Lesson 9: Variability in Samples

- Let's explore how samples can be different.

9.1: Selecting Samples

Coins are usually stamped with the year and location of the mint where they were made. D represents the mint in Denver, Colorado, and a blank or P represents the mint in Philadelphia, Pennsylvania.

Diego has a jar containing 36 coins. Select a sample of 5 coins by rolling your number cube once to represent the row and rolling again to find the column. For example, rolling a 3 and then a 5 would represent selecting the coin marked 2000 P. Repeat this process to collect a sample of 5 coins.

	coin 1	coin 2	coin 3	coin 4	coin 5	sample mean year	sample proportion minted in Denver
sample 1							
sample 2							
sample 3							

1. Find the mean for the sample of 5 coins.

2. Find the proportion of the sample of 5 coins that were minted in Denver.

3. Repeat the process to find 2 more samples of 5 coins, then compute the mean and proportion that were minted in Denver.

9.2: Examining Sample Statistics

Use the data from the warm-up to answer the questions.

1. Not all the samples have the same mean and proportion. Why not?

2. Examine the histogram for the mean year of coins from the samples. What do you notice?

3. Based on the mean years from the samples, estimate the mean year for all the coins in Diego's jar. Explain your reasoning.

4. Examine the histogram of the proportion of coins in each sample that were minted in Denver. What do you notice?

5. Based on the sample proportions found by the class, estimate the proportion of coins minted in Denver for all the coins in Diego's jar. Explain your reasoning.

9.3: Variability of Sample Estimates

A political campaign sends volunteers out into the various parts of the state to get a sense of how well their candidate will do in an upcoming election. Thirty volunteers each get a random sample of 10 people in the state and find the proportion of people who are expecting to vote for their candidate. The sample proportions are summarized in the histogram.
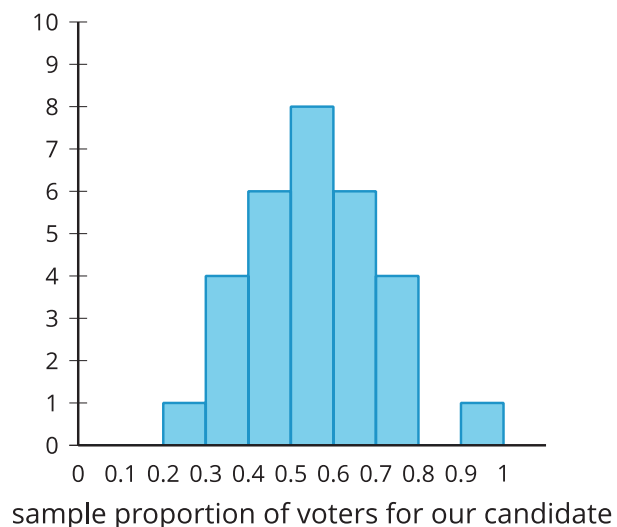
The mean of these sample proportions is 0.55, and the standard deviation is 0.15.

1. Recall that, for normally distributed data, about 95% of the data is within 2 standard deviations of the mean. What percentage of sample proportions are within 2 standard deviations of the mean for these data? Does this match what we expect from approximately normally distributed data? Explain or show your reasoning.

2. Estimates for population characteristics are usually given along with a **margin of error**. The margin of error is the maximum expected difference between the estimate of the population characteristic and the actual population characteristic. Each of the sample proportions are good estimates of the population proportion, so we should give a margin of error that contains about 95% of the sample proportions to be reasonably sure that the actual population proportion is in the range between the mean minus the margin of error to the mean plus the margin of error. What margin of error should be given along with the estimate of 0.55 for the population proportion? Explain or show your reasoning.

Are you ready for more?

The margin of error we constructed here was constructed at a 95% confidence level. Since 95% of the time our sample proportion is within 2 standard deviations of the true proportion for the population, by using a margin of error of 2 standard deviations we will capture the true proportion in our interval 95% of the time.

1. What would happen to the margin of error if we were okay with only capturing the true proportion 90% of the time?

2. What would happen to the margin of error if needed to capture the true proportion 99% of the time?

3. Why might someone choose a different confidence level?

Lesson 9 Summary

In many cases, it is difficult to collect data from an entire population, so using data from a small subset of the larger group is needed. The trade-off is that the incomplete information from such samples can only provide estimates of characteristics for the population.

For example, a researcher may wish to know how many fish of each type are present in a lake. It would be hard to collect all the fish in the pond to know all of the information, so a small group of fish might be caught to estimate the populations of the lake as a whole. Depending on how the fish are caught, the types of fish caught may be reflective of the entire lake or not.

To understand how varied the lake's fish are, the researcher may want to take several samples of fish from the lake. After taking many samples of fish, the researcher may find that the number of bass in a sample ranges from 2 to 20 and the number of catfish in the sample ranges from 5 to 7. Since the samples include a lot of different possibilities for the bass, the researcher might indicate that they have low confidence in an estimate for the bass population. On the other hand, the number of catfish in each sample is fairly consistent, so the researcher may be able to provide more confidence in an estimate for the catfish population in the lake.

To give a sense of the variability and confidence in estimates, a margin of error is usually given along with the estimate. A **margin of error** is the maximum expected difference between an estimate for a population characteristic and the actual value of the population characteristic. Estimates from samples tend to be approximately normal, so it is reasonable to expect that about 95% of the estimates are within 2 standard deviations of the mean from the estimates. In this unit, we will use 2 standard deviations as the margin of error.

From the fish example, the researcher may use the sample to estimate that there are 800 bass in the lake with a margin of error of 300 bass. This means that the researcher is fairly confident that the number of bass in the lake is somewhere between 500 and 1,100. The researcher may estimate that there are 650 catfish in the lake with a margin of error of 50. This means that the researcher is fairly confident that the number of catfish in the lake is between 600 and 700. Since the number of catfish in the samples was fairly consistent, there is a much smaller margin of error for the catfish population than there was for the bass population.