Chapter 3

Inference with Proportions

This ActiveStats document contains a set of activities for Introduction to Statistics, MA 207 at Carroll College. This is a non-calculus based statistics class which serves many majors on campus. This document is intended for the classroom teacher to support students in active engagement with statistics on a daily basis. This document is not designed to be given to students as is. Rather, it is a teacher resource.

The activity set is designed to work alongside the OpenIntro *Introductory Statistics with Randomization and Simulation* textbook by Diez, Barr, and Cetinkaya-Rundel. The chapters in ActiveStats are numbered to align with OpenIntro, though the subsections may differ. OpenIntro is an open source curriculum with accompanying data sets. OpenIntro is the textbook resource to direct students to for out-of-class reading assignments and review. We also use the Cartoon Guide to Statistics as a supplement for assigned reading.

Data sets for ActiveStats can be found at mathquest.carroll.edu/activestats/data/ or on the class Moodle page.

3.1 Sampling distributions and the Central Limit Theorem

Reading Assignment 3.1.

- For this preview, re-read the jury problem from pages 137 to 142 of the Cartoon Guide to Statistics. Now, we are going to add a few technical details. Read pages 143 - 145 for the details. When you're done reading please answer the questions in the preview.

▲

Preview Activity 3.2. Hypothesis testing

1. Which of these describes the null hypothesis?
   (a) A statement that chance alone causes the variance that we see
   (b) A statement that tells us how likely our data is
   (c) A statement about what we think might be true
2. Which of these describes the alternative hypothesis?

(a) A statement that chance alone causes the variance that we see

(b) A statement that tells us how likely our data is

(c) A statement about what we think might be true

3. Which of these describes the meaning of the p-value?

(a) The probability of finding our data assuming that the null hypothesis is true

(b) The probability of finding our data

(c) The probability of finding the null hypothesis

(d) The probability of finding the alternative hypothesis assuming that the null hypothesis is true

---

Teacher Note: If your class is large enough, you can use your live class data for this introduction to hypothesis testing. If you have a smaller class (less than 40), use the provided data.

Example 3.3. Facial prototyping and reasoning with proportions
(adapted from a presentation by Dr. Allan Rossman)

Researchers investigating facial prototyping have asked whether people tend to associate names with faces for people they have not actually met. The following image is taken from a research paper by Lea, Thomas, Lamkin, & Bell (2007). Research participants were asked who is on the left, Bob or Tim?

Who is on the left, Bob or Tim? Image from Lea, Thomas, Lamkin, & Bell, 2007

Gather data from the class. Who is on the left, Bob or Tim?

In this scenario, our null hypothesis is that there is no such thing as facial prototyping. In statistical terms, we expect 50% of particpants to choose Tim on the left and 50% to choose Bob on the left. i.e.

$$H_0 : p = 0.5$$

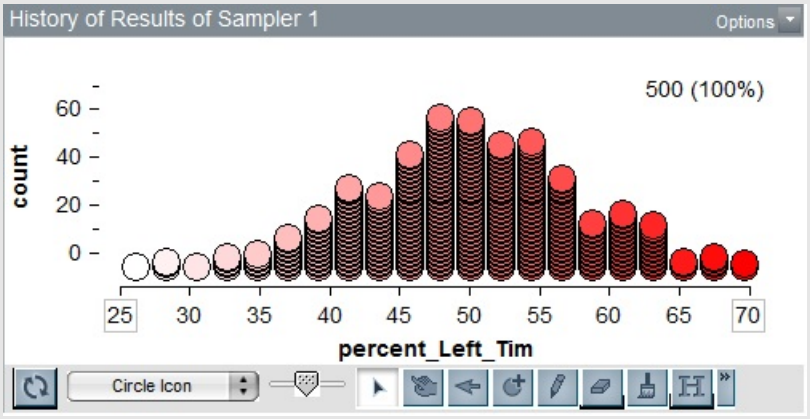
The alternate hypothesis is that facial prototyping is a genuine phenomenon and that a noticeable majority of the participants will agree on which face is on the left. i.e.

$$H_A : p \neq 0.5$$

Example: In a class of 46 students, 36 said that Tim was on the left and only 10 said that Bob was on the left. In other words, 78% of the class thought the left face would be Tim and only 22% thought it would be Bob.

- Is this evidence of facial prototyping?
- If there is no such thing as facial prototyping, how many votes for Tim on the left would you expect?
- Create a model of the situation if there is no such thing as facial prototyping. Build a sampling distribution in TinkerPlots.
- How rare is such an extreme sample result (if facial prototyping is not happening)?

Below is a sampling distribution for the null hypothesis that there is no such thing as facial prototyping ($p = 0.5$). The sampling distribution is built with 500 randomly generated samples of size 46 with $p = 0.50$. It is not unusual to get between 40% and 60% of participants selecting Tim on the left. However, not a single sample out of 500 samples was as extreme as the 78% (or 22%) we saw in the sample data. Therefore our p-value is less than $\frac{1}{500}$, i.e. p-value < 0.002.



Sampling distribution, with spinner at p=0.5 & 500 samples of size 46.

In other words, it would be *very* unusual to get a result as extreme as 78% of people spontaneously agreeing that Tim is on the left if people do not engage in facial prototyping. This provides evidence that the researchers can use to back up their theory that facial prototyping is a real phenomenon.

Now for a bit of theory

We've leveraged simulations and sampling distributions to get a sense for what is reasonable and what is extremely rare. Now we are going to make some generalizations about

sampling distributions and connect them to normal distributions. You'll notice that the sampling distribution from the prior activity looks rather bell shaped. In fact, it is approximately a normal distribution. We will describe how to find the mean and standard deviation for these very useful normal distributions.

The standard deviation for the sampling distribution is called the *standard error* or SE for the sampling distribution. The formula for the standard error will depend on the scenario. In this chapter we will examine scenarios based on proportions and in the next chapter we will focus on means.

Definition 3.4. Central Limit Theorem for Proportions

Given a population with a proportion p, the set of all possible samples of size n forms a sampling distribution that approaches a normal distribution with a mean of p and a standard error of $SE = \sqrt{\frac{p(1-p)}{n}}$. As the sample size, n, increases, the sampling distribution gets closer and closer to a normal distribution.

If we do not know the population proportion p, we can approximate the standard error using the sample proportion $\hat{p}$ instead of p. In this case $SE \approx \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$. This typically happens when generating a confidence interval.

How big does the sample size n need to be for this to be close enough to a normal distribution? What other conditions need to be met for us to assume the sampling distribution is a normal distribution? For proportion problems, the conditions are:

The Fine Print - (Conditions to check)

- The sample observations must be independent and generally not more than 10% of the overall population
- There should be at least 10 successes and 10 failures in our sample. i.e. If we ask a yes/no question, we should have at least 10 yes and 10 no responses in our sample. This is often written as $np \geq 10$ and $n(1-p) \geq 10$.

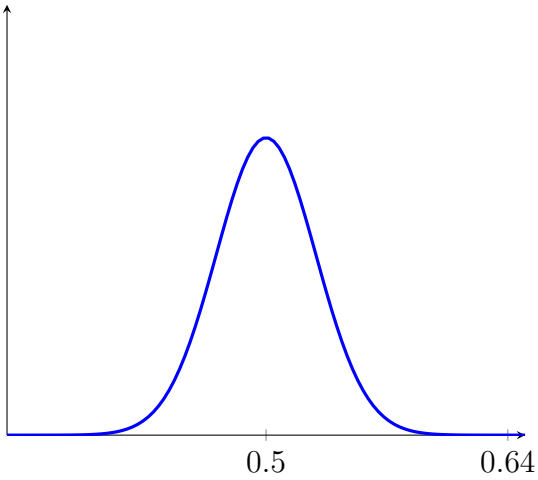
Activity 3.5. Continuing with the facial prototyping, with a larger n

When researchers notice a phenomenon with a small sample size, they often conduct a follow-up study with a larger sample size. Let's suppose our facial prototyping researchers decide to conduct a follow-up study with 300 participants. They use the same Tim and Bob images as the prior activity and find that 192 participants select Tim as the photo on the left.

Rather than building a simulation, this time we will use the formulas from the Central Limit Theorem to describe the bell curve that we would have created if we had done a simulation.

1. The null hypothesis is that facial prototyping is not a real phenomenon. State the null hypothesis in terms of a proportion of people who will select the first name.
2. State the alternative hypothesis

3. If the null hypothesis is true, what is the mean and standard error for the sampling distribution? Label this below.



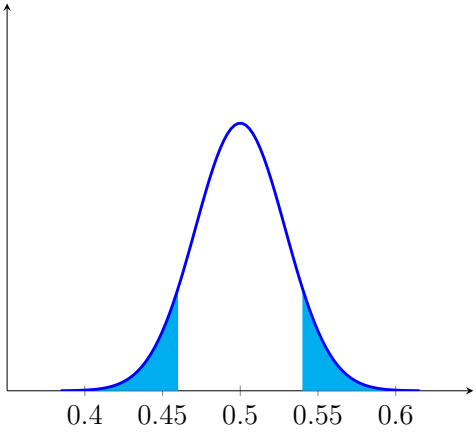
4. Use Excel to find the probability of a sample as extreme as the data above, given the mean and standard error you just found.
   Hint: Recall that the Excel command *norm.dist(x, mean, SE, 1)* gives you the area to the left of x. How do you find the area to the right?

5. Because we are interested in both extremely high and extremely low selection of the name Tim, we will want to find the probability of being in the extremes of either tail of our normal distribution. To compute this, double your result from the previous step to find the total p-value for this two-tailed test.

6. Two students wanted to try this out with their own names: Heather and Jennifer. They use their own pictures and find 300 volunteer participants. They find that 162 of the 300 participants assigned the left picture to Heather and the other 138 selected Jennifer. Why will the null hypothesis and alternative hypothesis be the same as before?

7. Will the mean and standard error change? If so, find the new mean and standard error. If not, explain why they will stay the same as the Bob and Tim example.

8. Use Excel to find the probability of a sample as extreme as 162 out of 300, i.e. $\hat{p} = 0.54$.



9. You calculated a much higher p-value this time than you did with the Tim and Bob version of the experiment. Why might this have happened? Which of the following sound reasonable?

   - Facial prototyping is actually not a real phenomenon and the researchers just got lucky with Tim and Bob.
   - The phenomonon might be real, but it only works for the names Tim and Bob.
   - The names Jennifer and Heather are associated with similar facial prototypes.
   - Other reasons or confounding variables. What is your explanation?

10. Suppose the researchers descide to classify two names as "significantly different profile types" if testing them provides a certain p-value. What p-value would you select as the cut-off? Explain your reasoning in complete sentences.

▲

Activity 3.6. A set of scenarios to practice finding the mean and standard error.

For each of the scenarios, determine whether the sampling distribution will be basically a normal distribution. If so, state the mean and standard error for the sampling distribution.

1. According to the CDC, approximately 9.3% of people in the US have diabetes. Random samples of 500 people are taken to determine the rates of diabetes.

   (a) Does this meet the requirements for a normal distribution? (Explain why)

   (b) If so, state the mean and standard error for the sampling distribution.

2. According to the ASPCA, approximately 44% of households in the US have dogs. Random samples of 80 households are taken to determine dog ownershp.

   (a) Does this meet the requirements for a normal distribution? (Explain why)

   (b) If so, state the mean and standard error for the sampling distribution.

3. The odds of finding a four-leaf clover are quite low. While exact figures are hard to find, one researcher claims that only 1 in 1000 clovers will be four-leaf clovers. A random sample of 2000 clovers from one field are selected in a search for an accurate rate of the appearance of four-leaf clovers.

   (a) Does this meet the requirements for a normal distribution? (Explain why)

   (b) If so, state the mean and standard error for the sampling distribution.

   Teacher Note: This example is not appropriate for a normal distribution. It fails the minimum count of 10 requirement for successes. It also likely fails the "independent" requirement. The appearance of an extra leaf on a clover is a genetic mutation. Once a genetic mutation has occured, it is more likely to repeat in the same field.

▲

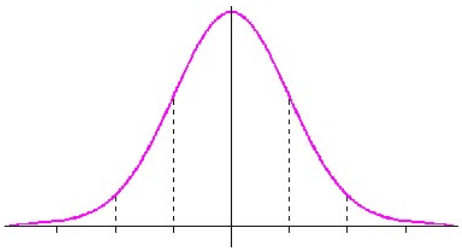
Activity 3.7. Finding p-values

1. According to the CDC, approximately 9.3% of people in the US have diabetes. A researcher is investigating whether this rate is lower in a community of recent immigrants from Africa. She gathers a random sample of 500 recent immigrants and finds that the rate of diabetes in her sample is 8.2%.

   (a) State the null and alternate hypotheses.

   (b) Find the p-value. Use the mean and standard error you computed in the prior activity rather than creating a simulation in TinkerPlots. **Solution:** z-score = -0.847, p-value = 0.1985

   (c) State your conclusions as a complete sentence. **Solution:** With a p-value of 0.1985, we do not have sufficient evidence to claim that recent immigrants from Africa have a different rate of diabetes than the general population in the U.S.

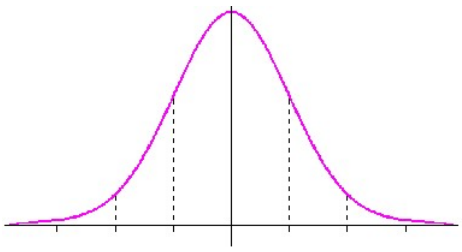


2. According to the ASPCA, approximately 44% of households in the US have dogs. A researcher wants to know if the rate of dog ownership is higher in rural areas. He takes a random samples of 80 rural households and finds that 51 had a dog.

   (a) State the null and alternate hypotheses.

   (b) Find the p-value. Use the mean and standard error you computed in the prior activity rather than creating a simulation in TinkerPlots. **Solution:** z-score = 3.559, p-value=0.000186

   (c) State your conclusions as a complete sentence. **Solution:** With a p-value of less than 0.01, we have strong evidence to reject the claim that 44% of rural households have a dog. The proporiton of dog ownership is higher than 44% and we could follow-up with a confidence interval about that rate, in a later section.



▲

3.2 Confidence Intervals for one proportion

Reading Assignment 3.8. • Read Sect 3.1 in the OpenIntro Statistics textbook (p123-128). Take notes on new formulas and vocab. Then answer the preview questions.

▲

Preview Activity 3.9. Confidence Intervals

1. When we measure a proportion we can then use it to estimate a standard error for the normal sampling distribution. The standard error is a proxy for the standard deviation of the sampling distribution, and we use this proxy simply because we have no other information. The formula used to calculate the standard error is found on page 124. If we happen to take a poll and find that 32% of a sample of size 100 support a given presidential candidate, then what is the standard error? **Solution:** SE=0.046648

2. Now let's calculate a 95% confidence interval for the poll mentioned in the previous problem. To do so we need a critical z-score. For a 95% confidence interval the text approximates this z-score as: **Solution:** z-score=1.96.

3. The margin of error (MOE) is the critical z-score times the standard error. Find the error for a 95% confidence interval based on the information given in problem 1 using your z-score answer from problem 2: **Solution:** MOE=0.0914.

4. The answer that you just got for problem 3 is called the margin of error. We use this to calculate the upper and lower bounds of the confidence interval. To find the lower bound we take our proportion (the 0.32) and subtract the margin of error. For the upper bound we add the margin of error. Find the lower and upper bound. **Solution:** Lower Bound=0.22857. Upper Bound =0.41143.

5. Finally, this gives us a range estimate for what we think the true population proportion is. Summarize this process in your notes so you can find it later.

Definition 3.10. What is a confidence interval?
In order to estimate a population mean or proportion based on a sample of data, we could provide a point estimate (sample mean $\bar{x}$ or sample proportion $\hat{p}$.) However, it is unlikely that our point estimate from our sample will be perfectly accurate. Different samples would result in slightly different averages. So instead of providing a point estimate, we will provide a likely interval for the population parameter. This is done through a confidence interval.

"A plausible range of values for the population parameter is called a **confidence interval.** Using only a point estimate is like fishing in a murky lake with a spear, and using a confidence interval is like fishing with a net. We can throw a spear where we saw a fish, but we will probably miss. On the other hand, if we toss a net in that area, we have a good chance of catching the fish." –OpenIntro Stats, Sect 2.8

To construct a 95% confidence interval, we can use the following formula (if we think

the underlying sampling distribution is basically normal)

$$\text{95\% CI:} \qquad \text{point estimate} \pm 1.96 \times SE$$

The number 1.96 is the critical z-score for capturing the middle 95% of data in a normal distribution. With the 68%, 95%, 99.7% Rule in Chapter 2, we had stated that 95% of data is within approximately 2 standard deviations from the mean. Now we refine that to being within 1.96 standard deviations to be more precise. While 95% confidence intervals are the most frequent, we can also find 90% and 99% confidence intervals by changing the critical z-score to 1.64 or 2.58, respectfully.

$$\text{90\% CI:} \qquad \text{point estimate} \pm 1.64 \times SE$$

$$\text{99\% CI:} \qquad \text{point estimate} \pm 2.58 \times SE$$

Using norm.s.inv() in Excel, you can find the critical z-score for any level of confidence you prefer. The general formula for a confidence interval becomes

$$\text{CI:} \qquad \text{point estimate} \pm z_{critical} \times SE$$

The second half of the equation, $z_{critical} \times SE$, is called the **margin of error** or MOE.

Standard error measures the variability within data. Standard error is the standard deviation for sampling distributions. The formula for standard error when working with a single proportion is

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

where p is the population proportion and n is the sample size. However, when working with confidence intervals, we do not know what the population proportion p actually is. So we use $\hat{p}$ as an approximation, which alters the SE formula to

$$SE \approx \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}.$$

Warning: If you shift from calculating a p-value to computing a confidence interval, you will need to recalculate SE using $\hat{p}$. Why? Because if you reject the null hypothesis, that means you have rejected p and you should stop using it. Switch to $\hat{p}$.

Example 3.11. According to the ASPCA, approximately 44% of households in the US have dogs. A researcher wants to know if the rate of dog ownership is higher in rural areas. He takes a random samples of 80 rural households and finds that 51 had a dog.

1. In the prior section you found that the null hypothesis was $H_0 : p = 0.44$

and the alternate hypothesis was $H_A : p > 0.44$

2. Next you found the p-value=0.000186, which led you to reject the null hypothesis. The rates of dog ownership are higher in rural areas.

3. If we rejected the null hypothesis that $p = 0.44$, then we should tell the reader what proportion of rural households do have dogs. It would be unsatisfying to just say that it is not 44%. While we could claim that the population proportion for dog ownership in rural areas is $p = \frac{51}{80}$ (or 0.638), this is also only an estimate based on one sample of 80 households. Instead, we'll offer a 95% confidence interval for where we believe the population proportion would be.

4. How do we create a 95% confidence interval? We can use what we know about sampling distributions for samples of size 80 with a center of $\hat{p} = 0.638$.

   - $center = \hat{p} = 0.638$
   - $SE \approx \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} = \sqrt{\frac{0.638(1-0.638)}{80}}$
   - The structure of a confidence interval is $\hat{p} \pm z_{crit} \times SE$
   - To find the middle 95% of data, use the z-critical score 1.96, which can be found using the command norm.s.inv(0.025) in Excel.

Teacher Note: What data sets would capture your students' attention? If your students are drawn from a common major, this may be a good source for data. Feel free to swap out with another real world data set. If an election is on the horizon, you may want to look at current events and political polling.

Activity 3.12. Cards Against Humanity, partnered with Survey Sampling International, has engaged in some public opinion polling. The results of these polls can be found at ThePulseOfTheNation.com. They ask burning questions, such as

- Do you believe your current job is likely to be replaced by robots?
- Do you believe in climate change?
- Would you rather be smart and sad, or dumb and happy?
- Do you believe in ghosts?

Choose 3 of the above questions and create a 95% confidence interval for each. To get started:

1. Open the data set PulseOfTheNationSept2017Data from the class website. You will want to use Pivot Tables in Excel to generate counts for these confidence intervals. Responses of DK/REF refer to Don't Know and Refuse to answer. Exclude these responses from your calculations. This will result in different sample sizes depending on how many participants actually provided answers.

2. Once you find your confidence interval, state your conclusion as a complete sentence that includes the context of the problem. For example: Based on the data from the

study, we are 95% confident that between 44.2% and 47.8% of American adults believe that ketchup is a vegetable.

▲

Activity 3.13. Confidence intervals and hypothesis tests using real world data.

Use the NCBabySmoke data from North Carolina (adapted from OpenIntro Stats) for the following problems.

1. Create a 95% confidence interval for the proportion of new moms who are married.
2. Create a 99% confidence interval for the proportion of newborns who are premies.

column name	description and units
fage	father's age
mage	mother's age
mature	under 35 vs. 35 or older
weeks	length of pregnancy
premie	premie or full term
visits	number of doctor visits
marital	married or not married
gained	weight gained by mom (lbs)
weight	weight of baby (lbs)
lowbirthweight	low is $\leq$ 5.5 lbs
gender	baby's gender
habit	smoking habit of mom
whitemom	white or not white

3. Are premies 50% girls and 50% boys, or are premie boys more common (in NC)? Conduct a hypothesis test and then follow-up with a confidence interval if appropriate. Note: For this question, you will have considerably less than 1000 babies. Use a pivot table to get a count of premies vs. full term babies, and to sort boys and girls.
4. According to www.childtrends.org, approximatley 8% of pregnant women in the US reported smoking in 2014. Is the rate of smoking higher than this for new moms in NC? Conduct a hypothesis test and then follow-up with a confidence interval if appropriate.

▲

3.3 Comparing two proportions

Reading Assignment 3.14.

- Read pages 157 - 167 in the Cartoon guide. Take good notes since there are many technical details. Once you're done answer the questions on the preview.
- Read Section 3.2 in the OpenIntro Textbook.

▲

Preview Activity 3.15.

1. The standard error formula for confidence intervals on two proportions is given in the middle of the left-hand side of page 164. Suppose that we redo the aspirin study and find that 3.2% of the people in a 100-person placebo group had a heart attack and that 1.9% of the people in an 89-person control group had a heart attack. Use the standard error formula to calculate the standard error for the difference between these proportions (be careful with parenthesis): **Solution:** $SE_{unpooled} = 0.022786$

2. When doing a hypothesis test on the difference of two proportions we have to calculate the pooled standard error. This assumes that the two groups come from the same population so we need to consider the total number of heart attacks in the total group (treatment & control). Let's assume that we have the same data as in problem 1 above. To find the pooled proportion use the following formula:

$$p_{pooled} = \frac{x_1 + x_2}{n_1 + n_2} = \frac{\hat{p}_1 n_1 + \hat{p}_2 n_2}{n_1 + n_2}$$

Calculate the pooled proportion for the data given above: **Solution:** $p_{pooled} = 0.02588$

$$SE_{pooled} = \sqrt{\hat{p}_{pooled}(1 - \hat{p}_{pooled})(\frac{1}{n_1} + \frac{1}{n_2})}$$

Calculate the pooled standard deviation for the data given above. You can find this formula on page 165: **Solution:** $SE_{pooled} = 0.02313$

This could be one or two days, depending on the types of activities we have.

Example 3.16. Open the NCBabySmoke data set from Moodle (from OpenIntro). Previously, we've explored this data set to ask probability questions and to work with single proportions. Now we are able to compare proportions from two subgroups of a population and determine if there is a statistically significant difference between those proportions. Let's explore whether there is a difference in smoking rates among mature and younger moms.

- Our null hypothesis in this situation is that there is no difference in smoking rates between mature moms and younger moms. Mathematically, that can be written as

$$H_0 : p_{mature} = p_{younger}$$
or
$$H_0 : p_{mature} - p_{younger} = 0$$

Usually the second option is used because it highlights the idea that the center of our sampling distribution will be 0 if we anticipate 0 difference between the two groups.

The alternate hypothesis is that there is a difference in smoking rates, but we don't have a theory of which group has a higher rate. Therefore this will be a two-tailed test.

$$H_A : p_{mature} - p_{younger} \neq 0$$

- Use a pivot table in Excel to organize your data. In the picture below, "mature" is put in the row tab and "habit" is put in the column tab. These can be swapped, of course. Count of habit is displayed within the pivot table.

Count of habit	Column Labels			
Row Labels	NA	nonsmoker	smoker	Grand Total
mature mom	1	121	11	133
younger mom		752	115	867
Grand Total	**1**	**873**	**126**	**1000**

- Find the proportion of mature moms and younger moms who smoke.

  $p_{mature} = \frac{11}{133} = 0.0827$ and $p_{younger} = \frac{115}{867} = 0.1326$

- Find the difference in proportions and then determine if that is statistically significant.

  The difference in rates of smoking based on age is $p_{mature} - p_{younger} = -0.0499$.

- In order to determine whether the difference of -0.0499 is significant, we need a way to measure the variability of paired samples. The standard error formula for a two proportion hypothesis test:

$$SE_{pooled} = \sqrt{p_{pooled}(1 - p_{pooled})(\frac{1}{n_1} + \frac{1}{n_2})}$$

where the pooled proportion is $p_{pooled} = \frac{x_1 + x_2}{n_1 + n_2} = \frac{p_1 n_1 + p_2 n_2}{n_1 + n_2}$

For this problem, $p_{pooled} = \frac{11 + 115}{133 + 867} = \frac{126}{999} = 0.126.$

Therefore $SE_{pooled} = \sqrt{0.126(1 - 0.126)(\frac{1}{133} + \frac{1}{867})} = 0.0309$

- Let's find the p-value. How rare is it to find a difference of -0.0499, if the standard error is 0.0309? Use the norm.dist function in Excel. Be sure to double the results because this is a two-tailed test.

$$2 \times norm.dist(-0.0499, 0, 0.0309, 1) = 2 \times 0.05317 = 0.1063$$

- With a p-value of 0.1063, we do not have sufficient evidence to say the difference we found was significant. We cannot claim that there is a significant difference in the rates of smoking between younger and more mature new moms.

Activity 3.17. Comparing proportions
Use the NCBabySmoke data set to answer each of the following questions.

column name	description and units
fage	father's age
mage	mother's age
mature	under 35 vs. 35 or older
weeks	length of pregnancy
premie	premie or full term
visits	number of doctor visits
marital	married or not married
gained	weight gained by mom (lbs)
weight	weight of baby (lbs)
lowbirthweight	low is $\leq$ 5.5 lbs
gender	baby's gender
habit	smoking habit of mom
whitemom	white or not white

1. Is there a difference in smoking prevalence between new moms who are married and not married?

   (a) State the null and alternative hypotheses

   (b) Find and label the proportions ($p_{married}$ and $p_{notmarried}$) and the sample size $n_{married}$ and $n_{notmarried}$ using a pivot table in Excel.

   (c) Find the difference in proportions, the standard error, and the p-value. Label each in your Excel sheet. Note: You will need to use the pooled proportion in your calculations for SE because your null hypothesis is that there is no difference between the two proportions.

   (d) State your conclusions in a complete sentence related to the context of the problem.

   (e) Are you surprised by the conclusions? Did you expect something different?

   (f) If you rejected the null hypothesis, what should you do to follow-up?

Solution:

Count of marital	Column Labels			
Row Labels	NA	nonsmoker	smoker	Grand Total
married		320	66	386
NA	1			1
not married		553	60	613
Grand Total	**1**	**873**	**126**	**1000**

	married	not married
x = smoker	66	60
n = total	386	612
p = proportion	0.170984456	0.0980392
p_diff	0.07294524	
p_pooled	0.126252505	
SE_pooled	0.021587825	
test stat	3.378999116	
p-value (2 tailed)	0.000727502	significant
SE_unpooled	0.022621101	
z_crit (95%)	1.96	
MOE	0.044337358	
lower	0.028607882	
upper	0.101553122	

2. Is there a difference in rates of low weight babies between smoking moms and non-smoking moms?

   (a) State the null and alternative hypotheses

   (b) Find and label the proportions ($p_{smoking}$ and $p_{nonsmoking}$) and the sample size $n_{smoking}$ and $n_{nonsmoking}$.

   (c) Find the difference in proportions, the standard error, and the p-value. Label each in your Excel sheet.

   (d) State your conclusions in a complete sentence related to the context of the problem.

   (e) Are you surprised by the conclusions? Did you expect something different?

   (f) If you rejected the null hypothesis, what should you do to follow-up?

Solution:

Count of habit	Column Labels			
Row Labels	NA	nonsmoker	smoker	Grand Total
low	1	92	18	111
not low		781	108	889
Grand Total	**1**	**873**	**126**	**1000**

	nonsmoker	smoker
x = low	92	18
n = total	873	126
p = proportion	0.105383734	0.1428571
p_diff	-0.037473409	
p_pooled	0.11011011	
SE_pooled	0.029831293	
test stat	-1.2561778	
p-value (2 tailed)	0.209051514	not significant

3. Generate two more research questions that you could ask of this data. Choose one question which could be answered with a one proportion hypothesis test and choose a second question that requires a two proportion hypothesis test.

▲

In order to follow-up with a confidence interval, we can use the same confidence interval strategy as before, but we will need a different formula for the standard error SE.

Confidence Interval:

$$\text{point estimate} \pm z_{critical} \times SE$$
$$(\hat{p_1} - \hat{p_2}) \pm z_{critical} \times SE$$

While we have a standard error formula for hypothesis testing, this formula assumes that the null hypothesis is true ($p_1 = p_2$). In that case we pooled the two proportions for the SE formula. But this is not an appropriate assumption for working with confidence intervals. Instead, we will use the unpooled standard error formula.

$$SE_{unpooled} = \sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}$$

Teacher Note: Near the end of the lesson, follow-up on the two problems above. The smoking vs. marital status has a significant difference in proporitons, so it should be followed up with a confidence interval. The 95% confidence interval is in the solutions above.

Activity 3.18. Let's revisit the Cards Against Humanity data set (Source: ThePulseOfTheNation.com) and investigate a little deeper.

1. In an earlier activity, you may have investigated the responses to "Would you rather be smart and sad, or dumb and happy?" You likely found that around 52.7% of the people who responded to the question indicated that they would rather be dumb and happy instead of smart and sad. You then created a confidence interval around this sample statistic to make a claim about the wider population.

   But do you suppose that the results are the same for all groups of people? For example, do men and women have different responses to this query? What about Democrats vs. Republicans? What about those who believe that climate change is real and those who do not? Let's test a few of these ideas.

   (a) Determine whether there is a statistically significant difference in the way men and women respond to the question: "Would you rather be smart and sad, or dumb and happy?" For the purpose of the question, exclude participants who responded DK/REF or Other. After you have stated the results of your hypothesis test, follow-up with a confidence interval if that is appropriate. Hint: Only follow-up with a confidence interval if you reject the null hypothesis.

   (b) Now let's consider beliefs about climate change. A pair of political scientists suspect that people who do not believe that climate change is real might also prefer not to be smart and sad. In order to conduct a hypothesis test about their theory, they ask you to lump together both groups that believe it is real (whether they think its caused by people or not) and separate the group that responded that it is not real at all.

   After you have stated the results of your hypothesis test, follow-up with a confidence interval if that is appropriate.

   Note: Statisticians frequently make choices about how to categorize responses. It is important for consumers of statistics to be aware of this and for statisticians to be transparent about the decisions they make. These choices may affect conclusions. We could have combined groups differently or we could have left out one of the three subgroups and compared only two of them.

2. **What do you notice, what do you wonder?** Choose another question you found interesting in the Pulse of the Nation data set. Determine if two subgroups of the population had different opinions about that question.

   Common subgroups of interest include men vs. women, Democrat vs. Republican, and college educated vs. high school only. *Note: At this point in the term, we are focusing on comparing two groups with proportions. We are not yet able to compare 3 or more groups, although such statistical tests are possible. If you interested in more, look into* χ^2 *tests.*

   Once you have chosen your question, use the following steps:

   (a) Clearly state your hypotheses.

(b) Find and label your proportions $(\hat{p_1}, \hat{p_2})$ and sample sizes (n_1, n_2) in your Excel sheet

(c) Find and label the difference in proportions $(\hat{p_1} - \hat{p_2})$, the pooled proportion $(\hat{p}_{pooled})$, the standard error (SE), and the p-value.

(d) State your conclusions in a complete sentence. Keep in mind that your classmates may have chosen different questions to investigate, so provide us iith enough information to understand what you discovered.

(e) If you reject your null hypothesis (that the two subgroups were the same), follow-up with a confidence interval to tell us how different the two subgroups are.

▲

3.4 Lab 6

Reading Assignment 3.19.

- For this preview activity you are going to read the article that the next lab is based on.

  Read the article: Gall-up: American's desire for a third political party

- Maybe read lab 6 ahead of time?

▲

Preview Activity 3.20. **Teacher Note:** Perhaps starts Lab 6 problem - we could do some guess and check on the algebra problem of sample size... that was a time sink last time

Now that you've read the article, consider these questions:

1. The 57% stated in the article is based off of a sample of Americans that participated in the phone survey. Does this tell us that exactly 57% of all Americans desire a third party? [yes or no]

2. In the "Survey Methods" section of the article (at the end), they state a margin of error of 4% with 95% confidence. Select the correct response:

   (a) If we gathered many samples in the same way, 95% of the resulting confidence intervals would contain the correct response.

   (b) There is a 95% probably of the true proportion being between 53% and 61%. [Isn't this also correct?]

   (c) The true proportion of Americans that want a third party is actually 57% and the margin of error just allows for some random variation.

3. The margin of error in a poll is affected by the sample size. Not surprisingly, the bigger the sample size, the smaller the margin of error. That is, if a pollster asks a larger number of people a polling question, the precision of their results improves.

   The margin of error formula for a 95% confidence interval is $MOE = 1.96 \times \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$. Since $\hat{p}$ is unknown in this discussion, let's assume it is $\hat{p} = 0.5$. For each of the following, assume we are looking for a confidence level of 95%.

   (a) How big will the margin of error be if the sample size is $n = 100$?

   (b) How big will the margin of error be if the sample size is $n = 500$?

   (c) How big will the margin of error be if the sample size is $n = 3000$?

   (d) If you want the margin of error to be 5%, what should you choose as your sample size? (To solve this, you may need to do some rearranging with algebra.)

4. Now you should contact your partner and start working on the lab early. This lab has the potential to be a bit long so an early start is a good idea!

Activity 3.21. Lab 6

Gallup poll context about the need for a third political party
Immigration context for two proportions, based on a Pew research study.

▲

3.5 χ^2 goodness of fit - Optional

Teacher Note: If you would like to contribute to this section, please let us know.

3.6 Summary

Summary 3.22. Student learning outcomes from Chapter 3

1. Students are able to make use of the Central Limit Theorem and have a conceptual understanding of sampling distributions for qualitative data.
2. Students are able to conduct hypothesis tests and create confidence intervals for one-proportion scenarios. Students are able to interpret the results into everyday language.
3. Students are able to conduct hypothesis tests and create confidence intervals for two-proportions scenarios. Students are able to interpret the results into everyday language.
4. Students can clearly articulate the meaning of a p-value.