Unit 8, 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 then 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. The samples are drawn with replacement, which means we allow for the opportunity to draw the same coin more than once into our sample.

| | 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 of the years 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, and compute the mean year and the proportion that were minted in Denver for each sample.

# 9.2 Variability of Sample Estimates

A manufacturer is worried that their product may not be consistently good enough to pass quality control inspections. They are going to take a random sample of 10 of their products and have a quality control expert examine the items to determine if they pass or fail.

Your teacher will give you a bag with paper slips inside. 7 are marked “pass” and 3 are marked “fail.”

1. One partner should hold the bag so that the other partner cannot see inside while they draw a slip of paper. The other partner should draw out a slip of paper and record whether it says “pass” or “fail,” then return the slip of paper to the bag. Repeat this process until 10 slips are drawn.
2. From the results of the first trial, what proportion of the simulation sample are marked “pass”?

3. Switch roles with your partner, and repeat the process until you have run 5 trials. For each trial, compute the proportion of the simulated sample that are marked “pass.” Pause for all of the class’s simulated sample proportions to be collected.

| simulated sample | 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|
| number of "pass" slips | | | | | |
| proportion of sample that passes | | | | | |

4. Create a dot plot based on the simulated samples from the class that shows the proportion that is marked “pass.” This distribution is called a **sampling distribution**.

5. What number can be used to describe the variability of the sample distribution?

6. Estimate a range of values that captures about 95% of the values in the distribution.

## Are you ready for more?

The range of values you wrote captures about 95% of the values.

1. What would happen to the range if you wanted to capture only 90% of the values?

2. What would happen to the range if you needed to capture 99% of the values?

3. Why might someone choose different percentages to capture?

## 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, called a sample, is needed. The trade-off is that the incomplete information from samples can provide only estimates of characteristics for the population.

For example, an ecologist may wonder what proportion of trees in a particular large forest contains bird nests. It would be hard to look at every tree in the forest, so the researcher might take a random sample of 1,000 trees to find the proportion that have nests. Let's say the ecologist found that a proportion of 0.04 trees in the sample have nests. That is a good number to give as an estimate for the percentage of trees in the forest that have nests, but because the sample was randomly selected, it would probably not be surprising if 0.028 or 0.05 of the trees have nests.

To give a sense of the variability and confidence in estimates, a margin of error is usually given along with the point estimate. A **margin of error** is the maximum expected difference between a point estimate for a population characteristic and the actual value of the population characteristic. For means and proportions, the distribution of point estimates derived from samples, called the **sampling distribution**, tends to be approximately normal and centered around the actual population characteristic, so it is reasonable to expect that about 95% of the point estimates are within 2 standard deviations of the actual population characteristic. In this unit, we will use 2 standard deviations of the sampling distribution as the margin of error.

The ecologist could take the data from their sample and use a computer to simulate drawing, with replacement, thousands more samples of the same size from the original sample to create a sampling distribution. If the standard deviation for the sampling distribution is 0.006, then they can use a margin of error of 0.012 along with the estimate of 0.04 for the population proportion and report an estimate of $0.04 \pm 0.012$ for the proportion of trees in the forest that have nests. This means that any value in the range of 0.028 to 0.052 would not be surprising for the proportion of trees in the forest that have nests.