Topic: Confidence Intervals for Proportions
Activity: Reese’s Pieces

Prerequisites: This lesson typically follows confidence intervals for means.

Materials: Bag of Reese’s Pieces large enough for each student to have a sample of 25. Small cups for dipping out the sample. Paper towels on which to pour the sample. Access to the Internet to run a Java Applet called Reese’s Pieces.

Goals:

- To reinforce the concepts of population parameter and sample statistics for proportions.
- To use simulation to explore the sampling distribution of proportions.
- To understand from simulated and theoretical viewpoints the concept of confidence interval for proportions.

Background Information:

The goal of a confidence interval is to estimate a population parameter based on a sample statistic. All confidence intervals have the form: point estimate ± margin-of-error.

Example 1: Colors of Reese’s Pieces

Consider the population of the Reese’s Pieces candies manufactured by Hershey. Suppose that you want to learn about the distribution of colors of these candies but that you can only afford to take a sample of 25 candies.

a) Take a random sample of 25 candies and record the number and proportion of each color in your sample.

	orange	yellow	brown
number			
proportion			

b) Is the proportion of orange candies among the 25 that you selected a parameter or a statistic?

c) Is the proportion of orange candies manufactured by Hershey’s process a parameter or a statistic? What symbol represents it?

d) Do you know the value of the proportion of orange candies manufactured by Hershey?

e) Do you know the value of the proportion of orange candies among the 25 that you selected?

These simple questions point out the important fact that one typically knows (or can easily calculate) the value of a sample statistic, but only in very rare cases does one know the value of a population parameter. Indeed, a primary goal of sampling is to estimate the value of the parameter based on the statistic.

Used and modified with permission by Lunsford-Espy-Rowell project, supported by NSF

f) Do you suspect that every student in the class obtained the same proportion of orange candies in his/her sample?

g) Pool the results of the class to construct a dotplot of the sample proportions of orange candies.

h) Did everyone obtain the same number of orange candies in their samples?

i) If every student was to estimate the population proportion of orange candies by the proportion of orange candies in his/her sample, would everyone arrive at the same estimate?

j) Based on what you have learned about random sampling and having the benefit of seeing the sample results of the entire class, take a guess concerning the population proportion of orange candies.

k) Again assuming that each student had access only to her/his sample, would most estimates be reasonably close to the true parameter value? Would some estimates be way off? Explain.

l) In what way would the dotplot have looked different if each student had taken a sample of 10 candies instead of 25? (If unsure, you can check this using the Reese's Pieces applet found at http://statweb.calpoly.edu/chance/applets/applets.html.)

m) In what way would the dotplot have looked different if each student had taken a sample of 75 candies instead of 25? (If unsure, you can check this using the Reese's Pieces applet found at http://statweb.calpoly.edu/chance/applets/applets.html.)

Our class results suggest that even though sample values vary depending on which sample you happen to pick, there seems to be a *pattern* to this variation. We need more samples to investigate this pattern more thoroughly, however. Since it is time-consuming (and possibly fattening) to *literally* sample candies, we will use the computer to *simulate* the process.

To perform these simulations we need to suppose that we know the actual value of the parameter. Let us suppose that 45% of the population is orange.

n) Use a web browser to go to: http://statweb.calpoly.edu/chance/applets/applets.html. Click on "Reese's Pieces." Enter 25 for the sample size, 1 for the number of samples, and .45 for *p*, the population proportion of candies that are orange. Click on "draw samples" and watch the candies roll out. What proportion of the sample is orange? _____ Click on "draw samples" two more times. Did you get the same sample proportion of orange each time? ____

o) Now turn off the "Animate" button and change the number of samples to 500. Click on "draw samples" and watch the sample proportions of orange accumulate. (Pretend that this is really 500 students, each taking 25 candies.) Do you notice any pattern in the way that the resulting 500 sample proportions vary? Explain.

p) Record the **mean** and **standard deviation** of these 500 sample proportions.

q) Roughly speaking, are there more sample proportions close to the population proportion (which, you will recall, is .45) than there are far from it?

r) Let us quantify the previous question. Click on “count samples” and choose the “count between” option. Use the applet to count how many of the 500 sample proportions are within ±.10 of .45 (i.e., between .35 and .55). Then repeat for within ±.20 and for within ±.30 (you can enter new values or drag the red lines). Record the results below:

	number of the 500 sample proportions	percentage of these sample proportions
within ± .10 of .45		
within ± .20 of .45		
within ± .30 of .45		

s) Forget for the moment that you have designated that the population proportion of orange candies be .45. Suppose that each of the 500 imaginary students was to estimate the population proportion of orange candies by going a distance of .20 on either side of her/his sample proportion. What percentage of the 500 students would capture the actual population proportion (.45) within this interval?

t) Still forgetting that you actually know the population proportion of orange candies to be .45, suppose that you were one of those 500 imaginary students. Would you have any way of knowing definitively whether your sample proportion was within .20 of the population proportion? Would you be reasonably “confident” that your sample proportion was within .20 of the population proportion? Explain.

This pattern displayed by the variation of the sample proportions from sample to sample is the *sampling distribution* of the sample proportion. Even though the sample proportion of orange candies varies from sample to sample, there is a recognizable long-term pattern to that variation. Thus, while one cannot use a sample proportion to estimate a population proportion exactly, one can be reasonably confident that the population proportion is within a certain distance of the sample proportion. This “distance” depends primarily on how confident one wants to be and on the size of the sample.

u) Use the applet to simulate drawing 500 samples of 75 candies each (so these samples are three times larger than the ones you gathered in class and simulated earlier). Look at a display of the sample proportions and calculate their mean and standard deviation.

v) How has the sampling distribution changed from when the sample size was only 25 candies?

w) Use the applet to count how many of these 500 sample proportions are within ±.10 of .45. Record this number and the percentage below.

x) How do the percentages of sample proportions falling within ±.10 of .45 compare between sample sizes of 25 and 75?

y) In general, is a sample proportion more likely to be close to the population proportion with a larger sample size or with a smaller sample size?

Now we proceed to a ***theoretical analysis*** of this situation. Let the random variable X be the number of orange candies in a random sample of n candies. Also let $\hat{p} = X/n$ be the sample proportion of orange candies.

z) What probability distribution does X have, and what are its parameters?

aa) Determine the expected value of $\hat{p}$. Is $\hat{p}$ an unbiased estimator of the parameter p?

bb) Determine the variance and standard deviation of $\hat{p}$.

cc) Under what conditions is the probability distribution of $\hat{p}$ approximately normal? Explain.

This analysis suggests that an approximate 100(1-α)% confidence interval for a population proportion p would be: $\hat{p} \pm z_{\alpha/2}\sqrt{\frac{p(1-p)}{n}}$.

dd) What's the problem with this procedure? [*Hint*: Which of those pieces do you not know from the sample data?]

A simple solution is to approximate p by $\hat{p}$. This procedure gives an approximate 100(1-α)% confidence interval for a population proportion p as: $\hat{p} \pm z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$. This procedure is valid as long as the data are a random sample, $n\hat{p} \geq 10$, and $n(1-\hat{p}) \geq 10$.

ee) Use the data from your original sample of size 25 and the theoretical formula to find the 95% confidence interval for the percent of orange Reese's Pieces in each large bag of Reese's Pieces. __________________ Use a complete sentence to explain the resulting confidence interval in the context of the problem.

ff) How does this answer compare with your answer in part "*t*"?

gg) Identify what you have learned about finding confidence intervals for population proportions.