Unit 7: Inference for Quantitative Data - Means

7.1 - Should I Worry About Error?

- if the mean data of two groups is different, the null hypothesis would be that the difference occurred by chance, while the alternative hypothesis suggests that the difference was caused by something else
  - to find how likely it is for the difference to occur, assume that the difference occurred by chance and then re-randomize the groups and calculate the difference in means until there is sufficient evidence to make an estimate
  - to determine if the evidence is convincing, estimate the probability of getting as strong or stronger evidence by chance alone by looking at the resulting graph from the previous step

> 📢 - This unit is about looking at quantitative data and calculating if it is strong enough to support a claim

---

7.2 - Constructing a Confidence Interval for a Population Mean

- z critical values are not applicable to population means
  - if we have to estimate the population mean, then we shouldn't already know the population standard deviation
- to estimate a population mean, use a t critical value and a sample standard deviation
- when estimating the mean of a single population, use a one-sample t interval for a population mean
  - check the following conditions
    - the data was collected using a random sample from the population or a randomized experiment AND either the data was sampled with replacement OR that the sample size is less than 10% of the population
    - the shape of the sampling distribution should be approximately normal, meaning that $n \geq 30$ OR that if $n < 30$, the sample data is free of strong skewness and outliers
      - if a graph is not given, you have to create one to prove that the distribution is normal
- calculating margin of error
  - the standard error of the statistic is usually calculated with the standard deviation, but we don't have the standard deviation, so we use a sample standard deviation
  - $SE_{\bar{x}} = \frac{s}{\sqrt{n}}$ where s is the sample standard deviation
  - for the critical value, we have to use the value coming from a t distribution and not a z distribution
    - **t distributions:** a different type of distribution than the normal, 3 z-score z distribution
      - the mean of a t distribution is 0
      - the variability of a t distribution is determined by its degrees of freedom, and to estimate a population mean, use the equation $\mathrm{df} = n - 1$
        - as degrees of freedom increase, the tails get thinner and the peak gets higher
      - has thicker tails and a lower peak compared to a normal z distribution
    - critical value represents the boundaries encompassing the middle C% of the t distribution with $n - 1$ degrees of freedom, where C% is the approximate confidence level
- the confidence interval is calculated by $CI = \bar{x} \pm t^* \frac{s}{\sqrt{n}}$
- if your degrees of freedom isn't found in the table, choose the largest degree of freedom df that is still less than your desired value
  - if my value was 39 degrees, but there wasn't a probability slot of 39 degrees, I would round down to the next largest probability slot which is 30 degrees

📢 - Z critical values don't work for population means, use a t critical value instead
- Confidence interval is calculated by $CI = \bar{x} \pm t^* \frac{s}{\sqrt{n}}$
- When estimating the mean of a single population, use a *one-sample t interval for a population proportion*

7.3 - Justifying a Claim About a Population Mean Based on a Confidence Interval

- interpret a confidence interval with the sentence starter "we are C% confident that the interval from [lower bound] to [upper bound] captures the [population parameter]
  - if the confidence interval contains the value being questioned, then there is not enough evidence to dispute the null hypothesis
    - if I have a 95% confidence interval of 10 grams to 12 grams, and I was looking for evidence that all of the bags are less than 11 grams, there would not be enough evidence to prove that
  - confidence level is when there is repeated sampling with the sample size, then C% of the C% confidence intervals created will capture the population mean
- want small margin of errors so the confidence interval is small, which means we need bigger sample size or smaller confidence level
- steps to construct and interpret a confidence interval for a population mean
  1. define the parameter - C% confidence interval for $\mu =$ [mean parameter] and the procedure - one sample t interval for μ
  2. check that the sampling is random, the sample size is less than 10% of the population size, and that the data graph or plot shows no strong skewness or outliers
  3. calculate the confidence interval with the sample mean and the margin of error
  4. interpret the confidence interval - "we are C% confidence that the interval from [lower] to [upper] captures the [mean parameter]"

📢 - The confidence level represents the percentage of intervals that will include the true mean between their bounds

---

7.4 - Setting Up a Test for a Population Mean

- the alternative hypothesis, for population mean questions, are usually about how the mean doesn't equal the null value
- the null hypothesis is that the mean equals the null value
- when testing a claim about a population mean, use a one-sample t test for a population mean
  - check the following conditions
    - the data was collected using a random sample from the population or a randomized experiment AND either the data was sampled with replacement OR that the sample size is less than 10% of the population
    - the shape of the sampling distribution should be approximately normal, meaning that $n \geq 30$ OR that if $n < 30$, the sample data is free of strong skewness and outliers

📢 - For population means, the alternative hypothesis is usually that the mean for a sample population doesn't equal the expected mean
- When testing a claim about a population mean, use a *one-sample t test for a population mean*

7.5 - Carrying Out a Test for a Population Mean

- calculate a test statistic and the p value to figure out how likely it is to get evidence that is as strong or stronger than the alternative hypothesis
  - $t = \frac{\bar{x}-\mu}{\frac{s}{\sqrt{n}}}$ to calculate the test statistic using the t distribution
  - degrees of freedom df is equal to $n-1$ where n is the sample size
  - use $\mathrm{tcdf}(\ )$ on your calculator and plug in the values to calculate the p value OR use the table, which could be less accurate
- interpret a p value with the sentence starter "assuming [null hypothesis is true], there is a p probability of getting a sample mean of $\bar{x}$ or greater/less by chance alone in the random sample of n [context about sample size]"
- the smaller the p value, the less likely it is going to occur by random chance alone
  - depends on whether the p value is bigger or smaller than the significance level α, which is usually 0.05
- to analyze paired data, where two sets of data are part of the same pair, you have to find the difference between the data first and then find th mean difference
- steps to performing a significance test for a popultation mean
  1. determine the null hypothesis (that there was no difference between sample and expected mean) and the alternative hypothesis (that there was a difference)
  2. verify the procedure (one-sample t test for μ_d) and check that there was random sampling and no strong skewness or outliers
  3. calculated a standarized test statistic, the degrees of freedom, and then p value
  4. interpret the p value according to the significance level

📢 - Because you're using t distributions, calculate your p value using $\mathrm{tcdf}(\ )$ on your calculator
- Whether you reject or accept your null hypothesis depends on the significance level, which is usually $\alpha = 0.05$
- If you are given paired data, find the difference for every pair and state which way they are being subtracted

---

7.6 - Confidence Intervals for the Difference of Two Means

- when estimating a confidence level for a difference between two population means, use a two-sample t interval for the difference in population means
  - check the following conditions
    - the data was collected using two independent random samples OR a randomized experiment AND that the data was either sampled with replacement OR the sample sizes were less than 10% of the population
    - the shape of the sampling distribution should be approximately normal, meaning that $n \geq 30$ OR that if $n < 30$, the sample data is free of strong skewness and outliers
- $SE_{\bar{x}_1-\bar{x}_2} = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$ to calculate standard error
- $CI = (\bar{x}_1 - \bar{x}_2) \pm t^* \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$ where $\bar{x}_1 - \bar{x}_2$ is the point statistic because we weren't given one initially
  - to find the degrees of freedom you should use your calculator by going to $\mathrm{2\text{-}SampTInt}$ and choose No for pooled, and it'll calculate the interval and degrees
  - to calculate critical value go to $\mathrm{invT}(\ )$ and type degrees of freedom and area, which is how much of the tail you want to leave out on either side
    - for example, if I had a 95% confidence interval, I would enter my area as 0.025 because there is 2.5% of tail on either end of the distribution

- When estimating the confidence interval for a difference in means, use a *two-sample t interval for a difference in means*
- Use your calculator when finding degrees of freedom and the critical value for t intervals

7.7 - Justifying a Claim About the Difference of Two Means Based on a Confidence Level

- interpret a confidence interval with the sentence starter "we are C% confident that the interval from [lower bound] to [upper bound] captures the difference (1 - 2) [value to be estimated]"
  - when interpreting, don't say that it captures the difference for this specific sample, say that it captures the difference for all parameters
  - the confidence level is not a probability
- if the two groups had a difference of 0, then there is no convincing evidence that one group, on average, is greater than the other
- steps to construct and interpret a confidence interval for a difference in means
  1. define the confidence interval - C% confidence for (parameter 1 - parameter 2) and identify the method used - 2-sample t interval for (parameter 1 - parameter 2)
  2. check that the sampling is random, sample size is less than 10% of the population, and if the distribution is normal for both samples
  3. use Stat > Tests > 2-SampTInt and plug in your numbers to find the interval and degrees of freedom
  4. use DISTR > invT to find critical value and plug in your degrees of freedom and your bottom tail percent
  5. interpret the confidence interval - "we are C% confident that the interval from [lower bound] to [upper bound] captures the difference between (parameter 1 - parameter 2) [value to be estimated]"

- If zero is included in the confidence interval, then there is a chance that the difference between two means is 0

7.8 - Setting Up a Test for the Difference of Two Population Means

- a null hypothesis claims that there is no difference between the two means $\mu_1 = \mu_2$
- an alternative hypothesis claims that there is a difference between groups, whether less than or greater than, or just different from the original
- make sure that both of the samples and their populations fit the conditions to construct a significance test, which includes random sampling, no strong skewness, and more than 30 samples of data

- State the null and alternative hypothesis before you start calculating values for the significance test

7.9 - Carrying Out a Test for the Difference of Two Population Means

- calculate the standardized test statistic with the equation $t = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$
- use the 2-SampTTest formula on your calculator to calculate the p value and the degrees of freedom

- interpret a p value with the sentence starter "assuming the [null hypothesis is true], there is a p probability of getting a difference in means greater/less than [context] by chance alone. Since the p value of [p-value] is greater/less than the significance level [α], we (fail to) reject the H_0"
  - if the p value is small, that means the situation is less likely to occur by chance
- steps for performing a significance test for a different of two population means
  1. define each population mean and determine the null (no difference) and alternative (difference) hypotheses
  2. verify the procedure (two-sample t test for the difference in means) and check that there was random sampling and greater than 30 test subjects for both samples
  3. use Stat > Tests > 2-SampTTest and plug in your values to calculate the standardized test statisic, p value, and degrees of freedom
  4. interpret the p value according to the significance level

> 📢 - Because the formula for calculating standardized test statistic is so long, use your calculator instead by going to Stat > Tests > 2-SampTTest
> - If the p value is small, that means the situation is less likely to occur by chance

---

7.10 - Skills Focus, Selecting, Implementing, and Communicating Inference Procedures

- when there are two treatment groups, you hvae to use a two-sample t test or two-sample t interval because you need to find the difference between means
- if each volunteer goes through both treatments, then it is a matched pairs experiment
  - the data for each treatment is linked to one person, so you have to measure the difference for *each person* and take the average of all of the differences, as opposed to taking the average of both groups first
  - use a paired t test for one difference or a paired t interval

> 📢 - Paired data requires specific procedures for estimating intervals and performing significance tests
> - If each person involved in the experiment goes through more than one treatment, that means the data is dependent on that one person and is therefore paired