

# Unit 28: Inference for Proportions



## SUMMARY OF VIDEO

It is nearly impossible to collect data about an entire population. Take, for example, all the salmon in one watershed. We can't count the number of eggs laid by every single spawning salmon. But we can count the eggs laid by a sample of some of these salmon. Then, using statistical inference, we can use the mean number of eggs in our sample to draw conclusions about the egg-laying population as a whole. As part of the inference procedure, we use probability to indicate the reliability of our results.

We can also use statistical inference to estimate a population proportion. For instance, suppose we wanted to know how many of the eggs laid by the salmon were fertilized. We could investigate the fertilization rate in our sample to get a sample proportion or sample percentage. Then we could use the sample proportion as an estimate of the unknown population proportion. But how good of an estimate is it? This will be the topic of this video – using information from samples to make inferences about population proportions.

Let's turn our attention to a completely different context: the workplace. Employers think about how to motivate their employees to do their best, most creative work. Psychologist Teresa Amabile has studied creativity for years. One of Amabile's discoveries from her earlier research is that creativity fluctuates, even for a given individual, as a function of the kind of work environment the individual is in. Building on that foundation, Amabile designed a study around the question of worker motivation. She recruited 238 people with creative jobs who were willing to keep track of their activities, emotions, and motivation levels every workday. Their electronic diaries had two components. One consisted of participants rating their motivation, emotions, and other subjective factors on a seven-point scale. The second component was an open-ended question where participants were asked to describe one event that stood out that day. It could be anything, as long as it was relevant to the work or the project. After several years, Amabile had nearly 12,000 diary entries. These entries validated her earlier findings that people were able to solve problems creatively and come up with new ideas on days they felt most motivated and excited about their work. So, the next question to ask was: What led to high levels of motivation?

Dipping into the diaries, Amabile was able to see that one factor, more than anything else, made people feel they were having a great day at work. That was simply making progress in meaningful work, even if the progress looked incremental. She called this the Progress Principle. It turned out that 76% of participants' best days had a progress event; whereas only 25% of their worst days had a progress event. Progress was paramount for people to feel positive and highly motivated – much more than other things like support from management and coworkers, feelings of doing important work, or collaboration, as can be seen from Figure 28.1.



Figure 28.1. Type of event recorded on workers best and worst days.

Amabile and her coauthor decided to survey managers to see whether they were aware of how important this feeling of progress was in motivating workers. She asked them to rate five different items in order of how much they felt they affected workers' motivation. If the managers just randomly chose one of the five options to rank as most important, we would expect 20% of them to pick progress. So, we let  $p$  be the proportion of all managers who would pick progress as the most important of the five items for motivating workers. Now, we can set up a test of hypothesis for the population proportion,  $p$ :

$$H_0 : p = 0.20$$

$$H_a : p \neq 0.20$$

As it turned out, only 35 out of 669 managers selected progress as the top motivational factor. That gives a sample proportion of just 0.0523, or a mere 5.23%. This seems pretty low compared to the 20% proportion from our null hypothesis. But is it low enough to reject the null hypothesis? To find out, we can turn to a z-test statistic:

$$z = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}},$$

where  $\hat{p}$  (pronounced p-hat) is our sample proportion,  $p_0$  is the null hypothesis proportion, and  $n$  is the sample size. Substituting our sample proportion and sample size we get:

$$z = \frac{0.0523 - 0.20}{\sqrt{\frac{(0.20)(1 - 0.20)}{669}}} \approx -9.55$$

That is a pretty extreme z-test statistic. If you compare it to a standard normal distribution, being 9.55 standard deviations from the mean is highly unlikely. As can be seen from Figure 28.2, the area under the curve that far out is not really visible! In fact, the  $p$ -value is 0.000. So, we have our answer: reject the null hypothesis and accept the alternative. The population proportion of all managers in the world who would select “Support for Making Progress” as the most important motivator is not 20%.

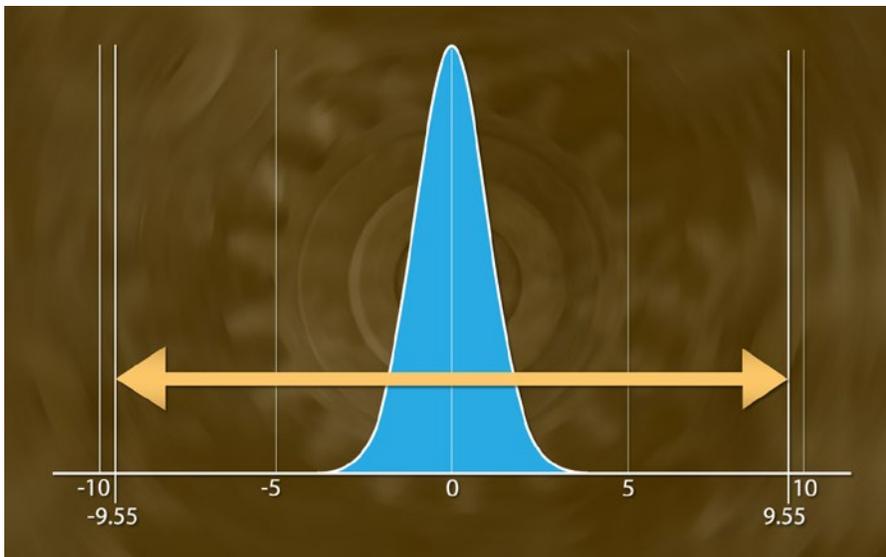


Figure 28.2. Determining a  $p$ -value from a standard normal density curve.

Now that we have rejected the null hypothesis, let’s calculate a confidence interval for the true population proportion. We know that the sample proportion of managers who selected progress was 0.0523, but we don’t know how close that is to the true population proportion. Just like in the confidence intervals for one mean, we can figure out a standard error to go with our point estimate. Here’s the formula for the confidence interval:

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

Suppose we decide that we want a 95% confidence interval. Then our value for  $z^*$  is 1.96 just as it was for z-confidence intervals for a population mean.

Next, we use our sample information to calculate the 95% confidence interval for the population proportion,  $p$ :

$$0.0523 \pm 1.96 \sqrt{\frac{0.0523(1-0.0523)}{669}},$$

$$0.0523 \pm 0.0169 \text{ or } 5.23\% \pm 1.69\%$$

So, our estimate is that only between 3.5% and 6.9% of managers in the overall population would rate progress as the number one motivational factor. A good question to ask is how could managers be so unaware of what really counted to their employees? What managers have said in response to that question is that it is just part of their employees' jobs – they are supposed to make progress. Managers don't typically think of progress as something that they need to worry about. But, according to Amabile, they actually do need to worry about it a lot. What Amabile saw in the diaries was that there were often little hassles happening in the work lives of most of the study participants that kept them from making as much progress as they would like. These were things that managers could have cleared away for them, without a lot of effort, if they had just been paying attention.

On some level the workers themselves might have recognized that their best days often went hand-in-hand with progress events. But the managers basically had no clue. It is the kind of finding that makes perfect sense once you know about it. Sometimes you just have to ask the right questions and know how to analyze the data.

# STUDENT LEARNING OBJECTIVES

- A. Identify inference problems that concern a population proportion.
- B. Know how to conduct a significance test of a population proportion.
- C. Be able to calculate a confidence interval for a population proportion.
- D. Understand that the z-inference procedures for proportions are based on approximations to the normal distribution and that accuracy depends on having moderately large sample sizes.

# CONTENT OVERVIEW

Up to this point, all the inference procedures we have discussed involve using sample means,  $\bar{x}$ , to make inferences about population means,  $\mu$ . In this unit, we focus on proportions. For example, what if we wanted to know what proportion of people own or use a computer at home, or have access to the Internet from home, or from work, or from school? In order to answer these questions, we need new inference procedures designed for proportions.

In inference, we start by defining the population – for our question on home-use of computers, the population will be all households in America. Of interest is the **population proportion**,  $p$ , of households in which some member owns or uses a computer at home. Now, we don't have access to every household in America, but we can take a sample. In a random sample of 2,500 households, 2,036 answered yes to the following question:

At home, do you or any member of this household own or use a desktop, laptop, netbook, or notebook computer?

From this information we can calculate the **sample proportion**, which we label as  $\hat{p}$ :

$$\hat{p} = \frac{2036}{2500} = 0.8144, \text{ or } 81.44\%$$

But how good is this estimate for  $p$ ? Remember, the sample proportion,  $\hat{p}$ , is a statistic. If we take another sample of 2,500 households, we will most likely get a different estimate for  $p$ . So, as a first step in developing inference procedures for population proportions, we need to know something about the **sampling distribution of the sample proportion**,  $\hat{p}$ .

## Sampling Distribution of a Sample Proportion

Suppose that a large population is divided by some characteristic into two categories, successes and failures, and that  $p$  is the population proportion of successes. A simple random sample of size  $n$  is drawn from the population and  $\hat{p}$  is the sample proportion:

$$\hat{p} = \frac{\text{number of successes in the sample}}{n}$$

As a statistic,  $\hat{p}$  varies over repeated sampling. Its sampling distribution has the following properties:

- Mean:  $\mu_{\hat{p}} = p$
- Standard deviation:  $\sigma_{\hat{p}} = \sqrt{\frac{p(1-p)}{n}}$ .
- Distribution: For large  $n$ ,  $\hat{p}$  has an approximately normal distribution.

Since, in the case of home use and/or ownership of computers, the sample size is large, 2,500, the sampling distribution of  $\hat{p}$  is approximately normal (as pictured in Figure 28.3.)

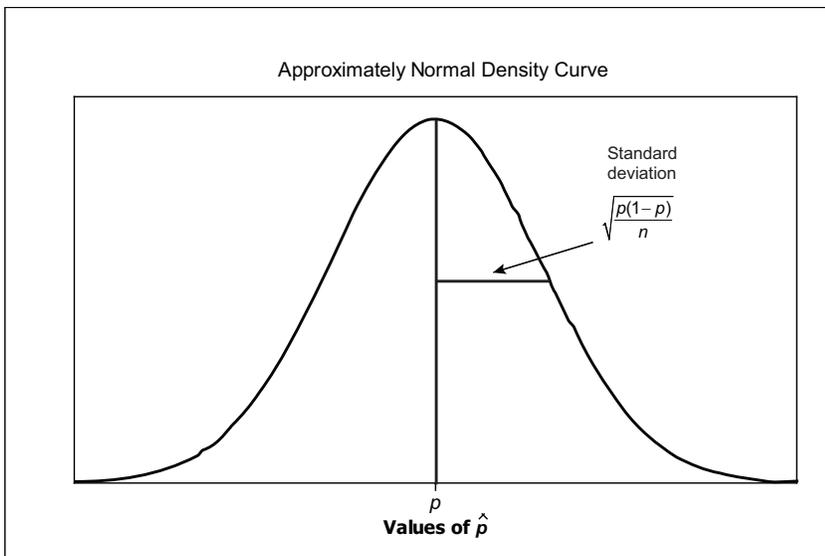


Figure 28.3. Sampling distribution of the sample proportion,  $\hat{p}$ .

Suppose that an online source claimed that 79% of American households had a member of the household who owned or used a computer at home. We would like to test that claim. To do so, we use the online source's claim about the population to set up the null and alternative hypotheses:

$$H_0 : p = 0.79$$

$$H_a : p \neq 0.79$$

Now, if the null hypothesis is true, then the distribution of  $\hat{p}$  from a sample with  $n = 2,500$  will have an approximately normal distribution with the following mean and standard deviation:

$$\mu_{\hat{p}} = 0.79$$

$$\sigma_{\hat{p}} = \sqrt{\frac{(0.79)(1-0.79)}{2500}} \approx 0.0081$$

Since we are dealing with an approximately normal distribution, we can express  $\hat{p}$  in standardized units (subtract the mean and divide by the standard deviation):

$$z = \frac{\hat{p} - 0.79}{0.0081}$$

If the null hypothesis is true,  $z$  will have a standard normal distribution. Now, go back to the results of the survey,  $\hat{p} = 0.8144$ , and express that value in standardized units:

$$z = \frac{0.8144 - 0.79}{0.0081} \approx 3.01$$

We calculate a  $p$ -value for the significance test by determining how likely it is to observe a value from the standard normal distribution that is at least 3.05 from the mean. In this case, we get a  $p$ -value of  $2(0.001306) \approx 0.003$ . Since this  $p$ -value  $< 0.05$ , we can reject the null hypothesis and conclude that the population proportion is not 0.79, or 79%.

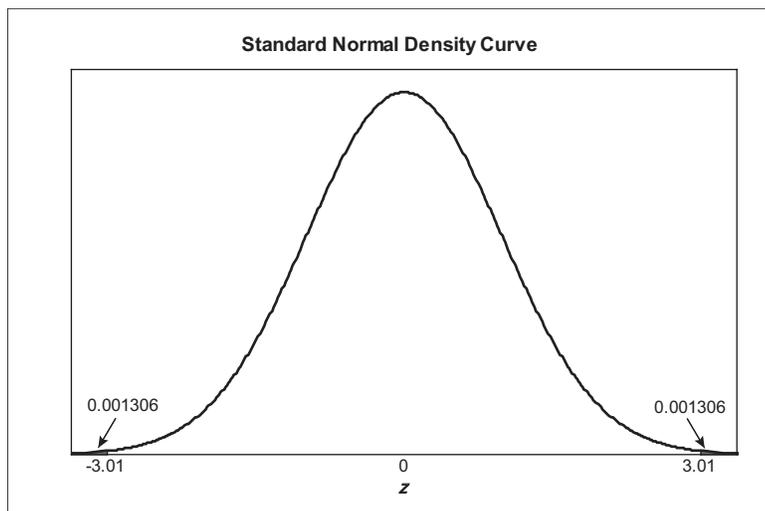


Figure 28.4. Calculating the  $p$ -value of a  $z$ -test statistic.

Before moving on, we summarize the basics of a significance test for population proportions.

### Significance Test for a Population Proportion

To test the null hypothesis  $H_0 : p = p_0$ , where  $p$  is the population proportion and  $p_0$  is the hypothesized value, we use the **z-test statistic**:

$$z = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}}$$

where  $\hat{p}$  is the sample proportion. When the null hypothesis is true and the sample size is large, the z-test statistic will have an approximate standard normal distribution.

Now that we have rejected the null hypothesis that members of 79% of American households own/use a computer at home, let's calculate a confidence interval for the true population proportion. The formula for a **confidence interval for a population proportion** follows the same pattern that was used to calculate a confidence interval for a population mean:

Point estimate  $\pm z^*$ (standard error of point estimate)

Point estimate  $\pm$  margin of error

Here's the formula for calculating a confidence interval for a population proportion.

### Confidence Interval for a Population Proportion

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

where  $\hat{p}$  is the sample proportion and  $z^*$  is the z-critical value (from a standard normal distribution) associated with the confidence level.

Suppose we decide on a 95% confidence interval for  $p$ . Then we use  $z^* = 1.96$ , just as we did in Unit 24, Confidence Intervals. All that is left is to substitute our observed sample proportion,  $\hat{p} = 0.8144$  into the formula (*Continued on next page*):

$$0.8144 \pm 1.96 \sqrt{\frac{(0.8144)(1 - 0.8144)}{2500}} = 0.8144 \pm 0.0152$$

81.44%  $\pm$  1.52%, or between 79.92% to 82.96%

So, now we are ready to use sample proportions to conduct significance tests and calculate confidence intervals for population proportions.

# KEY TERMS

Assume that a population is divided into two categories, successes and failures, based on some characteristic. The **population proportion**,  $p$ , is:

$$p = \frac{\text{number of successes in the population}}{\text{population size}}$$

Draw a sample of size  $n$  from this population. Then the **sample proportion**,  $\hat{p}$ , is calculated as follows:

$$\hat{p} = \frac{\text{number of successes in the sample}}{n}$$

If the sample size  $n$  is relatively large, the **sampling distribution of the sample proportion**,  $\hat{p}$ , is approximately normally distributed with the following mean and standard deviation:

- Mean:  $\mu_{\hat{p}} = p$ , where  $p$  is the population proportion.
- Standard deviation:  $\sigma_{\hat{p}} = \sqrt{\frac{p(1-p)}{n}}$ , where  $n$  is the sample size.

To test the null hypothesis  $H_0 : p = p_0$ , where  $p$  is the population proportion, we can use the **z-test statistic for proportions**. The formula for the z-test statistic is:

$$z = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}}$$

The z-test is used in situations where the sample size  $n$  is large.

In situations where the sample size  $n$  is large, a **confidence interval for the population proportion**,  $p$ , can be calculated from the formula:

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

where  $\hat{p}$  is the sample proportion and  $z^*$  is the z-critical value (from a standard normal distribution) associated with the confidence level.

# THE VIDEO

Take out a piece of paper and be ready to write down answers to these questions as you watch the video.

1. What is the general topic of this video?
2. In Teresa Amabile's earlier study of workers in creative jobs, how did participants of the study feel on the days when they were most able to solve problems creatively and come up with new ideas?
3. Describe the principle that Amabile dubbed the Progress Principle.
4. Managers were given five items, including progress, and asked to select the one that they felt most affected workers' motivation. If managers randomly selected one of the five items, what percentage of the managers would we expect to select progress?
5. What type of test statistic was used to test the null hypothesis  $H_0 : p = 0.20$ , where  $p$  is the population proportion?
6. In the video, a 95% confidence interval was calculated for the true population proportion of managers who would select progress as the most important motivational factor. After converting to percentages, were the values in this confidence interval below 20%, around 20%, or above 20%?

# UNIT ACTIVITY:

## PROPORTIONS OF BLUE EYES

In the activity for Unit 21, you completed Table 21.1 by simulating data for inheriting blue eyes (genes bb) from brown-eyed parents who carried a recessive gene for blue eyes (genes Bb). You will need those data for this activity. In this activity, the population consists of the children of brown-eyed parents, each of whom carries a recessive gene for blue eyes. In this case, the true population proportion is known, which is generally not the case, and  $p = 0.25$ . In this case, knowing the population proportion allows us to see how well the statistics perform.

Sample Number	Number of Blue-Eyed Children $n = 4$	Estimated Proportion Blue-Eyed Children $n = 4$	Running Total Number of Children	Running Total Number of Blue-Eyed Children
1				
2				
3				
4				
5				
6				
7				
8				
9				
10				
11				
12				
13				
14				
15				
16				
17				
18				
19				
20				
21				
22				
23				
24				
25				
26				
27				
28				
29				
30				

Table 28.1. Data on children's eye color.

1. In a copy of Table 28.1, enter the  $x$ -values from your completed Table 21.1 into the second column.
  
2.
  - a. For each sample of four children, calculate the sample proportion of blue-eyed children,  $\hat{p}$ . Enter the sample proportions in the third column of Table 28.1.
  - b. Notice that your sample proportions vary from sample to sample (even though the population proportion stayed the same). What was the smallest sample proportion? What was the largest?
  - c. To get a sense of the shape of the sampling distribution of the sample proportion, make a histogram of your values for  $\hat{p}$  (from column three). Use class intervals of width 0.25 for your histogram. Does your histogram indicate that the sample proportions have a normal distribution?
  
3.
  - a. Complete the fourth column of Table 28.1 by entering a running total of the number of children as samples are combined.  
This list should contain the following numbers: 4, 8, 12, . . . , 120.
  - b. Complete the fifth column of Table 28.1 by entering a running total of the number of blue-eyed children as samples are combined.
  
4. The confidence interval formula given in the Content Overview is for large sample sizes. After combining the data from the first 10 samples, you now have a sample of 40 children.
  - a. Give a point estimate for the population proportion,  $p$ , of blue-eyed children based on the 40 children from Samples 1 – 10.
  - b. Compute a 95% confidence interval for  $p$ . (Round to three decimals.)
  - c. How big is the margin of error in your confidence interval in (b)?
  
5. After combining the data from the first 20 samples, you now have a sample of 80 children.
  - a. Give a point estimate for the population proportion,  $p$ , of blue-eyed children based on your sample of 80 children.

- b. Compute a 95% confidence interval for  $p$ .
- c. How big is the margin of error in your confidence interval in (b)?
6. After combining the data from all 30 samples, you now have 120 children.
- a. Give a point estimate for the population proportion,  $p$ , of blue-eyed children based on your sample of 120 children.
- b. Compute a 95% confidence interval for  $p$ .
- c. How big is the margin of error in your confidence interval in (b)?
7. Compare the margins of error for the three confidence intervals that you computed in questions 4 – 6. What happened to the margin of error as the sample size increased?
8. From questions 4 – 6, we know that sample size affects the margin of error. How large a sample size  $n$  is needed to guarantee that the margin of error for a 95% confidence interval for  $p$  is less than 0.05? Complete parts (a) – (c) to find out.
- a. The margin of error,  $E$ , for a 95% confidence interval is calculated by the following formula:

$$E = 1.96 \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

Replace  $E$  by 0.05 and solve for  $n$ .

- b. If you solved for  $n$  correctly, you found that  $n$  is a multiple of  $\hat{p}(1-\hat{p})$ , which varies for different values of  $\hat{p}$ . Complete the second column of Table 28.2 by calculating the values of  $\hat{p}(1-\hat{p})$  for different values of  $\hat{p}$  (See next page).

$\hat{p}$	$\hat{p}(1-\hat{p})$
0.1	
0.2	
0.3	
0.4	
0.5	
0.6	
0.7	
0.8	
0.9	

Table 28.2 Values of  $\hat{p}(1-\hat{p})$  .

c. To find the value of  $n$  that guarantees a margin of error  $< 0.05$ , substitute the largest value you found for  $\hat{p}(1-\hat{p})$  into your equation in (a). Report the value of  $n$  needed to guarantee that the margin of error will be less than 0.05 (regardless of the value of  $\hat{p}$ ).

9. To conclude this activity, we know that the population proportion of blue-eyed children born to brown-eyed parents with a blue-eye recessive gene is  $p = 0.25$ . Which of your confidence intervals from questions 4 – 6 gave correct results? (In other words, which of your confidence intervals contained the true population mean?)

# EXERCISES

1. A random sample of 2,454 12<sup>th</sup>-grade students were asked the following question:

Taking all things together, how would you say things are these days – would you say you're happy or not too happy? Of the responses, 2,098 students selected happy. (These data were from a *Monitoring the Future* survey.)

- Determine the sample proportion of students who responded they were happy.
- Calculate a 95% confidence interval for the population proportion of 12<sup>th</sup>-grade students who are happy.
- Would a 90% confidence interval for the proportion of happy students be wider or narrower than the one you calculated for (b)? Justify your answer.

2. Currently, mothers in North America are advised to put babies to sleep on their backs. This recommendation has reduced the number of cases of sudden infant death syndrome (SIDS). However, it is a likely cause of another problem – flat spots on babies' heads. A study of 440 babies aged 7 – 12 weeks found that 46.6% had flat spots on their heads.

- The headline of the online news article reporting this story read: Nearly half of babies have flat spots, study finds. Conduct a test of hypotheses to test  $H_0 : p = 0.5$  against  $H_a : p \neq 0.5$  where  $p$  is the population proportion of North American babies aged 7 – 12 weeks who have flat spots on their heads. Report the value of your test statistic, the  $p$ -value, and your conclusion.
- Calculate a 95% confidence interval for the proportion of babies in this age group that have flat spots.
- Suppose you decide to use your confidence interval from (b) to make a decision between  $H_0 : p = 0.50$  and  $H_a : p \neq 0.50$ . Would your decision based on your confidence interval agree with your decision based on the  $z$ -test statistic from (a)? Explain.

3. An online article claims that 90% of American households in which a computer is owned/used have access to the Internet. However, an Internet provider questioned the claim. The Internet provider felt that the percentage should be higher. A phone survey contacted 1,910

households in which a computer was owned/used and respondents were asked if they could access the Internet from their home. A total of 1,816 of the households responded yes.

- a. Define the population.
  - b. Set up the null and alternative hypotheses.
  - c. Calculate the z-test statistic, determine the  $p$ -value, and state your conclusion.
4. Return to question 3. Calculate a 95% confidence interval for the population proportion  $p$ . Re-express your confidence interval as a percentage.

# REVIEW QUESTIONS

1. A sample of 5,462 eighth-grade students were asked whether or not they actively participated in sports, athletics, or exercising on a nearly daily basis. Of the students who responded, 2,998 said yes. (These data were from a *Monitoring the Future* survey.)
  - a. Determine the sample proportion of eighth-grade students who responded that they were involved nearly daily in some sort of physical activity.
  - b. A physical education teacher claimed that over 50% of all eighth-grade students in America actively participate in physical activity on a nearly daily basis. Set up a null hypothesis and an alternative hypothesis to test this claim.
  - c. Conduct a significance test for the population proportion. Report the value of the test statistic, the  $p$ -value, and your conclusion.
  
2. Polls taken a few days before the 2012 presidential election between Barack Obama and Mitt Romney did not indicate a clear winner. An NBC/*Wall Street Journal* poll showed that 48% of the sample intended to vote for Obama. The polling organization announced that they were 95% confident that the sample result was within  $\pm 2.6$  percentage points of the true percent of all voters who favored Obama.
  - a. Explain in plain language to someone who knows no statistics what “95% confident” means in this announcement.
  - b. The poll showed Obama leading Romney 48% to 47%. Yet NBC/*Wall Street Journal* declared the election was too close to call. Explain why.
  
3. A community college conducted a survey of student learning outcomes just prior to graduation. A sample of its students completed the survey. Student responses have been boiled down to two categories, agree or disagree, for the following three questions:
  - a. I have improved in my ability to take responsibility for my own actions.

Valid responses: 296; Agree: 255
  - b. I have improved in my ability to understand my society and the world.

Valid responses: 298; Agree: 234

c. I have improved in my awareness and appreciation of cultures other than my own.

Valid responses: 297; Agree: 174

For each of questions (a) – (c), determine a point estimate for the proportion of graduates from this college who would agree with the statement. Then calculate a 95% confidence interval for the population proportion.

4. Rasmussen Reports conducted a national survey of 1,000 adults from June 19-20, 2013. The poll found that 63% of Americans think that a government that is too powerful is a bigger danger than one that is not powerful enough.

a. Use the information from the report to calculate a 95% confidence interval for the proportion of Americans who would agree with the statement above. Restate your confidence interval in terms of percentages. What is the margin of error for your confidence interval?

b. The report concluded with the following statement: The margin of error is  $\pm 3\%$  with a 95% level of confidence. Compare this statement with the margin of error you calculated in (a).

c. Was a sample size of 1,000 large enough to guarantee that the margin of error was less than 3% even if the sample percentage had been as low as 50% or as high as 80%? Explain.

d. How large a sample size was needed to guarantee that the margin of error was below 3% regardless of the sample proportion?