the symbol used to represent it.

b. Use technology to conduct a test of whether or not the sample data provide evidence that the population proportion of households who own a pet cat differs from one-third. State the hypotheses, and report the test statistic and p-value. State your test decision for the $\alpha = .001$ significance level. Summarize your conclusion, stating your comments in the context of this study.

population proportion

c. Use technology to produce a 99.9% confidence interval for the population proportion of all American households that own a pet cat. Interpret this interval.

d. Is the confidence interval consistent with the test result? Explain.

yes

e. Do the sample data provide *very* strong evidence that the population proportion of households who own a pet cat is not one-third? Explain whether the p-value or the confidence interval helps you to decide.

we rejected null, so yes and not 1/3

f. Do the sample data provide strong evidence that the population proportion of households who own a pet cat is *very* different from one-third? Explain whether the p-value or the confidence interval helps you to decide.

below .001 by 10^-25

the interval is .322 something, so this was what were looking for.
its extremely strong evidence but the interval is not too different.

This activity illustrates that *statistical* significance is not the same thing as *practical* significance. A statistically significant result is simply one that is unlikely to have occurred by chance alone, which does not necessarily mean the result is substantial or important in a practical sense. Although there is strong reason to believe the actual population proportion of households owning a pet cat does indeed differ from one-third, that proportion is actually *quite close* to one-third—close enough to be not worth arguing over for most people. When you work with very large samples, an unimportant result can be considered statistically significant. The *p*-value tells you whether or not the difference could have arisen by chance alone, not whether or not it is interesting or important. Confidence intervals are useful for estimating the size of the effect involved and should be used in conjunction with significance tests.

Activity 18-3: Racquet Spinning 11-9, 13-11, 15-11, 17-18, 17-28, 18-3, 18-12, 18-13, 18-26, 18-27

Consider again the racquet spinning exercise from Exercises 11-9 and 15-11, and continue to suppose a goal is to determine whether or not the sample data provide evidence that the proportion of up results would differ from .5 in the long run.

a. Suppose you were to spin the racquet 200 times and obtain 113 up results. Use technology to determine the appropriate test statistic and *p*-value. Also report the value of the sample proportion of up landings and whether or not that sample proportion differs significantly from .5 at the $\alpha = .05$ level.

Sample proportion: Test statistic: *p*-value: Significant at .05?

supose pi=.25, and we test pi=/=.25 with threshold .05

if we instead construct a confidence interval with 1-threshold = 95% confidence, which is

(.221, .244), then we are sure with 95% confidence that pi isnt in this threshold.

By the duality, we are also sure that pi=/=.25 with a .05 threshold of statistical significance.

b. Repeat part a supposing you obtained 115 up results in 200 spins.

Sample proportion: Test statistic: *p*-value: Significant at .05? no

c. Repeat part a supposing you obtained 130 up results in 200 spins.

Sample proportion: Test statistic: *p*-value: Significant at .05?

d. In which pair of cases (a and b, a and c, or b and c) are the sample results most similar?

ab because simlar

e. In which pair of cases (a and b, a and c, or b and c) are the decisions about significance at the .05 level the same?

bc

it's not a sharp divide between sig and insig, its up for the reader to decide

The moral here is that it is unwise to treat standard significance levels as sacred. It is much more informative to consider the *p*-value of the test and to base your decision on the strength of evidence provided by the *p*-value. There is no sharp border between significant and insignificant, only increasingly strong evidence as the *p*-value decreases. Reports of significance tests should include the sample information and *p*-value, not just a statement of statistical significance or a decision about rejecting a hypothesis.

Activity 18-4: Women Senators 6-12, 18-4, 18-8

Suppose an alien lands on Earth, notices that the human species has two different sexes, and sets out to estimate the proportion of humans who are female. Fortunately, the alien had a good statistics course on its home planet, so it knows to take a sample of human beings and produce a confidence interval. Suppose the alien then happens upon

the members of the 2011 U.S. Senate as its sample of human beings, so it finds 17 women and 83 men in its sample.

a. Use this sample information to form a 95% confidence interval for the actual proportion of all humans who are female.

.09677 to .2432267

b. Is this confidence interval a reasonable estimate of the actual proportion of all humans who are female?

no bec not random

c. Explain why the confidence interval procedure fails to produce an accurate estimate of the population parameter in this situation.

no bec not ranodm

d. It clearly does not make sense to use the confidence interval in part a to estimate the proportion of women on Earth, but does the interval make sense for estimating the proportion of women in the 2011 U.S. Senate? Explain your answer.

yes, because it extends to the population of the senate lol

Watch Out

This example illustrates some important limitations of inference procedures:

- First, inference procedures do not compensate for the problems of a biased sampling procedure. If the sample is collected from the population in a biased manner, the ensuing confidence interval will be a biased estimate of the population parameter of interest.
- Second, confidence intervals and significance tests use *sample* statistics to estimate *population* parameters. If the data at hand constitute the entire population of interest, then constructing a confidence interval from these data is meaningless. In this case, you know without doubt that the proportion of women in the population of the 2011 U.S. senators is exactly .17, so it is senseless to construct a confidence interval from these data.

Activity 18-5: Hypothetical Baseball Improvements

18-5, 18-16, 18-17

Suppose a baseball player who has always been a .250 career hitter works extremely hard during the winter off-season. He wants to convince the team manager that he has improved, and so the manager offers him a trial of 30 at-bats with which to demonstrate his improvement.

a. State the null and alternative hypotheses to be tested on the resulting data, in symbols and in words.

h0 .25
ha >.25

population prop
his true current batting avg

18

Two kinds of errors can be made with a test of significance: The null hypothesis can be rejected when it is actually true (called a **Type I error**), and the null hypothesis can fail to be rejected when it is actually false (called a **Type II error**). The significance level α of a test puts an upper bound on the probability of a Type I error.

b. Describe (in words) what a Type I error means in the context of the baseball player.

the player by chance overperforms on the 30 trial test despite not improving

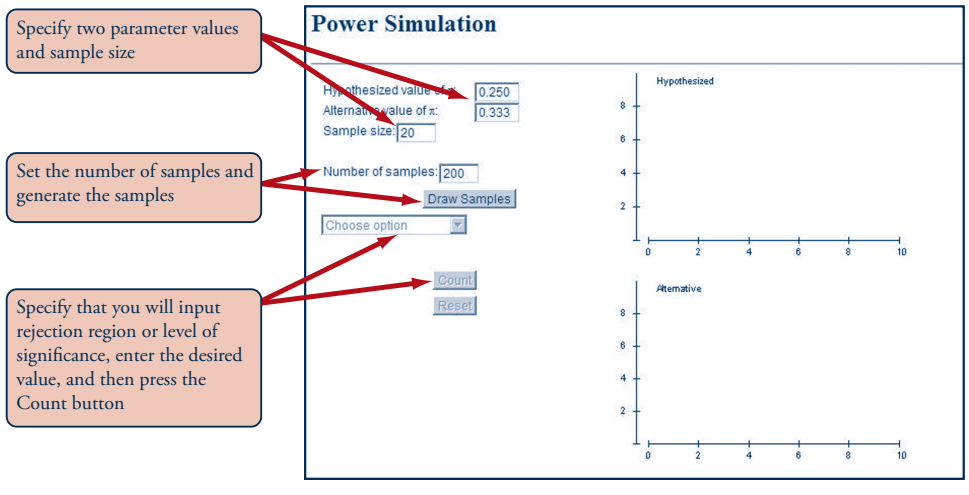
c. Describe (in words) what a Type II error means in this context.

the player by chance underperforms despite improving

A Type I error is sometimes referred to as a *false alarm* because the researcher mistakenly thinks that the parameter value differs from what was hypothesized. Similarly, a Type II error can be called a *missed opportunity* because the parameter really did differ from what was hypothesized, yet the researchers failed to realize it.

Now suppose that the player has genuinely improved to the point where he now has a .333 probability of getting a hit during an at-bat.

d. Open the `Power Simulation` applet. Set the hypothesized value of π to **.250** and the alternative value to **.333**. Set the sample size to **30**, and ask for **200** samples. Click **Draw Samples**. The top dotplot (Hypothesized) shows the distribution of the number of hits in the 30 trials for the 200 samples if the hitter's success probability is .250.



Describe this distribution.

e. Based on the dotplot of the simulated data, approximately how many hits would the player have to get in 30 at-bats so that the probability of a .250 hitter doing that well by chance alone is less than .05? [*Hint:* Use the pull-down menu to select **Level of Significance** and set α to .05. Click on **Count**.]

mound shape symmetical, low 3 high 14 c enter 4 13+ lol

f. The bottom dotplot (Alternative) uses .333 as the success probability. Comment on the amount of overlap between the two distributions.

center 10

g. In what percentage of the 200 samples did the .333 hitter exceed the number of hits that you identified in part e? (This would be the percentage of the 200 samples in which the .333 hitter would do well enough in 30 at-bats to convince the manager that a .250 hitter would have been very unlikely to do that well simply by chance.)

h. Is it very likely that a .333 hitter will be able to establish that he is better than a .250 hitter in a sample of 30 at-bats? Explain. [*Hint:* Refer to your answer to part g and to the overlap between these two distributions that you noted in part f.]

lots of overlap, hard to tell diff

The **power** of a statistical test is the probability that the null hypothesis will be rejected when it is actually false (and therefore should be rejected). Particularly with small sample sizes, a test may have low power, so it is important to recognize that failing to reject the null hypothesis does not mean accepting it as being true. Power equals one minus the probability of a Type II error.

i. Use your simulation results to report the approximate power of this significance test involving the baseball player. [*Hint:* Focus on your answer to part g.]

power increses with N(sampel size)

j. Repeat parts d–i assuming that the player has a sample of 100 at-bats in which to establish his improvement. Write a paragraph summarizing your findings about the power of the test with this larger sample size.

18

Increasing the sample size is one way to obtain a more powerful test, i.e., one that is more likely to detect a difference from the hypothesized value when a difference is actually there. With very large sample sizes, even minor differences can be detected, which reinforces the distinction between statistical and practical significance.

k. If the player had improved to the point of being a .400 hitter, would you expect the test to be more or less powerful than when his improvement was at the .333 level? Explain.

bigger difference easier to spot

more power

l. If your analysis had used the $\alpha = .10$ rather than $\alpha = .05$ significance level, would you expect the test to be more or less powerful? Explain.

asking for less evidence to reject H0, inc power this also inc type1 error

m. In addition to sample size, list two other factors that are directly related to the power of a test.

Self-Check
Activity 18-6: *West Wing* Debate 16-14, 18-6

Recall from Exercise 16-14 that the popular television drama *The West Wing* held a live debate between two (fictional) candidates for president on Sunday, November 6, 2005. Immediately afterward, an MSNBC/Zogby poll found that 54% (of real people) favored Democratic Congressman Matt Santos, played by Jimmy Smits, whereas 38% favored Republican Senator Arnold Vinick, played by Alan Alda. The poll was conducted online with a sample of 1208 respondents; the Zogby company screens the online respondents to try to ensure they are representative of the population of adult Americans.

a. Describe the relevant population and parameter.

b. Use technology to produce 90%, 95%, and 99% confidence intervals for the population proportion who favored Santos.

c. Comment on how these intervals compare (midpoints and widths).

d. Do these intervals suggest that more than half of the population favored Santos? Explain.

e. State the null and alternative hypotheses for testing whether or not the sample data provide strong evidence that more than half of the population favored Santos.

f. What do the intervals in part b say about the *p*-value for testing the hypotheses in part e? [*Hint:* Be careful because you should have a one-sided alternative in part e, but the duality result holds for a two-sided alternative.]

g. Describe what a Type I error would mean and what a Type II error would mean in the context of the hypotheses in part e.

h. Would a test of the hypotheses in part e have more power if 55% of the population actually favored Santos or if 52% of the population actually favored Santos? Explain.

i. Would the sample proportion of .54 favoring Santos be more impressive (i.e., more favorable to Santos) if the sample size were 10,000 or 100? Explain.

Solution

a. The population of interest is all adult Americans who are familiar with these fictional candidates. The parameter (call it π) is the proportion of this population who would have supported Santos if they had been asked.

b. The 90% CI for π is .54 ± .024, which is (.516, .564).
The 95% CI for π is .54 ± .028, which is (.512, .568).
The 99% CI for π is .54 ± .037, which is (.503, .577).

c. The midpoints are all the same, namely .54, the sample proportion of Santos supporters. The 99% CI is wider than the 95% CI, and the 90% CI is the narrowest.

d. Because all three intervals contain only values greater than .5, they do suggest, even with 99% confidence, that more than half of the population favored Santos.

e. H_0: $\pi = .5$ (half of the population favored Santos);
H_a: $\pi > .5$ (more than half of the population favored Santos).

f. Because all three intervals fail to include the value .5, you know that the *p*-value for a two-sided alternative would be less than .10, .05, and .01. Because you have a one-sided alternative in this case (and the sample result is in the conjectured direction), you know that the *p*-value will be less than .01 divided by 2, or .005.

g. A Type I error occurs when the null hypothesis is really true but is rejected. In this case, a Type I error would mean that you conclude that Santos was favored by more than half of the population when in truth he was not favored by more than half. In other words, committing a Type I error means concluding that Santos was ahead (favored by more than half) when he wasn't really. A Type II error occurs when the null hypothesis is not really true but is not rejected (you continue to believe a false null hypothesis). In this case, a Type II error means that you conclude Santos was only favored by half of the population when in truth he was favored by more than half of the population. In other words, committing a Type II error means concluding that Santos was not ahead when he really was.

h. The test would be more powerful if Santos really was favored by 55% rather than 52%. The higher population proportion would make it more likely to reject the null hypothesis that only half of the population favored Santos, because the distribution of sample proportions would center around .55 rather than .52 (farther from .5).

i. The larger sample (10,000) would produce stronger evidence that more than half of the population favored Santos. With less variability in the sampling distribution, the *p*-value would be much smaller.

Watch Out

This entire topic has been about issues to watch out for and the subtle ways in which confidence intervals and significance tests are often misinterpreted or misunderstood. Keep in mind that

- A statistically significant result may not be practically significant, especially with large sample sizes.
- Reporting a *p*-value is more informative than simply reporting a test decision at one particular significance level.
- A test may have insufficient power to reject a null hypothesis even when it is wrong, so you should never *accept* a null hypothesis, especially with small sample sizes.
- When you have data for the entire population, do not apply an inference technique (confidence interval or significance test).

Wrap-Up

This topic aimed to deepen your understanding of confidence intervals and tests of significance so you can better understand the relationship between them and avoid misinterpreting them. You discovered a **duality** between intervals and tests: when a confidence interval includes a particular value, then that value will not be rejected by the corresponding two-sided significance test. You also explored the distinction between practical and statistical significance. For example, a significance test based on a large sample of households convinced you that the proportion of American households with a pet cat is not one-third, but a confidence interval showed that the population proportion is actually quite close to one-third in practical terms. You also learned that common significance levels such as .05 and .01 are useful, but not sacred.

You also investigated the concepts of power and types of error associated with statistical tests. In the baseball example, a **Type I error** means to decide that the player has improved when he really has not (a *false alarm*), and a **Type II error** means to decide that the player has not improved when he really has (a *missed opportunity*). **Power** is the probability of rejecting a null hypothesis that is actually false, and you saw that sample size plays a large role in determining how powerful a test is. Small samples typically lead to tests with low power, meaning you are unlikely to detect a difference or improvement, even when there really is one.

Finally, the alien example reminded you that intervals and tests are bound to produce misleading findings when the sampling method is biased. In fact, statistical inference applies only when a sample has been drawn from a population in the first place. In cases where you have access to the entire population of interest (such as the 100 U.S. senators), you can describe the population but should not apply these statistical inference procedures to the data.

In Brief

Some useful definitions to remember and habits to develop from this topic are

- Confidence intervals and significance tests have different goals. Intervals estimate the value of an unknown population parameter, also indicating the amount of uncertainty in the estimate. Significance tests assess how much evidence the sample data provide against a particular hypothesized value for the population parameter.
- Whenever a two-sided test rejects a particular hypothesized value at a certain significance level α, then the confidence interval at the corresponding confidence level (for example, 95% confidence for $\alpha = .05$) will not include that hypothesized value.
- Whenever a test gives a statistically significant result, it is useful to follow up with a confidence interval (CI). A CI can help determine whether the result is also of practical significance.
- Especially with large sample sizes, a statistically significant result may fail to be practically important.
- Reporting a *p*-value is much more informative than simply providing a yes/no statement of whether or not the sample result is significant at a certain α level.
- A Type I error occurs when you reject a null hypothesis that is actually true. A Type II error occurs when you fail to reject a null hypothesis that is actually false.

- Sample size affects the power of a test. All else being equal, larger samples produce more powerful tests than smaller samples.
- Power is also influenced by the significance level α and by the actual value of the parameter. Using a smaller (i.e., more stringent) significance level *reduces* the power of the test. Having an actual parameter value that differs more from the hypothesized value *increases* the power of the test.
- Never state a conclusion as "accept the null hypothesis." Failing to have enough evidence to reject the null hypothesis does not necessarily mean that you have proven the null hypothesis to be true.
- Be very cautious about generalizing to a larger population, with either a confidence interval or a significance test, if the sample is not selected randomly from that population.
- Do not apply these inference procedures when you have access to data from the entire population.

You should be able to

- Describe what a significance test reveals about a confidence interval, and vice versa. (Activity 18-1)
- Appreciate the distinction between statistical and practical significance, and explain how to assess each of these based on *p*-values and confidence intervals. (Activity 18-2)
- Recognize the limitation of presenting test results based on fixed significance levels rather than *p*-values. (Activity 18-3)
- Identify situations in which statistical inference should not be applied because the sampling method is biased or because data for the entire population are available. (Activity 18-4)
- Describe what the two types of errors and statistical power mean in a particular context. (Activity 18-5)
- Perform simulation analyses to estimate power and probability of Type II error. (Activity 18-5)
- Describe the impact of sample size, significance level, and alternative value on power. (Activity 18-5)

In the next two topics, you will learn a confidence interval and a test of significance for a population *mean* rather than a population *proportion.* In other words, you will analyze a quantitative variable rather than a categorical one. You will find that whereas the details of implementing the procedures necessarily change, the basic structure, reasoning, and interpretation do not.

Exercises

Exercise 18-7: Charitable Contributions

16-18, **18-7**

Recall from Exercise 16-18 that the 2004 General Social Survey found that 1052 from a random sample of 1334 American adults claimed to have made a financial contribution to charity in the previous year.

a. Find a 90% confidence interval for the proportion of all American households that made a financial contribution to charity in the previous year.
b. Repeat part a with a 99% confidence interval.
c. Based on these confidence intervals, without carrying out a test of significance, indicate

whether or not this sample proportion differs significantly from .75 at the $\alpha = .01$ level. Explain your reasoning.

d. Based on these confidence intervals, without doing a test of significance, indicate whether or not this sample proportion differs significantly from 80% at the $\alpha = .10$ level. Explain your reasoning.

Exercise 18-8: Women Senators

6-12, 18-4, 18-8

Recall from Activity 18-4 that the 2011 U.S. Senate consists of 17 women and 83 men.

a. Treat these numbers as sample data and calculate the test statistic for the significance test of whether or not the population proportion of women is less than .50.

b. Use the test statistic and the Standard Normal Probabilities Table (Table II) or technology to calculate the *p*-value of the test.

c. If the goal is to decide whether or not women constitute less than half of the entire U.S. Senate of 2011, does this test of significance have any meaning? Explain.

Exercise 18-9: Distinguishing Between Colas

13-13, 15-13, 17-24, 18-9

Recall the cola taste test described in Exercise 17-24. Each subject is presented with three cups, two of which contain the same brand of cola and one of which contains a different brand. The subject is to identify which one of the three cups contains a different brand of cola than the other two.

a. Report (in symbols and in words) the null hypothesis of the test, corresponding to the conjecture that a subject is simply guessing at identifying the one cup of soda that differs from the other two.

b. Suppose one particular subject (Randy) is actually able to identify the different brand 50% of the time in the long run. Would this subject's sample data necessarily lead to rejecting the null hypothesis? Explain.

c. Suppose that Randy participates in $n = 50$ trials. Verify that the null hypothesis would be rejected at the .05 significance level if the subject obtained a sample proportion of correct identifications of .46 or greater.

d. Is there greater than a 50/50 chance that Randy, with his actual 50% success rate, will get a sample proportion of .46 or greater? Explain, based on a CLT calculation.

e. Would the probability in part d increase, decrease, or remain the same if the sample size were 100 instead of 50? Explain.

f. Would the probability in part d increase, decrease, or remain the same if Randy's probability of a correct identification were 2/3 instead of 1/2? Explain.

Exercise 18-10: Voter Turnout

4-19, 18-10

A random sample of 2613 adult Americans in 1998 revealed that 1783 claimed to have voted in the 1996 presidential election.

a. Use these sample data to construct a 99.9% confidence interval for π, the proportion of all adult American eligible voters who voted in that election.

b. Even though this truly was a random sample, do you really have 99.9% confidence that this interval captures the actual proportion who voted in 1996? Explain.

c. The Federal Election Commission reported that 49.0% of those eligible to vote in the 1996 election had actually voted. Is this value included within your interval?

d. Do you think this interval succeeds in capturing the proportion of all eligible voters who would claim to have voted in 1996? Explain how this parameter differs from that in part c.

e. Explain why the confidence interval renders it unnecessary to conduct a significance test of whether π differs from .49 at the .001 level.

Exercise 18-11: Phone Book Gender

4-17, 16-16, 18-11

Recall from Exercise 4-17 the sample data collected from a random page of the San Luis Obispo County telephone book: 36 listings had both male and female names, 77 had male names, 14 had female names, 34 had initials only, and 5 had pairs of initials. Before collecting the data, the authors conjectured that fewer than half of the names in the phone book would be female.

a. A total of how many first names were studied? (Ignore listings with only initials.) How many of them were female names? What proportion of the names were female?
b. Identify the observational units in this study.
c. Identify the population and the parameter of interest.
d. Do the sample data support the authors' conjecture that fewer than half of the names in the phone book would be female at the $\alpha = .05$ level? Report the details of the test and write a short paragraph describing your findings and explaining how your conclusions follow from the test results. Also discuss the technical conditions.
e. Accompany your test with a 95% confidence interval. Interpret the interval, and comment on how it relates to the test result.
f. Do the sample data provide evidence that fewer than half of the residents of San Luis Obispo County are female? Explain your answer, and be sure to discuss how this question differs from the one you addressed in part d.

Exercise 18-12: Racquet Spinning

11-9, 13-11, 15-11, 17-18, 17-28, 18-3, **18-12,** 18-13, 18-26, 18-27

Recall from Exercise 11-9 and Activity 15-11 that 100 spins of a tennis racquet produced 46 up results and 54 down results. Suppose you want to test whether or not a spun tennis racquet is equally likely to land up or down, that is, whether or not it would land up 50% of the time in the long run.

a. Identify with a symbol and in words the *parameter* of interest in this experiment.
b. Considering the stated goal of the study, is the alternative hypothesis one-sided or two-sided? Explain.
c. Specify the null and alternative hypotheses for this study, both in symbols and in words.
d. Use technology to calculate the relevant test statistic and *p*-value.
e. Use technology to calculate how the test statistic and *p*-value would have been different if the 100 spins had produced 54 up and 46 down results.
f. In either of these cases (46 of one outcome and 54 of the other), do the sample data provide strong evidence that a spun tennis racquet would *not* land up 50% of the time in the long run? Would you reject the null hypothesis at the .05 significance level?

Exercise 18-13: Racquet Spinning

11-9, 13-11, 15-11, 17-18, 17-28, 18-3, 18-12, **18-13,** 18-26, 18-27

Reconsider the previous exercise. Now suppose that the goal of the study is to investigate whether or not a spun tennis racquet tends to land up less than 50% of the time.

a. Restate the null and alternative hypotheses (in symbols).
b. Use technology to determine the test statistic and *p*-value of the one-sided test, assuming that 46 of the 100 sample spins landed up. Report the test statistic and *p*-value of the test, and comment on how they compare to the ones found with the two-sided test.
c. Still supposing that the goal of the experiment is to investigate whether or not a spun tennis racquet tends to land up less than 50% of the time, use technology to determine the test statistic and *p*-value of the one-sided test assuming that 54 of the 100 sample spins landed up. Again report the test statistic and *p*-value of the test, and comment on how they compare to the values found with the two-sided test.
d. For the sample results in part c, explain why the formal test of significance is unnecessary.

Exercise 18-14: Penny Activities

16-9, 16-10, 16-11, 17-9, **18-14,** 18-15

Recall Exercise 16-9 and the data you gathered on penny flipping, spinning, and tilting.

a. For each of those three activities, use your sample data to test whether or not the population proportion of heads differs significantly from one-half at the .05 significance level.
b. Comment on whether or not your test results are consistent with your 95% confidence intervals found in Exercise 16-9.

Exercise 18-15: Penny Activities

16-9, 16-10, 16-11, 17-9, 18-14, **18-15**

Statistics professor Robin Lock has asked his students to flip, spin, and tilt pennies for many years, and he has compiled a running total of the results:

- 14,709 heads in 29,015 flips
- 9,197 heads in 20,422 spins
- 10,087 heads in 14,611 tilts

a. For each of the three activities, determine a 95% confidence interval for the actual probability (long-run proportion) of heads.
b. Explain why it makes sense that these intervals are so narrow.
c. Based on these intervals, for which activities would you reject at the .05 significance level that the probability of heads is .5? Explain.
d. For each of the three penny activities, conduct the test of whether or not the probability of heads differs significantly from one-half. Report the test statistics and *p*-values. Do the results agree with your answer to part c?
e. With regard to penny *flipping,* would you regard the difference from one-half as *practically* significant even if it is statistically significant? Explain.

Exercise 18-16: Hypothetical Baseball Improvements

18-5, **18-16,** 18-17

Reconsider the baseball player from Activity 18-5. Suppose that he actually became a .400 hitter (i.e., had a .4 probability of getting a hit during an at-bat). Investigate whether the power for a significance test with a sample size of 30 at-bats is greater or less than it was when he became a .333 hitter. To do this, repeat your simulation analysis of Activity 18-5 for a .400 hitter. Write a paragraph reporting your findings concerning the change in the test's power.

Exercise 18-17: Hypothetical Baseball Improvements 18-5, 18-16, **18-17**

Consider once again the baseball player from Activity 18-5 who became a .333 hitter. Investigate whether the test is more or less powerful at higher significance levels by repeating your simulation analysis of Activity 18-5 using $\alpha = .10$ rather than $\alpha = .05$ as the significance level. Write a paragraph reporting your findings concerning the change in the test's power resulting from this change in significance level.

Exercise 18-18: Flat Tires

17-3, 17-4, 17-10, 17-16, **18-18,** 24-17

Reconsider the flat-tire scenario from Activities 17-3 and 17-4.

a. Report the null and alternative hypotheses being tested.
b. Describe what a Type I error and a Type II error mean in this context.
c. Perform a simulation using the `Power Simulation` applet to approximate the power of this test with a sample size of 100 and a significance level of .05, assuming that the right-front tire is actually chosen 50% of the time.
d. Would the power in part c increase, decrease, or remain the same if the sample size were 200 and all else remained unchanged? Explain.
e. Would the power in part c increase, decrease, or remain the same if the significance level were .10 and all else remained unchanged? Explain.
f. Would the power in part c increase, decrease, or remain the same if the right-front tire were actually chosen 40% of the time and all else remained unchanged? Explain.

Exercise 18-19: Emotional Support

4-15, **18-19**

Recall from Exercise 4-15 that sociologist Shere Hite received mail-in questionnaires from 4500 women, 96% of whom claimed that they give more emotional support than they receive from their husbands or boyfriends (Moore, 1995). Also recall that an ABC News/*Washington Post* poll of a random sample of 767 women found that 44% claimed to give more emotional support than they receive.

a. Determine the margin of error for each of these surveys. Also report each survey's 95% confidence interval for the proportion of all American women who feel they give more emotional support than they receive.
b. Are these two confidence intervals similar? Do they overlap at all?
c. Which survey has the smaller margin of error, that is, the narrower confidence interval?
d. Which of these two confidence intervals do you have more confidence in? Explain.

Exercise 18-20: Veterans' Marital Problems 17-23, **18-20**

Recall from Exercise 17-23 the study that investigated whether or not Vietnam veterans tend to get divorced at a higher rate than men aged 30–44 at that time (Gimbel and Booth, 1994).

a. State the null and alternative hypotheses for this study.
b. Describe what a Type I error would mean in this context.
c. Describe what a Type II error would mean in this context.

Exercise 18-21: Hiring Discrimination

17-21, **18-21**

Recall from Exercise 17-21 the legal case in which the U.S. government sued the Hazelwood School District on the grounds that it discriminated against African-American teachers in its hiring practices. Describe what a Type I error and a Type II error would mean in this context. Would you consider one of these to be a more serious error than the other? Explain.

Exercise 18-22: Employee Sick Days

17-26, **18-22**

Recall from Exercise 17-26 that the Human Resources manager for a large company suspects that employees might be abusing the company's sick day policy by taking an inordinate number of sick days on Mondays and Fridays in order to enjoy a long weekend. She decides to investigate this issue by taking a random sample of sick days and determining the proportion of these sick days that were taken on a Monday or Friday. She is considering a random sample of either 100 sick days or 200 sick days.

a. Which sample size should she choose to have a more powerful test?
b. Explain what it means to have a more powerful test, using nontechnical language that the manager could understand even if she has not studied statistics.

Exercise 18-23: Friend or Foe

Lab 1, **18-23,** 18-24, 18-25

In a study reported in the November 2007 issue of *Nature*, researchers investigated whether infants take into account an individual's actions toward others in evaluating that individual as appealing or aversive, perhaps laying the foundation for social interaction (Hamlin, Wynn, and Bloom, 2007). In one component of the study, 10-month-old infants were shown a climber character (a piece of wood with google eyes glued onto it) that could not make it up a hill in two tries. Then they were alternately shown two scenarios for the climber's next try, one where the climber was pushed to the top of the hill by another character (helper) and one where the climber was pushed back down the hill by another character (hinderer). The infant was alternately shown these two scenarios several times. Then the child was presented with both pieces of wood (the helper and the hinderer) and asked to pick one to play with. The researchers worked with a sample of 16 infants, whom they trusted to be representative of the population of all infants. They found that 14 of the 16 infants chose the helper over the hinderer.

a. Calculate the sample proportion of infants who chose the helper toy. Is it larger than one-half, as the researchers expected?
b. Define the relevant parameter in this study.
c. State the appropriate null and alternative hypotheses for testing whether infants in general tend to choose the helper toy more than the hinderer toy.
d. Show that the technical conditions for the one-proportion z-test are not satisfied.

Exercise 18-24: Friend or Foe

Lab 1, 18-23, **18-24,** 18-25

Reconsider the previous exercise. Although the technical conditions of the z-test are not satisfied, you can approximate the p-value of the test with simulation. Such a simulation is equivalent to tossing a fair (50/50) coin 16 times, once for each infant, letting heads represent a choice for the helper toy and tails represent a choice for the hinderer toy.

a. Use the `Coin Tossing` applet to simulate tossing a fair coin 16 times. (Click on the **Toss Coin** button.) Record the number of heads, representing the number of infants who choose the helper toy.
b Now set the number of repetitions to **1000** and click on **16 Tosses**. The graph will now have 1000 dots, each dot displaying the number of heads in a set of 16 tosses. Describe this distribution of *number of heads per sample.*
c. Enter **14** in the As extreme as box and press **Count** to count how many of these 1000 simulated samples resulted in 14 or more Heads. Report the proportion of these 1000 repetitions with 14 or more Heads.
d. Based on these simulation results, would you say that a result as extreme as the researchers

obtained (14 of 16 infants choosing the helper toy) is very surprising, under the assumption that the choice of toys is really 50/50 between the helper and hinderer? Explain.

e. Based on your answers to parts c and d, would you say that the sample result (14 of 16 infants choosing the helper toy) provides strong evidence that infants really do prefer the helper toy? Explain.

Exercise 18-25: Friend or Foe

Lab 1, 18-23, 18-24, 18-25

Reconsider the previous two exercises. Using an appropriate probability model (called the binomial distribution, as discussed in online Topic 32), it can be shown that the *p*-value for testing these hypotheses is .002.

a. Interpret this *p*-value: The probability is .002 that _____, assuming that _____.

b. Summarize the conclusion about the research question that you would draw from this *p*-value. Also, describe the reasoning process that leads to this conclusion.

Exercise 18-26: Racquet Spinning

11-9, 13-11, 15-11, 17-18, 17-28, 18-3, 18-12, 18-13, 18-26, 18-27

Recall from Exercise 15-11 that a tennis racquet was spun 100 times, landing in the up position for 46 of those spins.

a. Using the $\alpha = .10$ significance level, does this sample result lead to rejecting the hypothesis that the racquet is equally likely to land up or down? Report the *p*-value along with your test decision.

b. Produce and interpret a 90% confidence interval for the parameter.

c. Is the confidence interval consistent with the test decision? Explain.

d. Explain how the confidence interval helps you understand why it's not appropriate for the test decision to *accept* the null hypothesis.

Exercise 18-27: Racquet Spinning

11-9, 13-11, 15-11, 17-18, 17-28, 18-3, 18-12, 18-13, 18-26, 18-27

Reconsider the previous exercise.

a. Describe what committing a Type I error would mean in this context.

b. Describe what committing a Type II error would mean in this context.

c. Based on the test decision that you reached in part a of the previous exercise, which type of error (I or II) *could* you be making? Explain.

Exercise 18-28: Pop vs. Soda

4-25, 18-28

Recall from Exercise 4-25 that the pop-vs.-soda website (`popvssoda.com`) asked people visiting the site to vote for their preferred name for a generic cola drink: pop, soda, coke, other. A total of 293,772 responses were received from the United States, of which 108,707 answered "pop," 120,130 answered "soda," 46,883 answered "coke," and 18,052 answered "other."

a. Determine a 99.9% confidence interval for the population proportion who would answer "soda."

b. Based only on the confidence interval in part a, what can you say about the *p*-value for a significance test of whether the population proportion who would answer "soda" differs from .5? Explain your answer.

c. Explain why the confidence interval in part a is so narrow, despite a very high confidence level.

d. Does the very large sample size in this study ensure that the sample is likely to be representative of the population? Explain.

Exercise 18-29: Seat Belt Use

4-26, 18-29

In a study of 612 drivers who entered convenience stores in El Paso, Texas (Parada, et al., 2001), 75% reported that they always wear seat belts.

a. Describe the relevant population and parameter for this study.

b. Determine and interpret a 99% confidence interval for the parameter.

Researchers observed the 612 drivers as they drove into the convenience store parking lot, and they saw that 61.5% were wearing their seat belts at the time.

c. Is .615 within the 99% confidence interval?

d. Comment on a limitation of surveys that is revealed by the study.

e. Use the observed percentage of 61.5% who were wearing a seat belt to produce a 99% confidence interval for the proportion of drivers (who frequent convenience stores in El Paso, Texas) who actually wear a seat belt.

TOPIC 19

Confidence Intervals: Means

19

How many M&M candies do college students tend to grab? For how long did an "average" student at your school sleep last night? How much weight does a typical college student carry in his or her backpack? Although these questions concern different contexts, they nevertheless sound quite similar. In this topic, you will learn how to construct a confidence interval to address these and other questions.

Overview

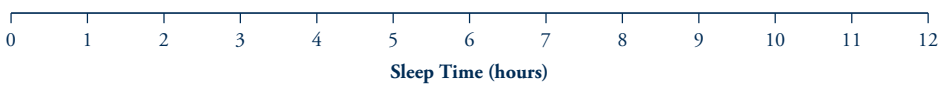
In Topic 16, you explored confidence intervals for a population *proportion*. In this topic, you will turn from binary categorical variables to quantitative variables, so now the population *mean* is the parameter of interest. You will investigate and apply confidence interval procedures for estimating a population mean. Although some of the procedure's details are different, and you will work with a new probability model called a *t*-distribution, you will find that the reasoning, structure, and interpretation of a confidence interval remain unchanged.

Preliminaries

1. Guess the average amount of sleep that a student in your class got last night. Then make guesses for the least and most hours of sleep that a student in your class got last night. (Activity 19-4)

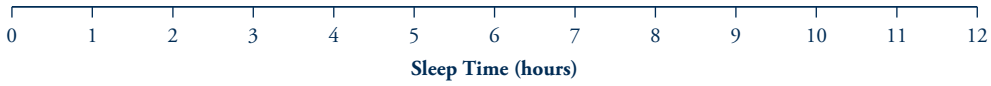
   Average: Least: Most:

2. Mark on the following scale an interval that you believe with 90% confidence to include the mean amount of sleep (in hours) that students at your school got last night. (Activity 19-4)



sleep hours
3
3.5
8
5
7
5
8
7
7
6
6
4
6.5
7
6
4
8.5

3. Mark on the following scale an interval that you believe with 99% confidence to include the mean amount of sleep (in hours) that students at your school got last night. (Activity 19-4)



4. Which of these two intervals is wider? (Activity 19-4)

5. Guess the average number of hours per week that third or fourth graders spend watching television. (Exercise 19-16)

Collect Data

1. Record the times that the students in your class went to bed last night and woke up this morning. Also calculate and record the amount of sleep time, in hours (e.g., 2 hours and 15 minutes = 2.25 hours). (Activity 19-4)

Student	Bedtime	Wake Time	Sleep Time
1			
2			
3			
. . .			

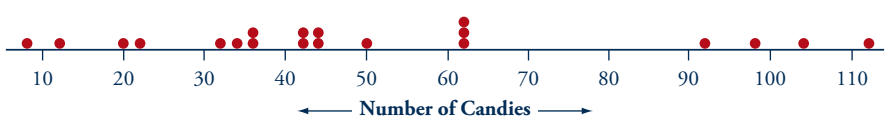
In-Class Activities

Activity 19-1: M&M Consumption 19-1, 19-3, 19-31, 22-30, 22-31, 22-37

Brian Wansink, author of the popular book *Mindless Eating*, conducts research into factors that influence people to eat without conscious regard for how much they are eating. One of his studies involved a sample of 20 undergraduates from the University of Illinois at Urbana-Champaign. Participants were allowed to take as many M&M candies as they wished from a 24-ounce bowl for consumption during a lab study session that day. After they selected the M&Ms individually, the researcher determined how many M&Ms each person had selected (`www.MindlessEating.org`). Suppose you plan to offer a study session and need to know how many candies to bring and therefore would like to estimate how many candies students tend to take on average.

a. Is the variable *number of M&Ms selected* categorical or quantitative?

Here is a dotplot and summary statistics of the results.



n	Mean	SD	Min	Q_L	Median	Q_U	Max
20	50.15	30.31	7.00	32.35	42.50	62.00	111.00

b. Are the numbers 50.15 and 30.31 parameters or statistics? What symbols would you use to represent them?

19

c. Identify in words a suitable parameter in this study. What symbol represents this parameter?

d. Do you know the value of the parameter in this study? Is it more likely to be close to 50.15 or far from it? Explain.

As you have seen, you can form an interval estimate of a parameter by starting with the sample statistic and going two standard deviations to either side of it. The Central Limit Theorem for a sample mean tells you the standard deviation of a sample mean $\bar{x}$ is $\sigma/\sqrt{n}$, where σ represents the population standard deviation and n the sample size.

e. Do you know the value of σ for this M&M study? Explain what σ represents in your own words.

When the population standard deviation σ is known, a confidence interval for a population mean is given by $\bar{x} \pm z^* \sigma/\sqrt{n}$. However, a major drawback to using this procedure is that it requires you to know the value of the population standard deviation σ, which you almost never know. (After all, if you knew the population standard deviation, isn't it likely you would also know the population mean and therefore not need to estimate it?)

f. What is a reasonable substitute for σ that you can calculate from the sample data?

To estimate the standard deviation of a sample mean $\bar{x}$, replace the population standard deviation σ with the sample standard deviation s, producing $SE_{\bar{x}} = s/\sqrt{n}$. This is known as the **standard error** of the sample mean.

g. Calculate the standard error for this study. Describe in your own words what it measures.

6.77752204888

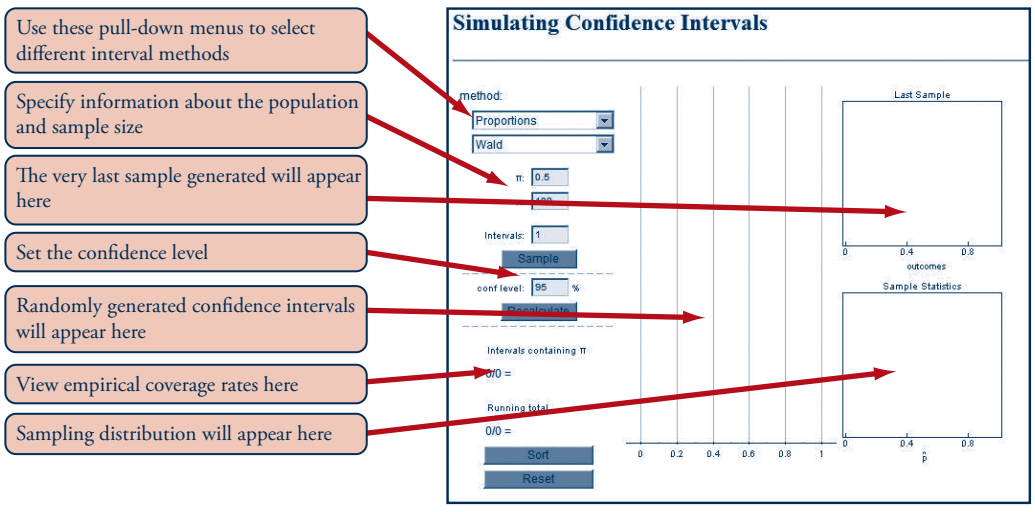
h. Add and subtract two standard errors to/from the sample mean to form an (informal) interval estimate for the population mean μ.

Taking a handful?

It seems reasonable to form a confidence interval for a population mean μ with $\bar{x} \pm z^* \sigma/\sqrt{n}$. But a problem emerges with this procedure, especially with small sample sizes, as you will discover through a simulation analysis.

i. Open the `Simulating Confidence Intervals` applet. Use the pull-down menus to switch to **means**, and then select the method called ***z* with sigma**. Set the population mean to $\mu = \mathbf{50}$ and the population standard deviation to $\sigma = \mathbf{30}$. Set the sample size to $n = \mathbf{20}$ and the confidence level to **95**%. Ask for **200** intervals, and click on **Sample**.

97%



94.9%

Notice that the applet generates 200 samples, calculates a 95% confidence interval for μ using the sample mean $\bar{x}$ from each sample and $\sigma = 30$, and colors the intervals green when they succeed in capturing the population mean and red when they fail. What percentage of these 200 intervals succeed in capturing the value of μ (indicated by the vertical line) between the two endpoints?

j. Continue to click **Sample** until you have produced 1000 intervals. What is the running total percentage of intervals that succeed in capturing the population mean? Is this value close to what you expected? Explain.

k. Now change the method to **z with s**. This method estimates σ with each sample's standard deviation s in calculating each interval. Click on **Sample** and continue to click until you have generated 1000 intervals. What percentage of these 1000 intervals succeed in capturing the population mean? Is this value very close to 95%?

93.5%

l. Now change the sample size to $n = 10$ and generate 1000 intervals. What do you notice about the widths of the intervals with the smaller sample size? What do you notice about the coverage rate—that is, what proportion of the 1000 intervals succeed in capturing the population mean?

19

91.8% success.

m. Now change the sample size to $n = 30$ and generate 1000 intervals. Again, comment on the widths and the coverage rate.

93.9%

The problem with simply replacing σ with s is that noticeably fewer than 95% of the intervals succeed in capturing the population mean when the sample size is small, even though they are being called 95% confidence intervals. (This happens because two unfortunate things happen with some samples: The sample mean is a little farther than expected from the population mean, and the sample standard deviation is a little less than the population standard deviation.) To compensate for this, you need to make the intervals a bit longer, so more of them will succeed and the overall percentage will be closer to the stated 95%. To do this, you use a different multiplier than the z^* critical value. You use what is called a ***t*-distribution**. The resulting t^* critical values (see Activity 19-2) will reflect the additional uncertainty introduced by estimating σ with s.

n. Use the pull-down menu to change from **z with s** to **t**, and generate 1000 intervals. Look at the coverage rates for 1000 intervals using $n = 10$, $n = 20$, $n = 30$. Is this close to 95% in each case?

close enoguh lmfao

Therefore, we will use a t-interval to estimate a population mean.

Confidence Interval for a Population Mean (*t*-interval):

$$\bar{x} \pm t^* \frac{s}{\sqrt{n}}$$

where t^* is the appropriate critical value from the t-distribution with $n - 1$ degrees of freedom for the desired confidence level.

Two **technical conditions** must be satisfied for this t-interval procedure to be valid:

there are three technical condition slmao

degrees of freedom is dsample size -1, so it means that if sampel size is 30, the degrees of freedom is 29.

z is rigid and doesn't change, while t changes in thsi case.

- The sample is a simple random sample from the population of interest.
- *Either* the sample size is large ($n \geq 30$ as a guideline) *or* the population is normally distributed.

Independent, so 10n≤N

The second technical condition stems from what you learned in Topic 14: The sampling distribution of the sample mean will be normal if the population itself is normal, or this sampling distribution will be approximately normal for any population shape as long as the sample size is large. A sample size of at least 30 is generally regarded as large enough for the procedure to be valid. If the sample size is less than 30, examine visual displays of the sample data to see if they appear to follow a normal distribution. If the sample appears to be roughly normal, then you can assume that the population follows a normal distribution.

.0007 that a random number chosen is an outlier in a normal curve

The *t*-procedure is fairly *robust* in that it tends to give reasonable results even for small sample sizes as long as the population is not severely skewed and does not have extreme outliers.

if you have a small sample, and get an outlier, you say "use caution, since outlier may suggest population may not be normal".

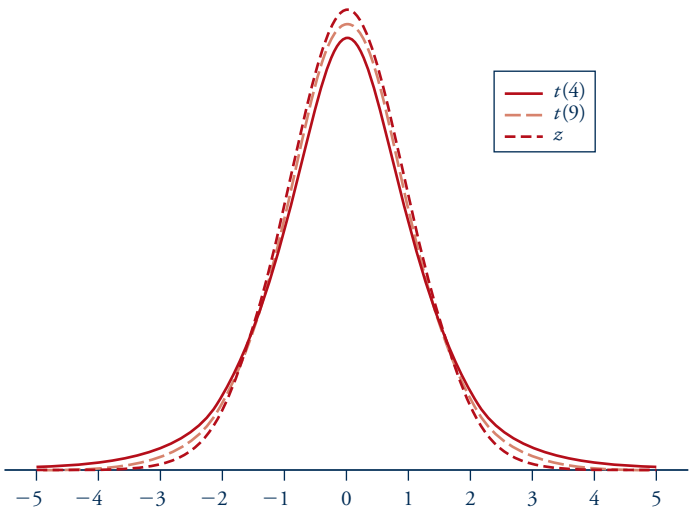
The reasoning and interpretation of these confidence intervals are the same as always: If you were to repeatedly take random samples from the population and apply this procedure over and over, then in the long run, 95% of the intervals generated would succeed in containing the population mean. This allows you to say that you are 95% confident that the one interval you actually constructed contains the actual value of the population mean.

Watch Out

- As mentioned in the last topic, it really is crucial to keep notation and terminology straight in your mind. Remember that μ denotes a population mean and $\bar{x}$ denotes a sample mean. The whole point of the confidence interval is to estimate the *unknown* value of μ, based on the *observed* value of $\bar{x}$.
- Similarly, remember that σ stands for a population standard deviation and s for a sample standard deviation. Be especially careful with the phrase "standard deviation." Not only is the population standard deviation σ different from the sample standard deviation s, but also the standard deviation of the sample mean ($\sigma/\sqrt{n}$) and the standard error of the sample mean ($s/\sqrt{n}$) are all different quantities as well.

Activity 19-2: Exploring the *t*-Distribution 19-2, 20-5, 20-6

The *t*-distribution is actually an entire family of distribution curves, similar to the normal distributions. Whereas a normal distribution is identified by its mean and standard deviation, a *t*-distribution is characterized by an integer number called its **degrees of freedom** (abbreviated df). These *t*-distributions are mound-shaped and centered at zero, but they are more spread out (i.e., they have wider, fatter tails) than standard normal distributions. As the number of degrees of freedom increases, the tails get lighter and the *t*-distribution gets closer and closer to a normal distribution.



Previously, you have used the Standard Normal Probabilities Table to find critical values; now you will learn how to use a *t*-table. A *t*-table (Table III) can be found in the Appendix. Part of this table is reproduced below. Notice that each line of the table corresponds to a different value for the degrees of freedom. Always start by going to the relevant line for the degrees of freedom with which you are working. In conducting inferences about a population mean, the degrees of freedom can be found by subtracting 1 from the sample size (df $= n - 1$). Next, note that across the top of the table are various values for the area to the right. Finally, observe that the body of the table gives values such that the probability of lying to the right of that value (equivalent to the area to the right of that value under the *t*-distribution) is given at the top of each column.

Suppose you need to find the critical value t^* for a 95% confidence interval based on a sample size of $n = 10$.

a. Draw a rough sketch of the *t*-distribution with 9 degrees of freedom. [*Hint:* It should look very much like a standard normal curve.]

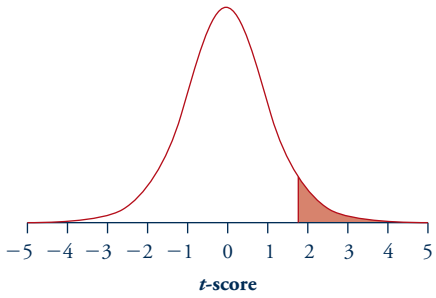
b. The critical value t^* for a 95% confidence interval is the value such that 95% of the area under the curve is between $-t^*$ and t^*. Shade this area on your sketch.

c. What is the area to the right of t^* under the curve? [*Hint:* This area is not .05.]

d. Look at the *t*-table (Table III) to find the value of t^* with that area to its right under a *t*-distribution with 9 degrees of freedom. Report this value.

2.262

Table III: *t*-Distribution Critical Values



The table reports the critical value for which the area to the right is as indicated.

Area to Right	0.2	0.1	0.05	0.025	0.01	0.005	0.001	0.0005
Confidence Level	60%	80%	90%	95%	98%	99%	99.80%	99.90%
df								
1	1.376	3.078	6.314	12.706	31.821	63.657	318.317	636.607
2	1.061	1.886	2.920	4.303	6.965	9.925	22.327	31.598
3	0.978	1.638	2.353	3.182	4.541	5.841	10.215	12.924
4	0.941	1.533	2.132	2.776	3.747	4.604	7.173	8.610
5	0.920	1.476	2.015	2.571	3.365	4.032	5.893	6.869
6	0.906	1.440	1.943	2.447	3.143	3.708	5.208	5.959
7	0.896	1.415	1.895	2.365	2.998	3.500	4.785	5.408
8	0.880	1.397	1.860	2.306	2.897	1.355	4.501	5.041
9	0.883	1.383	1.833	2.262	2.821	3.250	4.297	4.781
10	0.879	1.372	1.812	2.228	2.764	3.169	4.144	4.587
11	0.876	1.363	1.796	2.201	2.718	3.106	4.025	4.437

e. Is this critical value less than or greater than the critical value z^* from the standard normal distribution for a 95% confidence interval? Explain why this is helpful, based on your motivation for needing the *t*-distribution instead of the *z*-distribution.

because it accounts for the wider variablity of a smaller sample size.

Critical values t^* from the *t*-distribution are always greater than their counterparts from the *z*- (normal) distribution, reflecting the greater uncertainty introduced by estimating σ with *s*. This larger multiplier makes the intervals just long enough that 95% of them (or whatever the confidence level) succeed in capturing the value of the population mean.

f. Find the critical value t^* for a 95% confidence interval based on a sample size of $n = 20$. How does this value compare to the previous t^* value? Explain why this is appropriate for the interval procedure as well. [*Hint:* Think about whether a larger sample size would increase or decrease the uncertainty in estimating σ by s.]

g. Find the critical value t^* for a 90% and 99% confidence interval, based on a sample size of $n = 20$. Which is greater? Explain why this is appropriate.

h. Find the critical value t^* for a 95% confidence interval based on a sample size of $n = 130$. [*Hint:* When the desired value for degrees of freedom is not shown in the table, always round down to be *conservative,* because it's better to have a longer interval than necessary to ensure the achieved confidence level is at least as large as the stated confidence level.]

i. How has the t^* value changed as you increased the degrees of freedom by increasing the sample size? How does t^* with 129 degrees of freedom compare to z^*?

bruh idke

Note that when the sample size is large, we no longer have to pay much penalty for estimating the population standard deviation σ with the sample standard deviation s, and so the t^* values converge to the z^* values as the sample size increases. In fact, the last row of the t-table (labeled as infinite degrees of freedom) provides an alternative way to look up z^* critical values.

Activity 19-3: M&M Consumption 19-1, 19-3, 19-31, 22-30, 22-31, 22-37

Recall the M&M consumption data from Activity 19-1, with $\bar{x} = 50.15$ candies and $s = 30.31$ candies.

a. State and comment on the technical conditions for the validity of a one-sample t-interval with these data.

sample greater than 30 so not a valid normal curve, random selection no, 200 < all mnm consumer

be cautious of outliers as well.

b. Calculate and interpret a 95% *t*-confidence interval for the parameter of interest. Make sure your interpretation is in the context of this study and clearly describes the parameter and the population you are willing to generalize to.

n Mean SD Min QL Median QU Max
20 50.15 30.31 7.00 32.35 42.50 62.00 111.00

we are 95% confident that pop mean of all mnm consumers falls between xxx

35.964 to 64.336

c. Write a sentence interpreting what the phrase "95% confidence" means in this context. [*Hints*: Note that interpreting the *level* is different from interpreting the interval as you did in part b. Think back to the green and red intervals in the applet used in Activity 19-1. You should not use the words "confidence" or "sure" in your interpretation.]

on th elong trun 95% of the intervals capture the population mean

d. How does this interval compare to the informal interval you calculated in part h of Activity 19-1, particularly, the midpoint and width? Explain why the similarities and differences make sense.

Midpoint:

Width: got a little wider ig idk

e. Use technology to confirm your calculations of the 95% confidence interval (CI) in part b. Then use technology to produce a 90% confidence interval and a 99% confidence interval for the population mean number of candies.

90% CI for μ: 38.431-61.869

99% CI for μ: 30.76-69.54

f. Comment on how the midpoints and widths of these intervals compare.

th elarger the confiddence interval he greter the argin of the inteval

g. Suppose the sample size had been 200 instead of 20, but the sample statistics turned out exactly the same. How would you expect a 95% confidence interval to differ in this case from the 95% confidence interval in part b? Explain.

get smaller

h. Produce the confidence interval mentioned in part g, and describe how the interval has changed. [*Hint*: Comment on both its midpoint and width.]

45.924-54.376

get smaller

i. Suppose the sample size had been 200 but the sample distribution of number of candies had been sharply skewed. Would this *t*-interval calculation still be valid? Explain.

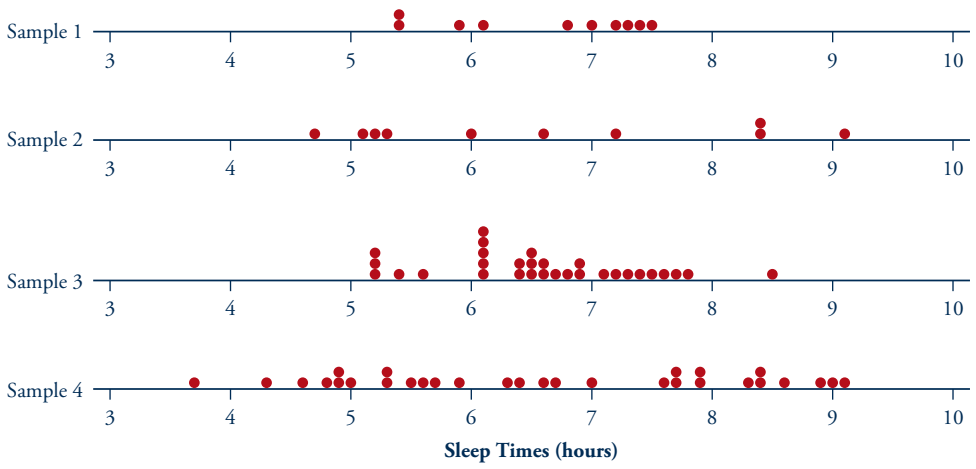
it wouldnt be valid because we're integrating an estimation of tha normal curve and this is not normla

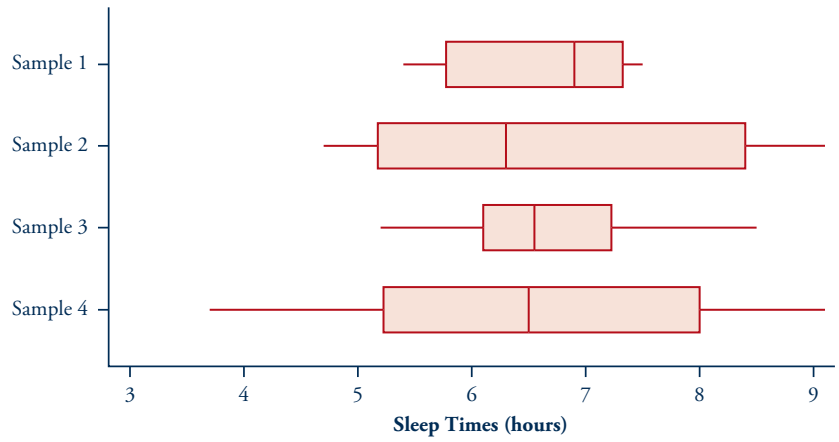
Watch Out

When checking the second technical condition for the *t*-interval, normality of the population is not necessary if the sample size is large. So, even if the sample data have a skewed distribution, the *t*-interval can still be used if the sample size is large.

Activity 19-4: Sleeping Times 8-29, 19-4, 19-5, 19-12, 19-19, 20-2, 20-7, Lab 5

Suppose you want to estimate the mean sleep times of all students at your school last night. Consider the four different (hypothetical) samples of sleep times presented in the dotplots and boxplots shown here.





a. The following descriptive statistics were calculated from these sample data. Fill in the Sample Number column by figuring out which statistics go with which plots.

Sample Number	Sample Size	Sample Mean	Sample SD
3	30	6.6	0.825
1	10	6.6	0.825
2	10	6.6	1.597
4	30	6.6	1.597

b. What do all of these samples have in common?

same mean

c. What strikes you as the most important difference between the distribution of sleep times in sample 1 and in sample 2?

one is much more variable

d. What strikes you as the most important difference between the distribution of sleep times in sample 1 and in sample 3?

the sample size

The following table gives a 95% confidence interval for the population mean sleep time, based on each sample's data (presented in the same order as in the previous table):

Sample Number	Sample Size	Sample Mean	Sample SD	95% CI
3	30	6.6	0.825	(6.29, 6.91)
1	10	6.6	0.825	(6.01, 7.19)
2	10	6.6	1.597	(5.46, 7.74)
4	30	6.6	1.597	(6.00, 7.20)

e. Comparing samples 1 and 2, which produces a more precise estimate of μ (i.e., a narrower confidence interval for μ)? Explain why this makes sense. [*Hint:* Refer to your answer in part c.]

5.96

f. Comparing samples 1 and 3, which produces a more precise estimate of μ (i.e., a narrower confidence interval for μ)? Explain why this makes sense. [*Hint:* Refer to your answer in part d.]

slightly skew
center at 5.97
spread is 1.65

In addition to sample size and confidence level, the sample standard deviation plays a role in determining the width of a confidence interval for a population mean. Samples with more variability (less precise measurements) produce wider confidence intervals.

Activity 19-5: Sleeping Times 8-29, 19-4, 19-5, 19-12, 19-19, 20-2, 20-7, Lab 5

Now consider the data on sleep times collected from the students in your class in the Collect Data section.

a. Use technology to produce graphical displays of the distribution. Write a few sentences commenting on key features of this distribution.

it shuld be representative bruh

b. Use technology to calculate the sample size, sample mean, and sample standard deviation. Record these values here, and indicate the symbol used to represent each value.

5.97 mean, standard deviation is 1.65, Q1 4.5 Q3 7

c. State the technical conditions required for the *t*-interval to be a valid procedure. For each condition, comment on whether or not it appears to be satisfied.

d. Use technology to construct a 90% confidence interval for estimating the mean sleep time for *all* students at your school on that particular night. Also write a sentence interpreting this interval in context.

one an be 96 confident that the poulation man of all stats students are are between 5.27 to 6.67

95% confidence.

e. Count how many of the sample sleep times fall within this confidence interval. What percentage of the sample does this represent?

6

f. Is the percentage in part e close to 90%? Should it be? Explain. [*Hint:* The issue here is not whether or not the sample was selected randomly from the population.]

no because small sample

Watch Out

Confidence intervals of this type estimate the value of a population *mean.* They do not estimate the values of *individual* observations in the population or in the sample. You do not expect 90% of the sample data to be in the interval, nor are you claiming that 90% of the population's sleep times are contained in this interval. Especially with large samples, it is possible that very few sample (or population) values might fall inside the interval. In this case, you are simply 90% confident that the value of the *population mean* sleeping time among all students at your school is somewhere inside this interval. This means that, in the long run, 90% of intervals constructed this way will capture the population mean. (Of course, this claim is only valid if you are willing to regard your sample as representative of the entire population on this issue, which would be problematic if, say, this was an early morning class.)

Self-Check

Activity 19-6: Backpack Weights 2-13, 10-12, 19-6, 20-1, 20-17, Lab 7

Refer to Exercises 2-13 and 10-12, which described a study (Mintz, Mintz, Moore, and Schuh, 2002) in which student researchers recorded the body weights and backpack weights for a sample of 100 students on the Cal Poly campus (`Backpack`).

a. Identify the observational units, variable (and its type), sample, and population in this study.

b. Use technology to create the *ratio of backpack weight to body weight* variable for each student in the sample. Produce and comment on graphical displays and numerical summaries of the distribution of this ratio for these 100 students.

c. Determine a 99% confidence interval for the population mean weight ratio among all Cal Poly students at the time this study was conducted.

d. Interpret this interval, and also explain what the phrase "99% confidence" means.

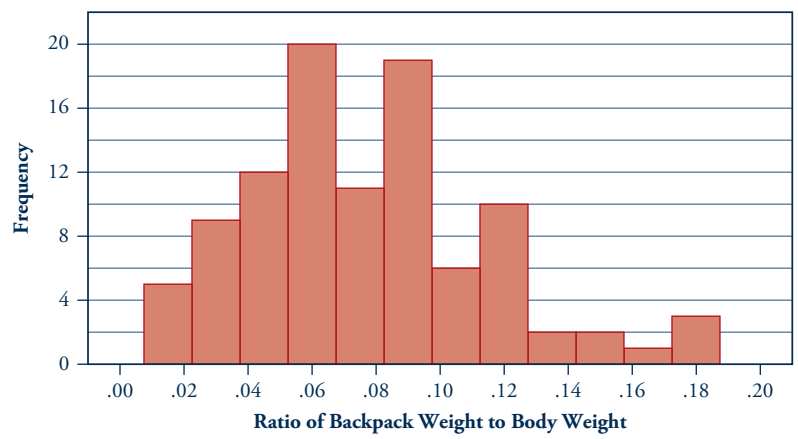
e. Comment on whether or not the technical conditions required for the validity of this interval are satisfied (as well as you can with the information provided).

f. Would you expect 99% of the students in the sample to have a weight ratio in this interval? Would you expect this of 99% of the students in the population? Explain.

Solution

a. The observational units are the students. The variable is the *ratio of backpack weight to body weight,* which is quantitative. The sample is the 100 Cal Poly students whose weights were recorded by the student researchers. The population is all Cal Poly students at the time the study was conducted.

b. The following dotplot reveals that the distribution of these weight ratios is a bit skewed to the right. The center is around .07 or .08 (mean $\bar{x} = .077$, median $= .071$). The five-number summary is (.016, .050, .071, .096, .181), so students in the sample carried as little as 1.6% of their weight in their backpacks and as much as 18.1% of their weight in their backpacks. The standard deviation of these ratios is $s = .037$.



c. Calculating a 99% CI for the population mean by hand using

$$\bar{x} \pm t^* \frac{s}{\sqrt{n}}$$

gives $.0771 \pm 2.639 \times .0366/\sqrt{100}$, which is $.0771 \pm .0097$, which corresponds to the interval from .0674 through .0868. (This t^* value should be based on 99 df, but we used 80 df here, the closest value less than 99 that appears in Table III.) Using technology gives a slightly more accurate 99% CI for μ of .0675 through .0867.

d. You are 99% confident that the mean ratio of backpack-to-body weights among all Cal Poly students at the time of this study is between .0674 and .0868. In other words, you are 99% confident that the average Cal Poly student carries between 6.74% and 8.68% of his or her body weight in his or her backpack. By "99% confident" you mean that 99% of all intervals constructed with this method would succeed in capturing the actual value of the population mean weight ratio.

e. The first condition is that the sample be randomly selected from the population. This is not literally true in this case, because the student researchers did not obtain a list of all students at the university and select randomly from that list, but they did try to obtain a representative sample. The second condition is either that the population of weight ratios is normal or that the sample size is large. In this case, the sample size is large ($n = 100$, which is much greater than 30), so this condition is satisfied, even though the distribution of ratios in the sample is somewhat skewed (and so presumably is the population).

f. You do not expect 99% of the sample, nor 99% of the population, to have a weight ratio between .0674 and .0868. You are 99% confident that the population *mean* weight ratio is between these two endpoints. In fact, only 18 of the 100 students in the sample have a weight ratio in this interval.

Watch Out

- Now that we have discussed two confidence interval procedures, the first question to ask when constructing a confidence interval is whether the underlying variable of interest is categorical or quantitative because this determines whether you are estimating a population proportion or a population mean. In the previous example, the *ratio* variable is quantitative. This may seem a little strange because this *ratio* variable gives numbers between 0 and 1, but it is not categorical (a numerical value is observed for each person). Another unsubtle clue is that part c clearly asks you to estimate a population *mean*.
- Be sure to carry many decimal places of accuracy in your intermediate calculations. If you round off too much in intermediate steps, your final calculations can turn out to be quite wrong. Go ahead and round your final answer to a few decimal places, but remember that a mean need not be an integer.
- Do not forget to divide by $\sqrt{n}$ in calculating the standard error portion of the CI for a population mean.
- As always, make sure that your interpretation of the calculated interval refers to context. Do not simply say that you're 99% confident that μ is within the interval without explaining what μ represents in the context of the study. This entails clarifying the variable and population of interest, as well as the parameter (e.g., mean weight ratio of all Cal Poly students).
- Remember the second technical condition is either/or, not both. If the sample size is large, the data do not have to be normally distributed. On the other hand, a small sample does not render the *t*-interval invalid; if the sample data appear to be roughly normal, then the *t*-interval is still considered valid even with a small sample size.

Wrap-Up

This topic introduced you to confidence intervals for a population *mean.* You found that the reasoning, structure, and interpretation of these intervals are the same as for a population proportion. For instance, these intervals have the basic form: *sample statistic* ± (*critical value*) × (*standard error of statistic*). And you still interpret "95% confidence" to mean that 95% of all intervals generated by this procedure succeed in capturing the population parameter (in this case, population *mean*) of interest. One difference is that the sample mean is the statistic of interest. Another is that the standard error of the statistic ($s/\sqrt{n}$) involves the sample standard deviation. Perhaps the most important difference is that the critical value is based on the ***t*-distribution** rather than the standard normal (z-) distribution.

You also studied the properties of these intervals. For example, you again found that a larger sample size produces a narrower interval, and a higher confidence level generates a wider interval (if all else remains the same). You also learned that more variability in the sample data leads to a wider confidence interval.

One common misunderstanding of these intervals is to think (wrongly) that 95% of the *data* should fall within the interval. The interval estimates the population *mean,* not individual data values. Especially with a large sample that produces a narrow interval, it is not uncommon for the interval to capture only a small percentage of the individual data values. You should not expect 95% of the sample values, or 95% of the population values, to fall within a 95% confidence interval for μ.

In Brief

Some useful definitions to remember and habits to develop from this topic are

- Remember to check the technical conditions before applying the procedure. This procedure requires either a large sample size *or* a normally distributed population. If the sample size is less than 30, examine visual displays of the sample data to see if they appear to follow a normal distribution.
- Be careful to consider whether or not the sample was randomly selected from the population of interest before generalizing the confidence interval to that population.
- A larger sample produces a narrower interval.
- A higher confidence level produces a wider interval.
- More sample variability produces a wider interval.
- Do not forget that this confidence interval procedure applies to a *quantitative* variable, not a categorical variable.
- Remember this confidence interval estimates a population *mean,* not individual data values.

You should be able to

- Explain why the t-distribution, rather than the normal (z-) distribution, is needed to conduct inference for a population mean. (Activity 19-1)
- Describe critical values from the t-distribution to be used in calculating confidence intervals, and understand related properties of the t-distribution. (Activity 19-2)

- Calculate and interpret a confidence interval for a population mean. (Activities 19-3, 19-5)
- Explain the effects of sample size and sample variability on a confidence interval. (Activity 19-4)
- Identify a common misconception that a confidence interval estimates individual values rather than a population mean. (Activity 19-5)

In the next topic, you will continue to work with quantitative variables, turning your attention to a test of significance about a population mean.

Exercises

Exercise 19-7: Body Temperatures

12-1, 12-19, 14-3, 14-18, 15-9, **19-7,** 19-8, 20-11, 22-10, 23-3

Recall from Activity 12-1 and Exercise 15-3 that the body temperatures (in degrees Fahrenheit) have been recorded for a sample of 130 healthy adults (Shoemaker, 1996). The sample mean body temperature is 98.249°F, and the sample standard deviation is 0.733°F.

a. Create and examine a dotplot or histogram of the sample data (`BodyTemps`). Also create and examine a normal probability plot. Does the *sample* distribution of body temperatures appear to be roughly normal?
b. Is normality of the *population* of body temperatures required for this *t*-procedure to be valid with these data? Explain.
c. Comment on whether or not the other technical conditions required for the validity of this *t*-procedure are satisfied.
d. Calculate a 95% confidence interval for the population mean body temperature, based on the sample results for these 130 healthy adults.
e. Write a sentence interpreting this interval. [*Hint*: Ask yourself what you believe to be in this interval with 95% confidence.] Then write a separate sentence interpreting what the phrase "95% confidence" means.
f. Based only on this confidence interval and assuming you have a representative sample, does it appear that 98.6°F is a plausible value for the mean body temperature for the population of all healthy adults? Explain.
g. Suppose the sample size had been 13 rather than 130, but the sample statistics turned out exactly the same. How would you expect a 95% confidence interval to differ in this case from the 95% interval in part d? Explain.
h. Produce the confidence interval mentioned in part g, and describe how the interval has changed. [*Hint*: Comment on both its midpoint and width.]

Exercise 19-8: Body Temperatures

12-1, 12-19, 14-3, 14-18, 15-9, 19-7, **19-8,** 20-11, 22-10, 23-3

Reconsider the previous exercise. Now consider the body temperatures of men and women separately.

a. Produce a 95% confidence interval for the population mean body temperature of a healthy male. Then do the same for the population of healthy females.
b. Interpret these intervals.
c. Comment on how the intervals compare to each other, addressing whether or not they suggest that men and women have different mean body temperatures.
d. Report the half-width of these two intervals. Which is wider? What aspect of the data causes the interval to be wider?
e. Compare the half-width of these intervals to the half-width of the interval based on the entire sample from the previous exercise. Explain why these two intervals are wider than that one.
f. Investigate and comment on whether or not the technical conditions required for the validity of this procedure appear to be satisfied for each gender.

Exercise 19-9: Social Acquaintances

9-8, 9-9, 10-13, 10-14, 19-9, 19-10, 20-12

Reconsider the data that you collected in Topic 9, based on the social-acquaintance exercise described by Malcolm Gladwell in *The Tipping Point.*

a. Use these sample data to produce a 90% confidence interval for the population mean number of acquaintances from this list among all students at your school.
b. Interpret what this interval says. Also include an explanation of what "90% confidence" means in this context.
c. Determine how many and what proportion of the sampled students have an acquaintance number that falls within this interval.
d. Is this proportion close to 90%? Should it be? Explain.

Exercise 19-10: Social Acquaintances

9-8, 9-9, 10-13, 10-14, 19-9, 19-10, 20-12

Reconsider the previous exercise. The file `AcquaintancesCP` contains data from 99 undergraduate students at Cal Poly who engaged in this exercise during the winter of 2006.

a. Determine a 90% confidence interval for the population mean number of acquaintances from this list among all students at Cal Poly in the winter of 2006.
b. Compare your class results with those from Cal Poly. Write a paragraph or two summarizing your comparison, including graphical displays and numerical summaries, as well as commenting on the confidence intervals.

Exercise 19-11: Nicotine Lozenge

1-16, 2-18, 5-6, 9-21, 19-11, 20-15, 20-19, 21-31, 22-8

Recall from Exercise 1-16 the experiment that investigated the effectiveness of a nicotine lozenge for subjects who wanted to quit smoking (Shiffman et al., 2002). Before the treatments began, subjects answered background questions, including how many cigarettes they smoked per day. Among the 1818 subjects in the study, the average was 22.0 cigarettes per day, and the standard deviation was 10.8 cigarettes per day.

a. Identify the population and parameter of interest.
b. Produce a 99% confidence interval for the population mean number of cigarettes smoked per day.
c. Based on this interval, does it seem plausible to assert that the population mean is 20 cigarettes (one pack) per day? Explain.

Exercise 19-12: Sleeping Times

8-29, 19-4, 19-5, 19-12, 19-19, 20-2, 20-7, Lab 5

Consider again the sleeping times collected in class and the confidence interval that you produced from that data. Describe how the confidence interval would have been different if the only change had been

a. A larger standard deviation among the sample sleeping times.
b. A smaller sample size.
c. A larger sample mean by 0.5 hours.
d. Each person's sleep time had been 15 minutes longer than reported.

Exercise 19-13: Critical Values

12-18, 16-2, 16-20, 19-13

a. Use the *t*-table to find the critical values t^* corresponding to the following confidence levels and degrees of freedom, filling in a table like the one shown here with those critical values:

	Confidence Levels			
Degrees of Freedom	80%	90%	95%	99%
4				
11				
23				
80				
Infinity				

b. Does the critical value t^* get larger or smaller as the confidence level increases (if the number of degrees of freedom remains the same)?
c. Does the critical value t^* get larger or smaller as the number of degrees of freedom increases (if the confidence level remains the same)?
d. Do the critical values from the *t*-distribution corresponding to infinitely many degrees of freedom look familiar? Explain. (Refer to Topic 16 if they do not.)

Exercise 19-14: Sentence Lengths

The following data are the lengths (measured as numbers of words) in a sample of 28 sentences from Chapter 3 of John Grisham's novel *The Testament:*

17	21	8	32	13	16	17	37	27	20	30	15	64	34
18	26	23	17	5	10	29	9	22	18	7	16	13	10

a. Create graphical display of these data. Write a few sentences commenting on key features of the distribution.
b. Use these sample data to produce a 95% confidence interval for the mean length among all sentences in this book.
c. Comment on whether or not the technical conditions necessary for the validity of this procedure seem to be satisfied.
d. Remove the outlier, and recalculate the 95% confidence interval. Comment on how this interval has changed. Did removing the outlier have much impact on the interval?

Exercise 19-15: Coin Ages

12-16, 14-1, 14-2, 19-15

Recall from Activity 14-1 the population of 1000 coins from which you selected a random sample.

a. As you did in Activity 14-1, use a table of random digits to select a random sample of 10 pennies from this population. Record their ages.
b. Use this sample to construct a 90% confidence interval for the mean age of the pennies in this population.
c. Do you think that the technical conditions for this procedure have been met? Explain.
d. The population mean age among these 1000 pennies is 12.26 years. Does the interval in part b succeed in capturing this population mean?
e. If you had planned to construct a 95% confidence interval rather than a 90% interval, would you have been more likely to capture the population mean?
f. If you had planned to take a random sample of 40 pennies (rather than 10) and then construct a 90% confidence interval, would you have been more likely to capture the population mean?

Exercise 19-16: Children's Television Viewing

1-15, 19-16, 20-4, 22-14, 22-15

Recall from Exercise 1-15 the study that investigated a relationship between watching television and obesity in third and fourth grade children (Robinson, 1999). Prior to assigning children to treatment groups, researchers gathered baseline data on their television viewing habits. Children were asked to report how many hours of television they watched in a typical week. The 198 responses had a mean of 15.41 hours and a standard deviation of 14.16 hours.

a. Are these values (15.41, 14.16) parameters or statistics? Explain.
b. Do you think that the technical conditions for the confidence interval for μ have been met? Explain.
c. Use this sample information to determine 90%, 95%, and 99% confidence intervals for the mean hours of television watched per week among all third and fourth graders.
d. Do any of these intervals include your guess from the Preliminaries section?
e. In this situation would it make much difference if you used the z^* critical values rather than t^*? Explain.

Exercise 19-17: Close Friends

19-17, 19-18, 22-1, 22-5, 22-22

The 2004 General Social Survey (GSS) interviewed a random sample of adult Americans. For one question the interviewer asked: "From time to time, most people discuss important matters with other people. Looking back over the last six months–who are the people with whom you discussed matters important to you? Just tell me their first names or initials." The interviewer then recorded how many names or initials the respondent mentioned. Results are tallied in the following table (`CloseFriends`):

Number of Close Friends	0	1	2	3	4	5	6	Total
Count (Number of Respondents)	397	281	263	232	128	96	70	1467

a. Identify the observational units and variable in the study. Is the variable categorical or quantitative?
b. This distribution is sharply skewed to the right, but a *t*-interval is still valid. Explain why.

c. Use technology to produce a 90% confidence interval for the mean number of close friends in the population of American adults.
d. Which two of the following are reasonable interpretations of this confidence interval and its confidence level:

- You can be 90% confident that the mean number of close friends in the population is between the endpoints of this interval.
- Ninety percent of all people in this sample reported a number of close friends within this interval.
- If you took another sample of 1467 people, there is a 90% chance that its sample mean would fall within this interval.
- If you repeatedly took random samples of 1467 people, this interval would contain 90% of your sample means in the long run.
- If you repeatedly took random samples of 1467 people and constructed *t*-intervals in this same manner, 90% of the intervals in the long run would include the population mean number of close friends.
- This interval captures the number of close friends for 90% of the people in the population.

e. For one of the incorrect interpretations in part d, explain why it is incorrect.
f. Describe how the interval would change if all else remained the same except

- The sample size were larger.
- The sample mean were larger.
- The sample values were less spread out.
- Every person in the sample reported one more close friend.

Exercise 19-18: Close Friends

19-17, **19-18,** 22-1, 22-5, 22-22

Refer to the previous exercise.

a. Produce and interpret a 90% confidence interval for the population proportion of people who would report having 0 close friends. [*Hint:* Note that the parameter being estimated is not a mean.]
b. Repeat part a for the proportion of people who would report having five or more close friends.
c. What information is gained in reporting a *t*-interval for the population mean rather than for these intervals?

Exercise 19-19: Sleeping Times

8-29, 19-4, 19-5, 19-12, **19-19,** 20-2, 20-7, Lab 5

Consider your analysis of students' sleep times from Activity 19-5. Let μ represent the mean sleep time in the population of all students at your school. Identify each of the following statements as legitimate or illegitimate interpretations of the 95% confidence interval that you produced. For each illegitimate interpretation, explain why it is incorrect.

a. You can be 95% confident that the interval contains the true value of μ.
b. If you repeatedly took random samples of college students and generated 95% confidence intervals in this manner, then in the long run 95% of the intervals so generated would contain the actual value of μ.
c. The probability is .95 that μ lies within the interval.
d. You can be 95% confident that the sleep time for any particular student falls within the interval.
e. Ninety-five percent of the students in the population have sleep times that fall within the interval.

Exercise 19-20: Planetary Measurements

8-12, 10-20, **19-20,** 27-13, 28-20, 28-21

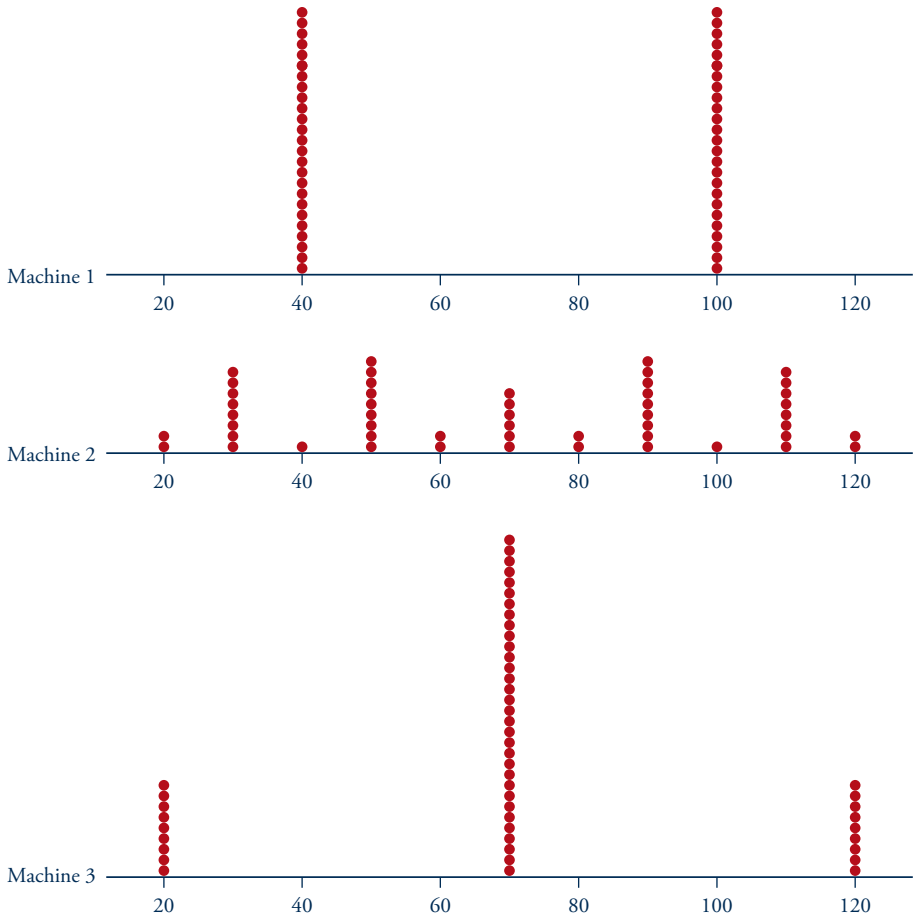
Reconsider the data presented in Exercise 8-12 that listed (among other things) the distance from the sun for each of the nine planets in our solar system. The mean of the distances turns out to be 1102 million miles, and the standard deviation is 1341 million miles.

a. Use these numbers to construct a 95% confidence interval.
b. Does this interval make any sense at all? If so, what population parameter does it estimate? Do you know the exact value of that parameter? Explain.

Exercise 19-21: Hypothetical ATM Withdrawals

9-24, **19-21,** 22-25

The following dotplots display samples of withdrawal amounts from three different automatic teller machines (`HypoATM`):



Withdrawal Amount (dollars)

a. Write a paragraph comparing and contrasting the three distributions of ATM withdrawal amounts.

b. Use technology to calculate the sample size, sample mean, and sample standard deviation of the withdrawal amounts for each machine. Also use technology to determine a 95% confidence interval for the population mean withdrawal amount among all withdrawals for each machine. Record the results in the following table.

	Sample Size	Sample Mean	Sample SD	95% CI for μ
Machine 1				
Machine 2				
Machine 3				

c. Summarize what this exercise reveals about whether or not a confidence interval for a mean describes all aspects of a dataset.

Exercise 19-22: House Prices

19-22, 26-1, 27-5, 27-25, 28-2, 28-12, 28-13, 29-3

The following 25 values are the prices and sizes of 25 houses in Arroyo Grande, California (`HousePricesAG`). These houses were selected as a random sample from houses listed on the website `zillow.com` on February 7, 2007.

Address	Price ($)	Size (square feet)
2130 Beach St.	311,000	460
2545 Lancaster Dr.	344,720	1030
415 Golden West Pl.	359,500	883
990 Fair Oaks Ave.	414,000	728
845 Pearl Dr.	459,000	1242
1115 Rogers Ct.	470,000	1499
579 Halcyon Rd.	470,000	1419
1285 Poplar St.	470,000	952
1080 Fair Oaks Ave.	474,000	1014

Address	Price ($)	Size (square feet)
690 Garfield Pl.	475,000	1615
1030 Sycamore Dr.	490,000	1664
620 Eman Ct.	492,000	1160
529 Adler St.	500,000	1545
646 Cerro Vista Cir.	510,000	1567
926 Sycamore Dr.	520,000	1176
227 S Alpine St.	541,000	1120
654 Woodland Ct.	567,500	1549
2230 Paso Robles St.	575,000	1540
2461 Ocean St.	580,000	1755
833 Creekside Dr.	625,000	1844

a. Examine graphical displays of both variables. Based on these graphs, do you think a *t*-confidence interval would be valid for both variables, only one variable, or neither? Explain.
b. Determine the appropriate t^* value for a 90% confidence interval based on this sample.
c. Regardless of your answer to part a, produce a 90% confidence interval for the population mean price and also a 90% confidence interval for the population mean size.
d. Interpret these intervals, including a statement of whether you believe the interval to be valid.

Exercise 19-23: Credit Card Usage

1-9, 16-12, **19-23,** 20-10

Refer to the study described in Exercise 16-12. The Nellie Mae organization found that in a random sample of undergraduate students taken in 2004, the average credit card balance was $2169. Take the sample size to be 1074, which corresponds to 76% of the sample of 1413 students who hold a credit card.

a. What additional information is needed to produce a 95% confidence interval for the mean credit card balance in the population of all undergraduate students who held a credit card in 2004?
b. The Nellie Mae report did not mention the sample standard deviation of these credit card balances. Suppose this standard deviation was $1000. Produce and interpret a 95% confidence interval for the population mean in this case.
c. Would you expect 95% of the population to have a credit card balance in this interval? Explain.

Exercise 19-24: Properties of Confidence Intervals

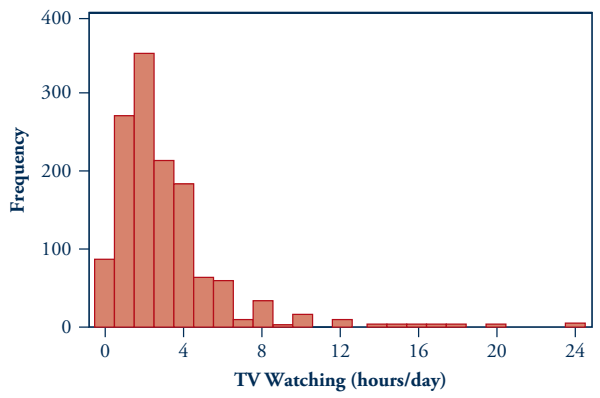
The following questions pertain to a *t*-interval $\bar{x} \pm t^* s/\sqrt{n}$ for a population mean μ

a. How does the critical value t^* change as the sample size gets larger?
b. As the sample size gets very large, what value does t^* approach?
c. The midpoint of this interval is always equal to what?
d. Write the expression for the margin-of-error of the interval.
e. Does the margin-of-error get larger, smaller, or stay the same as the sample size increases (assuming that all else remains the same)?
f. Does the margin-of-error get larger, smaller, or stay the same as the sample standard deviation increases (assuming that all else remains the same)?
g. Does the margin-of-error get larger, smaller, or stay the same as the sample mean increases (assuming that all else remains the same)?
h. Is it true that this *t*-interval procedure can only be applied when the sample data closely follow a normal distribution?
i. Is it true that this *t*-interval procedure can only be applied when the sample size is large?

Exercise 19-25: Television Viewing Habits

8-28, 10-29, **19-25**

In Exercise 8-28, you analyzed data from the 2008 General Social Survey, in which a random sample of 1324 adult Americans reported how many hours of television they watch per day. The following histogram and statistics were produced from the sample data:



$n = 1324$ $\bar{x} = 2.982$ hours $s = 2.659$ hours

a. This histogram reveals that the sample data are strongly skewed to the right. Does this indicate that the technical conditions for a *t*-interval are not satisfied? Explain
b. Regardless of your answer to part a, calculate and interpret a 95% confidence interval.
c. Is the proportion of sample data that fall within this interval close to 95%? Should it be? Explain.

Exercise 19-26: Facebook Friends

A college student was interested in estimating the average number of Facebook friends in the population of all students with Facebook accounts at her university. Taking a random sample from this population would have been extremely difficult, so she selected a convenience sample of 20 of her friends. Her sample data are displayed in Figure Ex. 19-26 and summarized with the following statistics:

Variable	Number of Friends
N	20
Mean	391.8
StDev	273.8
Minimum	39.0
Q1	176.0
Median	327.0
Q3	639.0
Maximum	947.0

a. State the technical conditions required for this confidence interval procedure to be valid, and comment on whether you consider them to be satisfied in this case.
b. Regardless of your answer to part a, determine a 90% confidence interval for the population mean μ.
c. Write a sentence interpreting this interval in context. Be sure to include a clear description of what μ represents here.
d. Would you expect 90% of student Facebook users at her university would have values (for number of Facebook friends) within this interval? Explain.

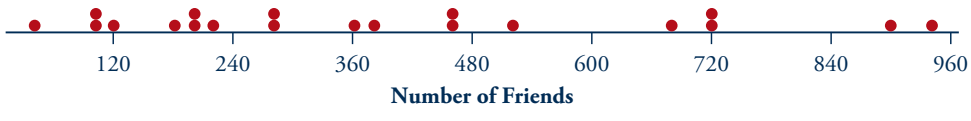


Figure Ex. 19-26

Exercise 19-27: Hockey Goals

Sample data for the total number of goals scored in the 43 National Hockey League (NHL) games played between Wednesday, November 25, and Sunday, November 29, 2009, were recorded. The intent was that this sample would be representative of the population of all NHL games in the 2009–2010 season, although the sample games were all chosen from a 5-day period early in the season. The following computer output was obtained from the sample data:

Variable	goals
N	43
Mean	5.558
StDev	2.472
SE Mean	0.377
95% CI	(4.797, 6.319)

a. Show how the standard error of the sample mean (with value 0.377) could have been calculated from other values provided here.
b. Interpret the confidence interval provided here: You're 95% confident of what?
c. If you had calculated this confidence interval by hand, what value of t^* would you have used?
d. The distribution of goals scored in this sample is noticeably skewed to the right. Does this call into question the validity of this confidence interval? Explain briefly.
e. For only 10 of the 43 games in the sample is the number of goals scored within this interval. Does this call into question the validity of this confidence interval? Explain briefly.

Exercise 19-28: Birth Weights

19-28, 19-29, 20-23, 20-24, 20-25, 29-24

The Odum Institute for Research in Social Science at the University of North Carolina made available data on all births that occur in the state of North Carolina. John Holcomb of Cleveland State University selected a random sample of 500 births in the year 2005; these data appear in the file `NCbirths`.

a. Examine graphical displays and descriptive statistics of the birth weights of these 500 babies. Write a paragraph summarizing the distribution of birth weights as revealed by these graphs and statistics.
b. Are the technical conditions for a *t*-confidence interval satisfied? Explain
c. Produce and interpret a 95% confidence interval for the population mean.

Exercise 19-29: Birth Weights

19-28, 19-29, 20-23, 20-24, 20-25, 29-24

Reconsider the previous exercise. Choose one of the other quantitative variables in that datafile: *mother's age, father's age, completed weeks of gestation, number of prenatal doctor visits, weight gained by mother.* For the variable that you choose, answer questions a, b, and c from the previous exercise.

Exercise 19-30: Christmas Shopping

15-2, 15-7, 15-18, 19-30

Recall from Activity 15-2 that a random sample of 2597 adult Americans in November 2009 revealed a sample mean of $343.31 that was spent on Christmas shopping over the Thanksgiving weekend.

a. Is $343.31 a parameter or a statistic? What symbol is used for this?
b. Determine the value of t^* for a 95% confidence interval for the population mean amount spent on Christmas shopping over that weekend.
c. What further information about the sample data do you need to determine this confidence interval?
d. Determine the endpoints of this confidence interval if the sample standard deviation turned out to be $150. Then repeat, assuming that the sample standard deviation turned out to be $200.
e. Comment on how the two intervals in part d compare. [*Hint*: Be sure to mention the midpoints as well as margin-of-error.]

Exercise 19-31: M&M Consumption

19-1, 19-3, 19-31, 22-30, 22-31, 22-37

Reconsider Activities 19-1 and 19-3, in which you determined a 95% confidence interval for the population mean number of M&Ms that would be chosen among all undergraduates who might have participated in that study.

a. Would you expect 95% of the *sample* values to fall within the 95% confidence interval? Explain.
b. Use technology (`MMconsumption`) to recalculate the 95% confidence interval, and then count how many of the 20 sample values fall within this interval. Is the answer close to 95%? Do you want to rethink your answer to part a?
c. Would you expect 95% of the *population* values to fall within the 95% confidence interval? Explain.

Exercise 19-32: On Your Own

6-30, 12-22, 19-22, 21-28, 22-27, 26-23, 27-22

a. Think of a situation in which you would be interested in producing a confidence interval to estimate a population *proportion*. Describe precisely the observational units, population, and parameter involved. Also describe how you might select a sample from the population. [*Hint*: Be sure to think of a *categorical* variable so dealing with a proportion is sensible.]
b. Repeat part a for a population mean rather than a proportion. [*Hint*: Be sure to think of a *quantitative* variable.]

TOPIC 20

Tests of Significance: Means

20

The golden ratio is a famous number not only in mathematics but also in art, music, and architecture. It has been suggested since ancient times that rectangles for which the ratio of width to length equals this golden value are aesthetically pleasing. In this topic, you will investigate whether or not beaded artwork of Shoshoni Indians appears to adhere to this famous number. Other questions that you will investigate include these: Do Cal Poly students tend to follow recommended practices of carrying less than 10% of their bodyweight in their backpacks? Do college students tend to get the recommended eight hours of sleep per night? Do schoolchildren watch more than two hours of television per day on average?

Overview

In the last topic, you studied confidence intervals for a population mean. Now you will turn to the other major type of statistical inference procedure, tests of significance, and apply these procedures to *quantitative* data and, therefore, a population *mean.* This procedure will lead you to one of the most famous of statistical techniques: the *t*-test.

Preliminaries

1. Guess how many points are scored (by the two teams combined) in a typical professional basketball game. (Exercise 20-9)

2. If you are looking to find evidence that rule changes have increased basketball scoring from the previous season, do you think that a sample of one day's games or a sample of one week's games would be more informative? (Exercise 20-9)

Collect Data

1. Guess your instructor's age or recall from Topic 8. (Exercise 20-13)

2. Gather and record data on your classmates' guesses of your instructor's age. (Exercise 20-13)

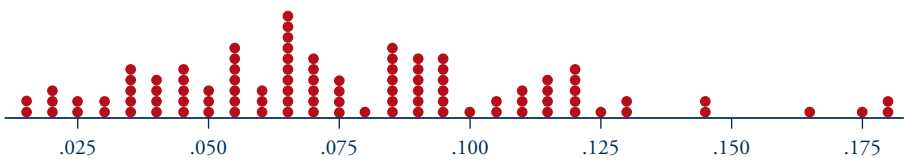
In-Class Activities

Activity 20-1: Backpack Weights 2-13, 10-12, 19-6, **20-1**, 20-17, Lab 7

Refer to Exercises 2-13, 10-12, and Activity 19-6, which described a study (Mintz, Mintz, Moore, and Schuh, 2002) in which student researchers recorded the body weights and backpack weights for a sample of 100 students on the Cal Poly campus (`Backpack`). The students wanted to see how well Cal Poly students adhered to recommendations to carry less than 10% of their body weight in their backpack.

In this activity, you will assess whether these data provide convincing evidence that the mean backpack-to-body weight ratio among all Cal Poly students is less than .10. You will follow the same six-step process for conducting a significance test that you learned in Topic 17.

Recall the summary statistics and dotplot of these ratios ($\bar{x} = 0.077$, $s = 0.0366$).



a. Do these summaries indicate that the sample data fall in the direction conjectured by these researchers?

b. *Step one*: Give a description of the parameter of interest in this study, being sure to identify the type of number, the variable, and the population. What symbol would you use to represent this parameter?

let M be the populatio mean backpack to bodyweght ratio during all cal poly students

c. *Step two*: State two competing claims about the parameter of interest. Keep in mind that the alternative hypothesis is what the researchers are hoping to show, and the null hypothesis is the "uninteresting" conjecture about the parameter.

Null hypothesis, H_0: mu = .1

Alternative hypothesis, H_a: mu<.1

How much do you carry?

d. *Step three*: What are the requirements for the Central Limit Theorem for a sample mean to be valid? Do you consider them met for this study? Explain.

r not random, hopefulyl representative???
n 100>30 so distribution means are normal
i 100*10 = 1000 < all cal poly students

e. Use the *standard error* of the sample mean to approximate the standard deviation of the above distribution. Calculate this value and interpret it in this context.

s/sqrtn = .00366

typical deviation the sample mean deviates from the population

20

f. *Step four*: Calculate the value of the test statistic. Using your answer to part e, determine how many standard errors the sample mean ($\bar{x} = 0.077$) falls from the hypothesized value ($\mu_0 = 0.10$).

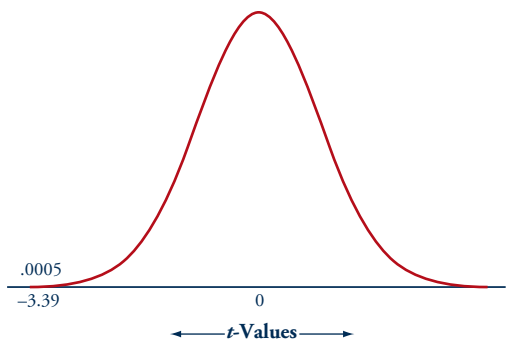
t=xbar-mu/(s/sqrtn) = -6.21

Because this calculation uses the standard *error* for the sample mean, based on the sample standard deviation s and not the population standard deviation σ, it is called a t-statistic. You calculate probabilities by comparing this test statistic value to the t-distribution with $n - 1$ degrees of freedom.

g. Draw a sketch of the t-distribution with 99 degrees of freedom. [*Hint:* It will look a lot like a normal distribution.] Mark the calculated value of the test statistic, and shade the area under the t-distribution to the left, in the direction of the alternative hypothesis, to represent the p-value.

tcdf(lower = -infty, upper = -6.28, df = 99) = 4.53107e-9

h. *Step five*: Calculate the p-value. Use the t-table to approximate the area to the left of your test statistic value. [*Note:* The table columns indicate the areas to the right of the given (positive) test statistic values. Due to the symmetry of the t-distribution, the area to the left of a negative t-value equals the area to the right of the corresponding positive t-value. You won't be able to find the (positive) test statistic value exactly in the table, but you can find two values on each side of the test statistic, or else find that the test statistic is off the chart. When the test statistic falls between two values, read off the probabilities from the top of the table that correspond to these two values. Then report that the p-value is between these two probabilities. If the test statistic value is off the chart, determine the bound on the p-value from the last probability listed. Your sketches should be very helpful.]



i. Use technology, possibly the `Test of Significance Calculator` applet as in Activity 17-3, to verify your calculations for this test and to calculate the *p*-value more exactly. [*Hint:* In the applet, use the pull-down menu to change the procedure from One Proportion to **One Mean**.]

Applet *p*-value:

tcdf(lower = -infty, upper = -6.28, df = 99) = 4.53107e-9

j. Interpret the *p*-value in the context of these data and hypotheses.

Assuming the mean backpack to backlift ratio among all uni students is .1, the chance of getting a sample mean f .077 or less is literally 0 lmfao

k. *Step six*: Make a test decision. Would you consider these data statistically significant at the .01 level? Explain.

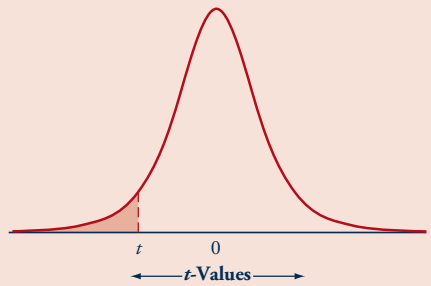
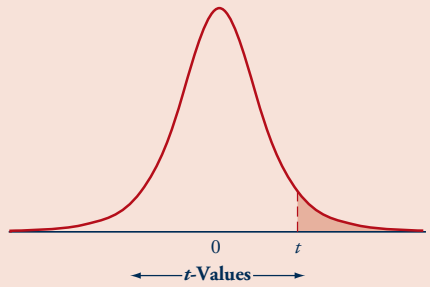
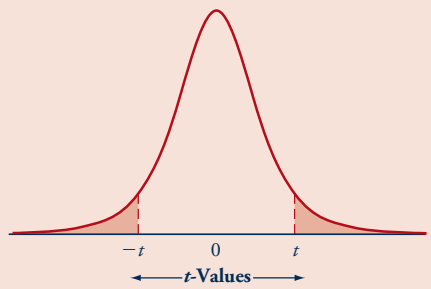
p<.01 and reject

l. *Step six continued:* Write a sentence or two summarizing your conclusion about whether or not the sample data provide convincing evidence that Cal Poly students' backpack-to-body weight ratios average less than the recommended .10. Include an explanation of the reasoning process showing that your conclusion follows from the test result.

we have very convincing evidence with p = literally fucking 0 to say that avg <.1

it most definetely is not

The following summarizes what you have done, **a test of significance for a population mean μ (*t*-test):**

1. Identify and define the population parameter of interest.	μ
2. State the null and alternative hypotheses based on the study question.	$H_0: \mu = \mu_0$ $H_a: \mu < \mu_0$ or $H_a: \mu > \mu_0$ or $H_a: \mu \neq \mu_0$
3. Check whether or not the technical conditions required for the procedure to be valid are satisfied.	• Simple random sample from population of interest • Either the population follows a normal distribution or the sample size is large ($n \geq 30$ as a guideline).
4. Calculate the test statistic, which measures the distance between the observed sample statistic and the hypothesized value of the parameter.	$t = \dfrac{\bar{x} - \mu_0}{s/\sqrt{n}}$
5. Calculate the *p*-value, which is the probability of obtaining such an extreme test statistic value when the null hypothesis is true. It is often very helpful to include a sketch of the sampling distribution, shading in the appropriate *p*-value area.	*p*-value = $\Pr(T_{n-1} \leq t)$  or $\Pr(T_{n-1} \geq t)$  or $2 \times \Pr(T_{n-1} \geq \lvert t \rvert)$  depending on the form of the alternative hypothesis.
6. Summarize your conclusion in context based on the magnitude of the *p*-value, including a test decision if a significance level α is provided	

Watch Out

- Remember hypotheses are always about *parameters,* not statistics. In this case, the relevant parameter is a population mean, denoted by μ, because the variable (*backpack-to-body weight ratio*) is quantitative.
- The alternative hypothesis should always be formulated before collecting the sample data, based on the research question.
- Note that the denominator of the test statistic is the standard error of the sample mean, which you also encountered in Topic 19 when forming a confidence interval for μ. Do not forget the $\sqrt{n}$ term in this expression.
- When calculating the *p*-value for a two-sided alternative, include the total area in *both* tails of the *t*-distribution beyond the value of the test statistic. But take advantage of the symmetry of the *t*-distribution and calculate this total area by doubling the area in the right tail.
- Think through the reasoning process of tests and *p*-values as you learned in Topic 17. A small *p*-value indicates that you are unlikely to obtain such extreme sample data if the null hypothesis is true, which provides evidence against the null hypothesis and in favor of the alternative hypothesis. The smaller the *p*-value, the stronger the evidence against the null hypothesis.
- Also keep in mind the duality between two-sided tests and confidence intervals (this time the standard error is the same as well).

Activity 20-2: Sleeping Times 8-29, 19-4, 19-5, 19-12, 19-19, 20-2, 20-7, Lab 5

Reconsider the data that you collected and analyzed on the sleep times for yourself and your classmates in Activity 19-5.

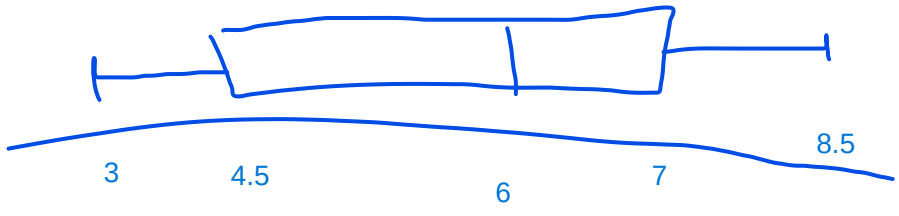
a. Conduct a significance test of whether or not your class data provide strong evidence that the mean sleep time among all students at your school that particular night differs from 7 hours. Clearly define the parameter of interest and report the hypotheses, test statistic, and *p*-value. Would you reject the null hypothesis at the .05 level? [*Hint:* You may need to use the symmetry of the *t*-distribution in finding the *p*-value.]

let mu equal the population mean of the number of hours that AP students sleeps

H0: mu = 7
Ha: mu =/= 7

b. Do you think the technical conditions for the validity of this procedure have been met? Explain.

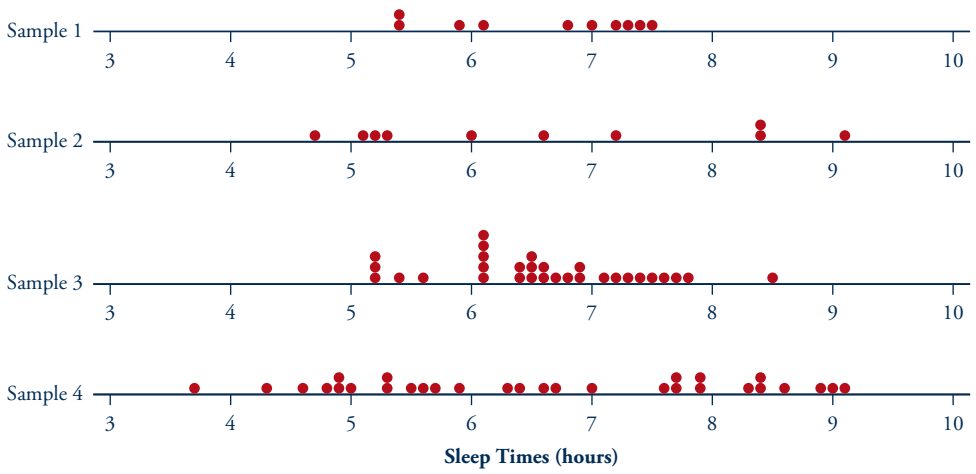
R random not quite because its in class but shuld be representative
N we look at the boxplot, symmetric and no boxplot and suggests normal popiulation distribution
I more than 170<number of AP students at NASH



test statistic is xbar-mu all over s/sqrtn

which is .02

Now consider the four samples of hypothetical sleep times presented in Activity 19-4. Dotplots and summary statistics are reproduced here. (The data are stored in the file `HypoSleep`.)



interval did not capture .07

Sample Number	Sample Size	Sample Mean	Sample SD	Test Statistic	*p*-value
1	10	6.6	0.825		
2	10	6.6	1.597		
3	30	6.6	0.825		
4	30	6.6	1.597		

when t increases p dec inverse relationship

c. Comparing samples 1 and 2, which sample do you think supplies stronger evidence that $\mu \neq 7$ (i.e., that the population mean sleep time differs from 7 hours). In other words, which sample (1 or 2) would produce a smaller *p*-value for the appropriate test of significance? Explain.

d. Comparing samples 1 and 3, which sample do you think supplies stronger evidence that $\mu \neq 7$ (i.e., that the population mean sleep time differs from 7 hours)? In other words, which sample (1 or 3) would produce a smaller *p*-value for the appropriate test of significance? Explain.

e. For each of these four samples, use technology to calculate the test statistic and *p*-value for testing that the population mean differs from 7 hours. Record these in the table following part b.

f. Which of the samples give(s) you enough evidence to reject the null hypothesis at the .05 level and conclude that the mean sleep time is, in fact, different from seven hours?

in conclusion, have the evidence to say that the mean sleep time of all NAsh students are likely not 7

g. Comment on whether or not your conjectures in part c and part d are confirmed by the test results.

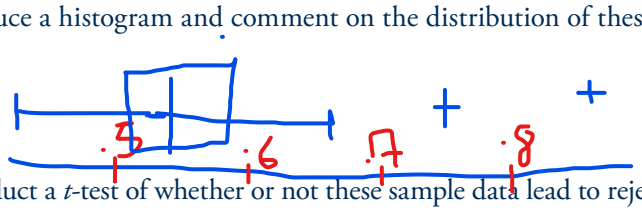
This activity should reinforce what you discovered in Activities 17-4 and 19-4 about effects of sample size and sample variation. An observed difference between a sample mean and a hypothesized mean is more statistically significant (i.e., less likely to occur by chance, indicated by a smaller *p*-value) with a larger sample than with a smaller one. Also, that difference is more significant if the sample data values themselves are less spread out (less variable, more consistent) than if they are more spread out. Naturally, the farther the sample mean is from the hypothesized mean, the more statistically significant the result.

Activity 20-3: Golden Ratio

The ancient Greeks made extensive use of the so-called golden ratio in art and literature, for they believed that a width-to-length ratio of 0.618 was aesthetically pleasing. Some have conjectured that American Indians also used the same ratio (Hand et al., 1993). The following data are width-to-length ratios for a sample of 20 beaded rectangles used by the Shoshoni Indians to decorate their leather goods.

0.693	0.662	0.690	0.606	0.570	0.749	0.672	0.628	0.609	0.844
0.654	0.615	0.668	0.601	0.576	0.670	0.606	0.611	0.553	0.933

a. Produce a histogram and comment on the distribution of these ratios.



b. Conduct a *t*-test of whether or not these sample data lead to rejecting the hypothesis that the population mean width-to-length ratio equals .618. Use the .10 significance level and a two-sided alternative. Report all aspects of the test, including a check of technical conditions, and summarize your conclusion in context.

0 mu=.618
a mu=/=.618
mu is population mean width to length ratio of shashowne indian beaded rectangles

r no but hopefully representative
n sampling distribution of xbar normal? no so use caution on results since outliers suggest pop not normal
i safe to assume 200 more lol

t=(x-mu)/(2/sqrtn) = 2.055 and p=.054. since its <.1 we reject H0 so "significant"
we have enough evidence with p=.054<.1 that the ratio differs from indians

Self-Check
Activity 20-4: Children's Television Viewing 1-15, 19-16, 20-4, 22-14, 22-15

Exercise 1-15 described a study on children's television viewing conducted by Stanford researchers (Robinson, 1999). At the beginning of the study, parents of third- and

fourth-grade students at two public elementary schools in San Jose were asked to report how many hours of television the child watched in a typical week. The 198 responses had a mean of 15.41 hours and a standard deviation of 14.16 hours.

Conduct a test of whether or not these sample data provide evidence at the .05 level for concluding that third- and fourth-grade children watch an average of more than two hours of television per day. Include all the components of a significance test, and explain what each component reveals. Start by identifying the observational units, variable, sample, and population.

Solution

The observational units are third- and fourth-grade students. The sample consists of the 198 students at two schools in San Jose. The population could be considered all American third- and fourth-graders, but it might be more reasonable to restrict the population to be all third- and fourth-graders in the San Jose area at the time the study was conducted.

The variable measured here is the *amount of television the student watches in a typical week,* which is quantitative. The parameter is the mean number of hours of television watched per week among the population of all third- and fourth-graders. This population mean is denoted by μ. The question asked about watching an average of two hours of television per day, so convert that to be 14 hours per week.

The null hypothesis is that third- and fourth-graders in the population watch an average of 14 hours of television per week (H_0: $\mu = 14$). The alternative hypothesis is that these children watch more than 14 hours of television per week on average (H_a: $\mu > 14$).

You should check the technical conditions for the *t*-test before you proceed:

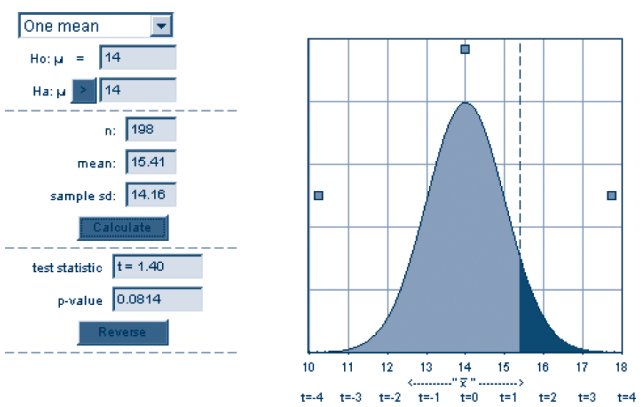
- The sample of children was not chosen randomly; they all came from two schools in San Jose. You might still consider these children to be representative of third- and fourth-graders in San Jose, but you might not be willing to generalize to a broader population.
- The sample size is large enough (198 is far greater than 30) that the second condition holds regardless of whether or not the data on television watching follow a normal distribution. You do not have access to the child-by-child data in this case, so you cannot examine graphical displays; however, the large sample size assures you that the *t*-test is nevertheless valid to employ.

The sample size is $n = 198$; the sample mean is $\bar{x} = 15.41$ hours; and the sample standard deviation is $s = 14.16$ hours. The test statistic is

$$t = \frac{15.41 - 14}{14.16/\sqrt{198}} \approx 1.401$$

indicating the observed sample mean lies 1.401 standard errors above the conjectured value for the population mean. Looking in Table III, using the 100 df line (rounded down from

the actual df of 198 − 1 = 197), reveals the *p*-value (probability to the right of $t = 1.401$) to be between .05 and .10. Technology calculates the *p*-value more exactly to be .081.



This *p*-value is not less than the .05 significance level. The sample data, therefore, do not provide sufficient evidence to conclude that the population mean is greater than 14 hours of television watching per week. This conclusion stems from realizing that obtaining a sample mean of 15.41 hours or greater would not be terribly uncommon when the population mean is really 14 hours per week. If you had used a greater significance level (such as .10), which requires less compelling evidence to reject a hypothesis, then you would have concluded that the population mean exceeds 14 hours per week.

Watch Out

- The first step is to decide whether your test is about a proportion or a mean. One way to judge is to ask whether the variable (as measured on the individual observational units) is categorical or quantitative. Also be on the lookout for more obvious clues, such as the word *average* appearing in the statement of this problem.
- Remember the hypotheses are about parameters, not statistics. Also remember the hypothesized value comes from the research question, not from the sample data. Also, the direction of the alternative hypothesis should be based on the research question and not on the data.
- When a sample is not chosen randomly, be cautious about generalizing the study results to a larger population. Try instead to think about what population the sample might be representative of, and even then be wary.
- Do not forget to include the $\sqrt{n}$ factor in calculating the value of the test statistic.

Wrap-Up

This topic introduced you to tests of significance concerning a population *mean.* You found that the structure, reasoning, and interpretation of these tests are the same for a mean as they were for a proportion. As always, the *p*-value is the key, because it reports how likely you would have been to observe such extreme sample data if the null hypothesis were true. So, a small *p*-value provides evidence against the null hypothesis that the population mean equals a particular hypothesized value.

Two ways in which the ***t*-test** for a mean is different from the *z*-test for a proportion are that the *p*-value is calculated from the *t*-distribution instead of the normal distribution, and the variability in the sample data also plays a role in determining the *p*-value. For example, learning that the average sleep time in a sample of students is 6.6 hours is not enough information to assess whether or not the population average sleep time is less than 7 hours, because you also need to know the sample size and how variable (as measured by the sample standard deviation) those sleep times are.

In Brief

20

Some useful definitions to remember and habits to develop from this topic are

- Specify the null and alternative hypotheses based on the research question, prior to seeing the sample data. With a quantitative variable, the hypotheses concern a population mean.
- Check the technical conditions before applying the procedure. As with a *t*-interval, the *t*-test requires either a large sample size or a normally distributed population. You can generally regard a sample of at least 30 as large enough for the procedure to be valid. If the sample size is less than 30, examine visual displays of the sample data to see whether they appear to follow a normal distribution.
- Consider whether or not the sample was randomly selected from the population of interest before generalizing the test conclusion to that population.
- Do not forget to begin your analysis with graphical and numerical summaries of the sample data. Do not jump into a *t*-test before examining the sample data first.

You should be able to

- Conduct all aspects of a *t*-test of significance concerning a population mean, from stating hypotheses and checking conditions, to calculating a test statistic and *p*-value, and drawing conclusions. (Activities 20-1, 20-2, 20-3)
- Check conditions for whether a one-sample *t*-test is valid to apply. (Activities 20-1, 20-3)
- Explain the impact of sample size and sample variability on all components of a test of significance. (Activity 20-2)

Thus far you have studied confidence intervals and tests of significance for both categorical and quantitative variables, but you have only considered cases with a single sample. In the next unit, you discover how to apply these inference procedures to research questions that call for comparing two groups. These procedures will be especially useful for analyzing data from randomized experiments.

Exercises

Exercise 20-5: Exploring the *t*-Distribution 19-2, **20-5,** 20-6

a. Use the *t*-table to find the *p*-values (as accurately as possible) corresponding to the following test statistic values and degrees of freedom, filling in a table like the one shown here with those *p*-values:

df	Pr($T \geq 1.415$)	Pr($T \geq 1.960$)	Pr($T \geq 2.517$)	Pr($T \geq 3.168$)
4				
11				
23				
80				
Infinity				

b. Does the *p*-value get larger or smaller as the value of the test statistic increases (if the number of degrees of freedom remains the same)?

c. Does the *p*-value get larger or smaller as the number of degrees of freedom increases (if the value of the test statistic remains the same)?

Exercise 20-6: Exploring the *t*-Distribution 19-2, 20-5, **20-6**

a. Use the *t*-table and/or your answers to Exercise 20-5 to find the following *p*-values:

df	Pr($T \leq -1.415$)	Pr($T \leq -1.960$)	$2 \times$ Pr($T \geq \lvert -2.517 \rvert$)	$2 \times$ Pr($T \geq \lvert -3.168 \rvert$)
4				
11				
23				
80				
Infinity				

b. Describe how the *p*-values in the first two columns of the table compare with those in the first two columns of the table from Exercise 20-5. Explain why this makes sense.

c. Describe how the *p*-values in the last two columns of the table compare with those of the last two columns of the table from Exercise 20-5. Explain why this makes sense.

Exercise 20-7: Sleeping Times

8-29, 19-4, 19-5, 19-12, 19-19, 20-2, **20-7,** Lab 5

Reconsider the hypothetical samples of sleep times (`HypoSleep`) presented in Activity 19-4 and analyzed in Activity 20-2. Suppose now that you are interested in testing whether or not the sample data provide evidence that the population mean sleep time is *less than* seven hours. Use technology to conduct the test of significance for each of the four samples with a *one-sided* alternative hypothesis. Record the *p*-value for each sample and comment on how these *p*-values compare with the two-sided values found in Activity 20-2.

Exercise 20-8: UFO Sighters' Personalities **20-8,** 22-17

In a 1993 study, researchers took a sample of people who claimed to have had an intense experience with an unidentified flying object (UFO) and a sample of people who did not claim to have had such an experience (Spanos et al., 1993). They then compared the two groups on a wide variety of variables, including IQ. Suppose you want to test whether or not the average IQ of those who have had such a UFO experience is higher than 100, so you want to test H_0: $\mu = 100$ vs. H_a: $\mu > 100$.

a. Identify clearly what the symbol μ represents in this context.

b. Is this a one-sided or a two-sided test? Explain how you can tell.

The sample mean IQ of the 25 people in the study who claimed to have had an intense experience with a UFO was 101.6; the standard deviation of these IQs was 8.9.

c. Does this information enable you to check the technical conditions completely? What needs to be true for this procedure to be valid?

d. Calculate the test statistic and draw a sketch with shaded area corresponding to obtaining a test statistic as extreme or more extreme than this one observed for the sample of 25 UFO observers.

e. Use the *t*-table or technology to determine (as accurately as possible) the *p*-value.

f. Write a sentence interpreting the *p*-value in the context of this sample and these hypotheses. Summarize the conclusion of your test in context.

Exercise 20-9: Basketball Scoring 20-9, 20-20

Prior to the 1999–2000 season in the National Basketball Association, the league made several rule changes designed to increase scoring. The average number of points scored per game in the previous season had been 183.2. Let μ denote the mean number of points per game in the 1999–2000 NBA season.

a. If the rule changes had *no effect* on scoring, what value would μ have? Is this a null or an alternative hypothesis?
b. If the rule changes had the desired effect on scoring, what would be true about the value of μ? Is this a null or an alternative hypothesis?

The following sample data are the number of points scored in the 25 NBA games played during December 10–12, 1999 (`NBAPoints`):

196	198	205	163	184	224	206	190	140	204	200	190	195
180	200	180	198	243	235	200	188	197	191	194	196	

c. Use technology to examine visual displays of this distribution. Write a few sentences commenting on these points per game, particularly addressing the issue of whether scoring seems to have increased over the previous season's mean of 183.2 points per game.
d. Use technology to calculate the mean and standard deviation of this sample of points per game. Use appropriate symbols to denote these values.
e. Comment on whether the technical conditions for the validity of the *t*-test have been satisfied.
f. Use the sample statistics you found in part d to compute the value of the *t*-test statistic from this sample.
g. Use the *t*-table or technology to approximate the *p*-value for this study.
h. Interpret the *p*-value in the context of these data and hypotheses.
i. Would you reject the null hypothesis at the .10 level? At the .05 level? At the .01 level? At the .005 level? Explain.
j. Write a sentence or two summarizing your conclusion about whether the sample data provide evidence that the mean points per game in the 1999–2000 season is higher than in the previous season. Include an explanation of the reasoning process by which your conclusion follows from the test result. Also mention any concerns you have about the technical conditions for this study.

Exercise 20-10: Credit Card Usage

1-9, 16-12, 19-23, **20-10**

Refer to the study described in Exercises 16-12 and 19-23. The Nellie Mae organization found that in a random sample of undergraduate students taken in 2004, the average credit card balance was $2169. Again take the sample size to be 1074, which corresponds to the 76% of the sample of 1413 students who hold a credit card. Suppose (for now) the sample standard deviation of these credit card balances is $1000.

a. Conduct a significance test of whether or not the sample data provide strong evidence (at the $\alpha = .05$ level) that the population mean credit card balance exceeds $2000. Report all aspects of the test, including a check of the technical conditions.
b. Repeat part a, assuming that the sample standard deviation of these credit card balances is $2000.
c. Which scenario produces a greater *p*-value? Explain why this makes sense.

Exercise 20-11: Body Temperatures 12-1, 12-19, 14-3, 14-18, 15-9, 19-7, 19-8, **20-11**, 22-10, 23-3

Reconsider the data from Exercise 19-7 on body temperatures for a sample of 65 healthy adult males and 65 healthy adult females (`BodyTemps`). For each gender, use technology to produce graphical and numerical summaries of the body temperatures. Then conduct a *t*-test of whether or not these sample data provide strong evidence that the population mean body temperature differs from 98.6 degrees. Provide all components of a hypothesis test, and summarize your conclusions. Also comment on how your conclusions are similar or different between the two genders.

Exercise 20-12: Social Acquaintances

9-8, 9-9, 10-13, 10-14, 19-9, 19-10, **20-12**

Reconsider the data that you collected in Topic 9, based on the social-acquaintance exercise described by Malcolm Gladwell in *The Tipping Point.*

a. Use the sample data from your class to test whether or not the population mean number of acquaintances differs from 30. Report the hypotheses, test statistic, and *p*-value. Also state the test decision using the $\alpha = .025$ significance level, and summarize your conclusion.
b. To what population would you be willing to generalize the result of your significance test in part a? Explain.

Exercise 20-13: Age Guesses

8-20, 20-13, 20-14

Consider the data collected in the Collect Data section concerning guesses of your instructor's age. Conduct a full analysis of the data with regard to the question of whether or not the guesses tend to average out to the actual age. (If your instructor prefers not to reveal his or her actual age, address whether or not the guesses tend to average out to the age that you personally guessed.) Include graphical and numerical summaries as well as an appropriate test of significance, and be sure to identify the population and parameter of interest very clearly. Write a paragraph or two describing and explaining your analysis and findings.

Exercise 20-14: Age Guesses

8-20, 20-13, 20-14

After 5 weeks of a 10-week term, a 44-year-old statistics instructor asked his students to guess his age. The responses are displayed in the dotplot in Figure Ex. 20-14.

a. Comment on the distribution of age guesses. Be sure to address whether students tended to misjudge his age on the high side or low side or whether their guesses averaged out to approximately the correct age. [*Hint:* Remember to mention center, spread, shape, and outliers.]
b. The sample size is 44; the sample mean is 41.182 years; and the sample standard deviation is 3.996 years. Are these parameters or statistics? Identify the appropriate symbol for each of these numbers.
c. Identify the relevant population and parameter of interest.
d. Conduct a significance test of whether or not these sample data provide evidence that the population mean age guess differs from the instructor's actual age of 44 years. Include all components of a test, including a check of technical conditions. State the test decision that you would make at the $\alpha = .01$ significance level, and summarize your conclusion in context.

Exercise 20-15: Nicotine Lozenge

1-16, 2-18, 5-6, 9-21, 19-11, 20-15, 20-19, 21-31, 22-8

Recall from Exercises 1-16 and 19-11 the experiment that investigated the effectiveness of a nicotine lozenge for subjects who wanted to quit smoking (Shiffman et al., 2002). Before the treatments began, subjects answered background questions, including how many cigarettes they smoked per day. Among the 1818 subjects in the study, the average was 22.0 cigarettes per day, and the standard deviation was 10.8 cigarettes per day.

a. Use these sample statistics to test whether or not the population mean differs from 20 cigarettes (one pack) per day.
b. Suppose that a smaller study produced the same sample mean and standard deviation, but with a sample size of only 100 subjects. Conduct the test from part a in this case.
c. Describe how your test results differ between part a and part b, and explain why this makes sense.

Exercise 20-16: Random Babies

11-1, 11-2, 14-8, 16-17, 20-16, 20-22

Recall the simulated data that you collected by shuffling and dealing cards to represent assigning babies to mothers at random in Activity 11-1. Consider the variable *number of matches.*

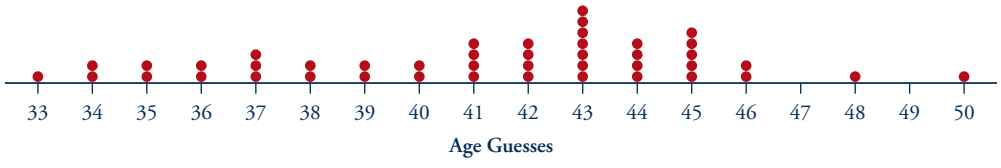


Figure Ex. 20-14

a. Report (by either recalling or recalculating) the sample size (number of repetitions conducted by yourself and your classmates), the sample mean number of matches per repetition, and the sample standard deviation of those numbers of matches.
b. Recall that your theoretical analysis in part g of Activity 11-2 revealed the long-run average number of matches to be exactly 1. Use the sample statistics from part a to test whether or not the simulated data provide strong reason to doubt that the population mean equals 1. Report the hypotheses (in words and in symbols), the sampling distribution specified by the null hypothesis, the test statistic, and the p-value. Also comment on the technical conditions and write a conclusion.

Exercise 20-17: Backpack Weights

2-13, 10-12, 19-6, 20-1, **20-17,** Lab 7

Recall the data on backpack weights and body weights of college students from Activity 19-6 (`Backpack`). Consider the ratio of backpack weight to body weight.

a. Is this ratio a categorical or quantitative variable?
b. Conduct a test of whether or not the sample data provide evidence that the population mean ratio differs from .10. Report all aspects of the test, including a check of technical conditions. Also summarize your conclusion.
c. Comment on whether or not your test result is consistent with the findings of the confidence interval found in Activity 19-6.

Exercise 20-18: Looking Up to CEOs

15-4, **20-18**

Recall from Activity 15-4 the study that took a random sample of male chief executive officers (CEOs) of American companies to test whether or not their average height exceeds the 69-inch average height for adult American males (Gladwell, 2005).

a. State the null and alternative hypotheses for this study.
b. Describe what a Type I error would mean in this context.
c. Describe what a Type II error would mean in this context.

Exercise 20-19: Nicotine Lozenge

1-16, 2-18, 5-6, 9-21, 19-11, 20-15, **20-19,** 21-31, 22-8

Recall the data that you analyzed in Exercise 19-11 on the number of cigarettes smoked per day by subjects in an experiment that investigated the effectiveness of a nicotine lozenge. Among the 1818 subjects in the study, the average was 22.0 cigarettes per day, and the standard deviation was 10.8 cigarettes per day.

a. Based on the 99% confidence interval for the population mean μ, what potential values of μ would be rejected at the .01 significance level? Explain.
b. Conduct a test of whether or not the population mean differs from 22.0 at the .05 significance level.
c. Based on the p-value in part b, what can you say regarding a 95% confidence interval for μ? Explain.

20

Exercise 20-20: Basketball Scoring

20-9, **20-20**

Reconsider the data that you analyzed in Exercise 20-9 concerning whether or not the mean points scored per NBA game was higher in the 1999–2000 season than it had been in the previous season. The November 29, 1999, issue of *Sports Illustrated* reported that for the first 149 games of the 1999–2000 season, the mean number of points per game was 196.2.

a. State in words and in symbols the hypotheses for testing whether or not the sample data provide strong evidence that the mean for the entire 1999–2000 season exceeds 183.2.
b. Do you have enough information to calculate the test statistic? Explain.
c. How large would the test statistic have to be to reject the null hypothesis at the .01 level?
d. If the sample standard deviation for these 149 games were close to the standard deviation for the 25 games that you analyzed in Activity 20-9, would the test statistic exceed the rejection value in part c? By a lot? Explain.
e. Even though the magazine did not provide all the information necessary to conduct a significance test, can you reasonably predict whether or not the test result would be significant at the .01 level? Explain, based on your answers to parts c and d.

f. Does the validity of this test procedure depend on the scores being normally distributed? Explain.
g. Discuss one advantage and one disadvantage to using this sample rather than the one used in Exercise 20-9.

Exercise 20-21: Pet Ownership

13-9, 15-14, 15-15, 18-2, 20-21

Recall from Activity 18-2 that in a survey of 80,000 households conducted by the American Veterinary Medical Association in 2001, 31.6% of households reported that they owned a pet cat. Of those 25,280 households that did own a pet cat, the mean number of cats per household was 2.1.

a. State (in words and in symbols) the hypotheses for testing whether or not the sample data provide strong evidence that the mean number of cats per cat-owning household exceeds 2.
b. What additional sample information do you need to conduct this test?
c. Supply a reasonable estimate for the missing sample statistic, and calculate the test statistic and p-value. Does the sample provide strong evidence that the mean number of cats exceeds two? Explain.
d. As a check on the sensitivity of this test, double your estimate for the missing sample statistic and repeat part c. Does your conclusion change substantially? Explain.
e. Using your estimate from part c, find a 99% confidence interval for the population mean number of cats per cat-owning household. Would you say that this mean value greatly exceeds two cats in a practical sense? Explain.

Exercise 20-22: Random Babies

11-1, 11-2, 14-8, 16-17, 20-16, 20-22

Recall again from Activity 11-1 and Activity 11-2 the class simulation of the random-babies activity. Let π be the long-term proportion of repetitions that result in no matches, and let μ be the long-term mean number of matches per repetition.

a. Using the simulated sample data, conduct a significance test of whether or not π differs from .4. Report the hypotheses, test statistic, and p-value. Is the result significant at the $\alpha = .05$ level?
b. If the test result in part a is not significant, does it follow that you accept π equals .4 exactly?
c. Recall from your theoretical analysis in Activity 11-2 the exact value of π. Explain how this relates to part b.
d. Using the simulated sample data, conduct a significance test of whether or not μ differs from 1.1. Report the hypotheses, test statistic, and p-value. Is the result significant at the $\alpha = .05$ level?
e. If the test result in part d is not significant, does it follow that you accept μ equals 1.1 exactly?
f. Recall from your theoretical analysis in Activity 11-2 the exact value of μ. Explain how this relates to part e.

Exercise 20-23: Birth Weights

19-28, 19-29, 20-23, 20-24, 20-25, 29-24

Recall from Exercise 19-28 the sample of 500 births in the state of North Carolina in the year 2005 (`NCbirths`). We said in Activity 12-2 that birth weights can be modeled with a normal distribution with mean 3300 grams, or about 7.275 pounds. Use these sample data to conduct a significance test of whether the population mean birth weight in the state of North Carolina in 2005 differs from 7.275 pounds. Be sure to follow and report all six steps of this test procedure.

Exercise 20-24: Birth Weights

19-28, 19-29, 20-23, 20-24, 20-25, 29-24

Following up on the previous exercise, consider testing whether the mean age among all mothers who gave birth in North Carolina in 2005 differs from 27.4 years, the mean age among all mothers in the United States in 2005 (*National Vital Statistics Reports*). The value of the t-test statistic turns out to be: $t = -1.82$.

a. Based on this test statistic, determine the p-value as accurately as possible.
b. Summarize the conclusion that you would draw from this test.
c. Based on the test statistic and p-value, what can you say about a 95% confidence interval for the mean age among North Carolina mothers? Explain.

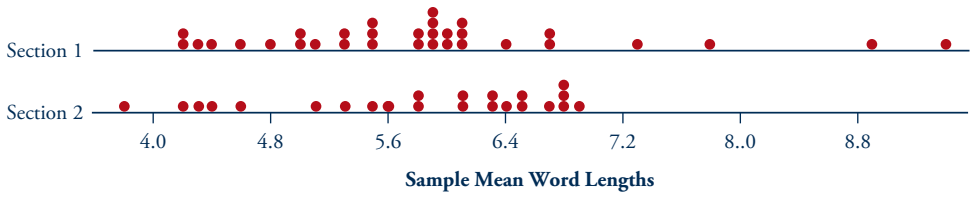


Figure Ex. 20-26

Exercise 20-25: Birth Weights

19-28, 19-29, 20-23, 20-24, 20-25, 29-24

Following up on the previous exercise, consider finding a confidence interval for the mean age among all *fathers* of newborn babies in North Carolina in 2005. Varying the confidence level, three confidence intervals turn out to be:

90%: (29.49, 30.56) 95%: (29.38, 30.66)
99%: (29.18, 30.86)

a. Based on these confidence intervals, what can you say about the *p*-value for a significance test of whether the population mean age among these fathers differs from 30 years? Explain.
b. Based on your answer to part a, summarize your conclusion from that test and *p*-value.

Exercise 20-26: Sampling Words

4-1, 4-2, 4-3, 4-4, 4-7, 4-8, 8-9, 9-15, 10-23, 14-6, 20-26

Recall from Exercise 10-23 that a statistics professor used the Gettysburg Address activity (Activity 4-1) with two sections of students. The dotplots in Figure Ex. 20-26 display the distributions of the average number of letters per word in the samples of size 10 that the students selected based on their own judgment, without using the random digit table:

Descriptive statistics for the sample mean word lengths in these two sections are:

Section	Sample Size	Sample Mean	Sample SD
1	32	5.856	1.210
2	23	5.765	0.956

a. Based on the sampling techniques used, did the professor believe his students' average would be larger or smaller than the population mean? Or did he have no prior suspicion?
b. For *each section*, conduct a significance test using the $\alpha = .01$ significance level of whether the population mean is greater than 4.29.
c. Recall that the actual population mean number of letters per word in the Gettysburg Address is 4.29. What does your answer to part b indicate about the sampling methods of these students? Explain.

20

Lab 5: Sleepless Nights

8-29, 19-4, 19-5, 19-12, 19-19, 20-2, 20-7, Lab 5

Goals

In this lab you will compare the amount of sleep obtained by students in your class to published recommendations for college students. You will apply tools you have learned to explore the data and then a simulation to compare your class results to a hypothesized sampling distribution. Then you will practice carrying out the six steps in a test of significance for a population mean.

Background

The following information is taken from the University of Michigan Wellness website (`http://www.uhs.umich.edu/sleep`, viewed April 4, 2011):

> **Night Owl Nation** College students, like Americans overall, are sleeping less. The college years are notoriously sleep-deprived due to all-night cram sessions, parties, TV, the net, and a general overload of activity. On average, college students today are going to bed 1–2 hours later and sleeping 1–1.6 hours less than they did a generation ago. As a result, sleep complaints and depression have increased dramatically among college students.
>
> **Why Do We Need Sleep?** Sleep maintains your circadian rhythms (the light-dependent 24-hour cycle that regulates body and mind), restores your body functions, and strengthens your immune system. It also helps you remember what you learn and prepares you for your next challenge.
>
> **How Much Sleep Do I Need?** Many adults function best with around 8 hours of sleep, but each person has unique needs. Sleep requirements depend on the environment, stress, health, age, and many other variables.
>
> But if you're like most college students, you're not getting enough sleep. On average, college students get only 6–6.9 hours of sleep per night.

Pre-lab Questions

Answer the following questions the best you can with your current knowledge. Then proceed to the online lab to analyze the data (using a Java applet) and to complete the lab report (using a word processing program).

a. For this study, identify the observational units and the response variable of interest.

b. Is the response variable quantitative or categorical?

c. What symbol will we use to refer to the average amount of sleep in our sample?

$\hat{p}$ π μ $\bar{x}$ σ s

d. What symbol will we use to refer to the average amount of sleep by all students at your school?

$\hat{p}$ π μ $\bar{x}$ σ s

e. What symbol will we use to refer to the standard deviation of the sleep amounts in our sample?

$\hat{p}$ π μ $\bar{x}$ σ s

f. What symbol will we use to refer to the standard deviation of the values (sleep amounts) in the population?

$\hat{p}$ π μ $\bar{x}$ σ s

g. If you consider the distribution of all sleep times that night in the population, what *shape* do you suspect this distribution will have? Briefly explain.

h. Do you suspect students at your school get more than 8 hours of sleep on a typical night or less than 8 hours? Or do you have no prior suspicion?

i. Do you suspect students at your school get more than 7 hours of sleep on a typical night or less than 7 hours? Or do you have no prior suspicion?

Now continue to the online lab where you will find instructions for completing the lab write-up.