STAT 218 – Handout – Week 3 Lecture 3
Sampling Distributions

Recall the important terms **parameter** and **statistic**.

- A number describing a population is a **parameter**.
    - A parameter's value is fixed but typically not known.
    - Symbols: μ for population mean, σ for population std dev, π for population proportion
- A number describing a sample is a **statistic**.
    - A statistic's value can vary from sample to sample.
    - A statistic is often used to estimate a population parameter.
    - Symbols: $\bar{x}$ for sample mean, s for sample std dev, $\hat{p}$ for sample proportion

Example 22-1: Candy colors
Suppose that you take a random sample of 25 Reese's Pieces candies and record the number and proportion of each color: orange, yellow, brown.

a) What are the observational units and variable?

b) Is the candy's color a numerical or categorical variable?

c) Is the proportion of orange candies among your 25 candies a parameter or a statistic? What symbol should we use for it?

d) Is the proportion of orange candies manufactured by Hershey a parameter or a statistic? What symbol should we use for it?

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

f) Would you know the value of the proportion of orange candies among the 25 candies that you selected?

g) Would every student obtain the same proportion of orange candies in their sample?

h) If every student was to estimate the population proportion of orange candies by the proportion of orange candies in their sample, would everyone arrive at the same estimate?

- The values of statistics vary from sample to sample. This phenomenon is called **sampling variability**. Fortunately, if we look at the results of many samples, there is a predictable pattern to this variability.
    - Because random sampling is *unbiased*, the actual value of the population proportion should be close to the center of these sample proportions.

We will use an applet called "Reese's Pieces" to simulate selecting many random samples. For now we will suppose that 45% of the population is orange.

i) Use the "Reese's Pieces" applet to draw 10,000 samples of 25 candies each, assuming that the population proportion of orange is .45. (Pretend that this is 10,000 students, each taking 25 candies and counting the number of orange ones.) Describe the distribution of the sample proportions obtained.

j) Is there an obvious pattern to the distribution of the sample proportions of orange candies? Is it approximately normal?

- Even though the sample proportion of orange candies varies from sample to sample, there is a recognizable long-term pattern to that variation. This pattern is called the **sampling distribution** of the statistic.

k) What are the mean and standard deviation of the 10,000 simulated sample proportions of orange candies?

l) Now assume that the population proportion of orange candies is .55. What do you think will change in the sampling distribution? Again use the applet to draw 10,000 samples of 25 candies each. How has the distribution changed?

shape:

center:

variability:

m) How do you predict the distribution of sample proportions to change if we take many random samples of 100 candies rather than 25?

n) Use the applet to select 10,000 samples of 100 candies each. How has the distribution of sample proportions changed (or not changed) from when the sample size was only 25 candies?

shape:
center:
variability:

- A larger sample size produces less variability in sample statistics.

Central Limit Theorem (CLT) for Sample Proportion:

Suppose that the proportion of a population having some characteristic is denoted by π and suppose that a random sample of size n is taken from the population. Then the sampling distribution of the sample proportion $\hat{p}$ is approximately normal with mean π and standard deviation of $\sqrt{\frac{\pi(1-\pi)}{n}}$. This approximation is generally considered to be valid as long as $n\pi \geq 10$ and $n(1-\pi) \geq 10$.

o) Draw a sketch to represent this result, first in general and then applied to the candy example with $n = 25$ and assuming that the population proportion of orange is 0.45. Is this consistent with what the simulation revealed?

Example 22-2: Voter registration
According to the California Secretary of State (https://elections.cdn.sos.ca.gov/ror/ror-odd-year-2023/complete-ror.pdf), the number of Californians who were registered to vote as of February 10, 2023 was 21,980,768. This report also indicated that 22.5% of registered voters in California expressed no party preference (NPP) when they registered to vote.

a) Is this number 0.225 a parameter or a statistic? Also identify it with the appropriate symbol.

Now consider selecting a simple random sample of 400 registered voters in California and determining the proportion of them who expressed no party preference.

b) Is this number a parameter or a statistic? Also identify it with the appropriate symbol.

c) How would the sample proportion who expressed no party preference vary from sample to sample? Describe its shape, center, and variability. Also draw a well-labeled sketch to illustrate how this sample proportion would vary. Also check the conditions for whether the approximation is reasonable.

d) Determine the (approximate) probability that less than 20% of the sample would have expressed no party preference. (First shade the appropriate region in your sketch, then calculate the relevant z-score, and then use the normal probability table.)

e) Now suppose that you decide instead to take a simple random sample of 1000 (rather than 400) registered voters in California. Without doing any probability calculations, describe how your answer to the previous question will change. Explain your reasoning.