Unit 5: Sampling Distributions

5.1 - Why Is My Sample Not Like Yours?

- the same population can return different sample means, and some are more likely to happen than others
- the sample means will cluster around the true mean
  - one statistic cannot show a pattern for the whole population

> 📢 - This unit focuses on using sampling proportions to predict the pattern of data from a population

---

5.2 - The Normal Distribution, Revisited

- three ways to describe how often something can occur: proportion, percent, probability
- revisiting how to solve a z-score problem
  1. given the mean μ_x, standard deviation σ_x, and variable x
  2. $z = \frac{x-\mu_x}{\sigma_x}$ to solve for z-score
  3. plug in z-score to probability table to determine probability
  - include the defined random variable x, an example of how you used the normal distribution, the parameters μ_x and σ_x, the value of interest (or the value of x), and the correct probability
- find the difference between parameters $X \pm Y$
  - $\mu_{X\pm Y} = \mu_X \pm \mu_Y$ to calculate the difference of the means
  - $\sigma_{X\pm Y} = \sqrt{\sigma_X^2 + \sigma_Y^2}$ to calculate difference in standard deviation
    - convert to variances first because X and Y are independent
  - the resulting distribution will be normal if the distributions of both of the independent random variables are also approximately normal
- make graphs of a binomial distribution to determine if the distribution is normal
  - the larger the sample size, the closer the distribution will be to normal

> 📢 - Probability can be calculated by finding the z-score and plugging it into the z-score table
> - Larger sample sizes are more likely to be normal

5.3 - The Central Limit Theorem

- **sampling distribution:** distribution of values of all possible samples of the same size from a population
  - if the parent population is normal, then the sampling distribution will be approximately normal
  - the standard deviation of the sampling distribution depends on how big the sample size is, where the bigger the sample size, the smaller the standard deviation of the sampling distribution
- **central limit theorem:** if the sample size of a population is large enough, then the sampling distribution of the population will be normal regardless of the distribution of the original population
- perform random reallocation to calculate whether or not an answer value was calculated due to chance
  - given a control group and an experimental group
  1. take all values from the control and experimental group and redistribute to each group at random
  2. find the means of each group and calculate the difference of the means

3. the sampling distribution of the difference of the means will result in a normal graph, where you can convert your answer value to a z-score and then a probability

📢 - **Sampling distribution:** distribution of values of all possible samples of the same size from a population
- **Central limit theorem:** when the sample size is large enough, the sampling distribution of any population will be approximately normal
- Randomization distributions reallocate values to experimental groups to determine if an effect was due to chance

---

5.4 - Biased and Unbiased Point Estimates

- an estimator is unbiased if, on average, the value of the estimator is equal to the population parameter
  - if the population range is 10, then in order for the estimator to be unbiased, the average range of all of the sample statistic ranges should be 10
- **point estimator:** a sample statistic of the corresponding population parameter

📢 - **Point estimator:** a sample statistic of the corresponding population parameter
- An estimator is unbiased if the average value of the estimator is equation to the population parameter, so for example, if the population parameter is 5, then the average of all possible estimators should be 5 for the estimators to be unbiased

---

5.5 - Sampling Distributions for Sample Proportions

- **sampling variability:** samples won't always be the same probability when compared to the original population
- determining mean and standard deviation of sample distributions
  - given the sample proportion of successes in a random sample $\hat{p}$, the size of the sample n, and the population with a proportion of successes p
    - $\mu_{\hat{p}} = p$ to find the mean of the sample distributions
    - $\sigma_{\hat{p}} = \sqrt{\frac{p(1-p)}{n}}$ to find the standard deviation of the sample distributions ASSUMING that the sample size is less than 10% of the population OR that the samples are with replacement
    - the sampling distribution of $\hat{p}$ will be approximately normal when $np \geq 10$ and $n(1-p) \geq 10$
- interpreting the mean and standard deviation of sample distributions
  - interpret in context of the problem, and be sure to specify that when explaining what the mean or standard deviation means
  - mean: on average, from the sample size n, the sample proportions will have a mean of $\mu_{\hat{p}}$
  - standard deviation: on average, from the sample size n, the sample proportions would vary from the mean $\mu_{\hat{p}}$ by $\sigma_{\hat{p}}$
    - make sure to include "typically" or "on average" in the sentence

📢 - **Sample proportion:** the proportion of successes to non-successes taken from a random sample from a population
- The mean and standard deviation of a sampling distribution show how the *average sample statistics* sizes up in comparison to the rest. It doesn't show how each value compares to other values, but rather the sample groups that have been randomly selected

---

5.6 - Sampling Distributions for Differences in Sample Proportions

- you can subtract the sample $\hat{p}_2$ from the sample $\hat{p}_1$ to find the difference, and plot this on a sampling distribution graph
- determining the mean and standard deviation of the sampling distribution of differences between two sample proportions
  - given the sample proportion of successes $\hat{p}_1$ of sample size n_1 selected from population 1 with a proportion of successes p_1, AND a sample of proportion of successes $\hat{p}_2$ of sample size n_2 selected from population 2 with a proportion of successes p_2
    - $\mu_{\hat{p}_1-\hat{p}_2} = p_1 - p_2$ to find the mean of the sampling distribution of the differences
    - $\sigma_{\hat{p}_1-\hat{p}_2} = \sqrt{\frac{p_1(1-p_1)}{n_1} - \frac{p_2(1-p_2)}{n_2}}$ to find the standard deviation of the sampling distribution of the differences ASSUMING that the sample size is less than 10% of the population or were sampled with replacement AND that the two samples are independent of each other
    - the sampling distribution of $\hat{p}$ will be approximately normal when $n_1p_1 \geq 10$, $n_1(1-p_1) \geq 10$, $n_2p_2 \geq 10$, and $n_2(1-p_2) \geq 10$
- interpreting the mean and standard deviation of the sampling distribution of differences of sampling proportions
  - mean: on average, the difference between the samples $\hat{p}_1$ and $\hat{p}_2$ will be $\mu_{\hat{p}_1-\hat{p}_2}$
  - standard deviation: on average, the amount that the difference between the samples $\hat{p}_1$ and $\hat{p}_2$ varies from the mean $\mu_{\hat{p}_1-\hat{p}_2}$ is $\sigma_{\hat{p}_1-\hat{p}_2}$

📢 - The sampling distribution for a difference can help find the probability that one item occurs or less than another

---

5.7 - Sampling Distributions for Sample Means

- **sample mean:** a sample of individual items taken from a population and averaged
- determining the mean and standard deviation of sampling distributions of sample means
  - given the mean of a sample $\bar{x}$ where the sample is of size n, which is selected from a population with the mean μ and standard deviation σ
    - $\mu_{\bar{x}} = \mu$ to find the mean of the sampling distribution of sample means
    - $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$ to find the standard deviation of sampling distribution of sample means ASSUMING that the sample size is less than 10% of the population
    - the sampling distribution of $\bar{x}$ will be approximately normal if the population distribution is approximately normal OR if the sample size of the mean is greater than 30 (central limit theorem)
- interpreting the mean and standard deviation of the sampling distribution of sample means
  - mean: on average, a sample of size n from this distribution of sample means will have a mean of $\mu_{\bar{x}}$
  - standard deviation: on average, a sample of size n from this distribution of sample means will differ from the average sample mean $\mu_{\bar{x}}$ by $\sigma_{\bar{x}}$

📢 - **Sample mean:** the mean of a sample of individual items taken from a population
- The difference between a sample proportion and a sample mean is that a proportion is how often something is successful (binomial) while the mean is the average amount of times something occurs in the sample

---

5.8 - Sampling Distributions for Differences in Sample Means

- the sampling distribution will be made up of the sample mean of population 2 $\bar{x}_2$ subtracted from the sample mean of population 1 $\bar{x}_1$
- determining the mean and standard deviation of the sampling distribution of differences between two sample means
  - given the sample mean $\bar{x}_1$ of sample size n_1 selected from population 1 with a mean of μ_1 and a standard deviation of σ_1, AND a sample mean $\bar{x}_2$ of sample size n_2 selected from population 2 with a mean of μ_2 and standard deviation of σ_2
    - $\mu_{\bar{x}_1-\bar{x}_2} = \mu_1 - \mu_2$ to find the mean of the sampling distribution of the differences
    - $\sigma_{\bar{x}_1-\bar{x}_2} = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$ to find the standard deviation of the sampling distribution of the differences ASSUMING the sample size is less than 10% of the population AND that the two samples are independent
    - the sampling distribution of $\bar{x}_1 - \bar{x}_2$ will be approximately normal if both population distributions are approximately normal OR if both sample sizes are greater than 30 (central limit theorem)
- interpreting the mean and standard deviation of the sampling distribution of differences between two sample means
  - mean: on average, the difference between the samples $\bar{x}_1$ and $\bar{x}_2$ will be $\mu_{\bar{x}_1-\bar{x}_2}$
  - standard deviation: on average, the amount that the difference between samples $\bar{x}_1$ and $\bar{x}_2$ varies from the mean $\mu_{\bar{x}_1-\bar{x}_2}$ is $\sigma_{\bar{x}_1-\bar{x}_2}$

📢 - The sampling distribution for differences of sample means show how much the two samples vary from the rest
- Make sure to specify which sample is being subtracted from which