STAT 201 Chapter 10

Compare Two Samples

Difference of Proportions

- We're often interested in comparing two groups of **independent** statistics.

- Like, compare the difference of proportion of self-satisfaction between **male** and **female** student in 201
- Like, compare the difference of proportion of whether like Harry Potter between students sitting in **first to third row** with the one sitting in **forth to sixth row**

Sampling Distribution for Proportion Difference

- **The sample proportion difference** is the sample proportion of group 1 minus the sample proportion of group 2
  - $\hat{p}_{\mathrm{d}} = \hat{p}_1 - \hat{p}_2$
- **The population mean of the sample proportion difference** is the population proportion of group 1 minus the population proportion of group 2
  - $\mu_{p_d} = p_1 - p_2$

Sampling Distribution for Proportion Difference

- **The standard error, or the standard deviation of the sample proportion difference,** is seen below:
- $\sigma_{\hat{p}_d} = \sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}$
  - where $n_1 \; and \; n_2$ are the number of people in each group

Confidence Intervals for Proportion Difference

- $\hat{p}_{\mathrm{d}} = \hat{p}_1 - \hat{p}_2$
- $\sigma_{\hat{p}_d} = \sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}$

- A confidence interval for $\mu_{p_d} = p_1 - p_2$ is given by:

$$\hat{p}_{\mathrm{d}} \pm z\left(\sigma_{\hat{p}_d}\right)$$

  - Where z is found the same way as we did for one proportion

Confidence Intervals for Proportion Difference

- $\hat{p}_{\mathrm{d}} = \hat{p}_1 - \hat{p}_2$
- Confidence interval for $\mu_{p_d} = p_1 - p_2$ is given by:

$$\hat{p}_{\mathrm{d}} \pm z\left(\sigma_{\hat{p}_d}\right)$$

- If the resulting interval, (a, b), has both a and b greater than 0 this suggests that group 1 has the greater proportion.

Confidence Intervals for Proportion Difference

- $\hat{p}_{\mathrm{d}} = \hat{p}_1 - \hat{p}_2$
- Confidence interval for $\mu_{p_d} = p_1 - p_2$ is given by:

$$\hat{p}_{\mathrm{d}} \pm z\left(\sigma_{\hat{p}_d}\right)$$

- If the resulting interval, (a, b), has both a and b less than 0 this suggests that group 2 has the greater proportion.

Confidence Intervals for Proportion Difference

- $\hat{p}_{\mathrm{d}} = \hat{p}_1 - \hat{p}_2$
- Confidence interval for $\mu_{p_d} = p_1 - p_2$ is given by:

$$\hat{p}_{\mathrm{d}} \pm z\left(\sigma_{\hat{p}_d}\right)$$

- If the resulting interval, (a, b), contains 0 this suggests that there may be no difference between the two groups.

Example 1 Confidence Interval

- We randomly selected 100 students from South Carolina, where 50 are from USC and 50 are from Clemson. We are interested in the proportion difference of whether students think they are attractive by themselves in two universities.
- We want to find a confidence interval of the proportion difference of attractive students with 95% confidence level. $\mu_{p_d} = p_1 - p_2 = p_{USC} - p_C$

Example 1 Confidence Interval

- We have sample proportions for USC and Clemson attractive students: $\hat{p}_1 = 0.64; \hat{p}_2 = 0.42$
- $\hat{p}_{\mathrm{d}} = \hat{p}_1 - \hat{p}_2 = 0.64 - 0.42 = 0.22$
- $\sigma_{\hat{p}_d} = \sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}} =$

$$\sqrt{\frac{0.64(1-0.64)}{50} + \frac{0.42(1-0.42)}{50}} = 0.0974$$

Example 1 Confidence Interval

- $\hat{p}_{\mathrm{d}} = 0.22$
- $\sigma_{\hat{p}_d} = \ 0.0974$
- $\hat{p}_{\mathrm{d}} \pm z\left(\sigma_{\hat{p}_d}\right) = 0.22 \pm 1.96(0.0974) = (0.03, 0.41)$
- We see that the sample proportion difference is 0.22. This means that, in our sample, the proportion of attractive students at USC is higher than the proportion of attractive students at Clemson

Example 1 Confidence Interval

- $\hat{p}_{\mathrm{d}} = 0.22$
- $\sigma_{\hat{p}_d} = \ 0.0974$
- $\hat{p}_{\mathrm{d}} \pm z\left(\sigma_{\hat{p}_d}\right) = 0.22 \pm 1.96(0.0974) = (0.03, 0.41)$
- We are 95% confident that the true population proportion difference lies on the interval (0.03, 0.41). This means we are 95% confident that the proportion of attractive students at USC is higher than the proportion of attractive students at Clemson.

Example 1 Confidence Interval

- **If the 95% confidence interval is (-0.22,0.15)** , because 0 is contained in this interval, which means we are 95% confident that there is no difference between the proportion of attractive students at USC and the one at Clemson.

Example 1 Confidence Interval

- **If the 95% confidence interval is (-0.42,-0.03)** , we say that we are 95% confident that the proportion of attractive students at USC is lower than the proportion of attractive students at Clemson.

Hypothesis Test for Proportion Differences: Step 1

- State Hypotheses:
  - **Null hypothesis**:
    - $H_o: \mu_{p_d} = p_1 - p_2 \leq 0$ (right tailed test)
    - $H_o: \mu_{p_d} = p_1 - p_2 \geq 0$ (left tailed test)
    - $H_o: \mu_{p_d} = p_1 - p_2 = 0$ (two tailed test)
  - **Alternative hypothesis:** What we're interested in
    - $H_a: \mu_{p_d} = p_1 - p_2 > 0$ (right tailed test)
    - $H_a: \mu_{p_d} = p_1 - p_2 < 0$ (left tailed test)
    - $Ha: \mu_{p_d} = p_1 - p_2 \neq 0$ (two tailed test)

Hypothesis Test for Proportion Differences: Step 2

- Check the assumptions
  - The variable must be categorical
  - The data are obtained using randomization
  - We want at least 5 of each category within each group.
    - Five that have the attribute and five that don’t have the attribute for each group, at least

Hypothesis Test for Proportion Differences: Step 3

- Calculate Test Statistic
  - The test statistic measures how different the sample proportion we have is from the null hypothesis
  - We calculate the z-statistic by assuming that p_0 is the population proportion difference

$$z^* = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p}_{pool}\left(1 - \hat{p}_{pool}\right)\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

Hypothesis Test for Proportion Differences: Step 4

- Determine the P-value
  - The P-value describes how unusual the sample data would be if H_o were true.

Alternative Hypothesis	Probability	Formula for the P-value
$H_a: p_1 - p_2 > 0$	Right tailed	P(Z>z*)
$H_a: p_1 - p_2 < 0$	Left tailed	P(Z<z*)
$H_a: p_1 - p_2 \neq 0$	Two-tailed	2*P(Z<-|z*|)

Hypothesis Test for Proportion Differences: Step 5

- Summarize the test by reporting and interpreting the P-value
  - Smaller p-values give stronger evidence against H_o
- If p-value$\leq (1 - confidence) = \alpha$
  - Reject H_o, with a p-value = _____, we have sufficient evidence that the alternative hypothesis might be true
- If p-value$> (1 - confidence) = \alpha$
  - Fail to reject H_o, with a p-value = _____, we do not have sufficient evidence that the alternative hypothesis might be true

Example 1 Hypothesis Test: Step 1

- We randomly selected 100 students from South Carolina, where 50 are from USC and 50 are from Clemson. We are interested in the proportion difference of whether students think they are attractive by themselves, or not, in two universities.
- $H_o: p_{USC} - p_C = 0$ vs. $H_a: p_{USC} - p_C \neq 0$

Example 1 Hypothesis Test: Step 2

- Check the assumptions
  - Categorical variable: attractive, not attractive
  - Students are randomly selected
  - In USC group, 32 attractive, 18 not attractive
  - In Clemson group, 21 attractive, 29 not attractive
  - All 32, 18, 21, and 29 are larger than 5

Example 1 Hypothesis Test: Step 3

- Calculate the z-score

$$z^* = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p}_{pool}\left(1 - \hat{p}_{pool}\right)\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} = \frac{0.64 - 0.42}{0.0998} = 2.2$$

- $\hat{p}_1 = \frac{32}{50} = 0.64;\ \hat{p}_2 = \frac{21}{50} = 0.42;\ \hat{p}_{pool} = \frac{32+21}{50+50} = 0.53$

Example 1 Hypothesis Test: Step 4

- Calculate p-value. We have two-tail test here

$$P(Z < -|z^*|) = P(Z < -2.2) = 0.0139$$

$$p - value = 2 \times P(Z < -|z^*|) = 2 \times 0.0139 = 0.0278$$

Example 1 Hypothesis Test: Step 5

- Summarize the test by reporting and interpreting the P-value
- Since p-value=0.0278<0.05= α, we reject H_o and we have sufficient evidence that the alternative hypothesis might be true
- This means we are 95% confident that the population proportion of students who think themselves are attractive in USC and Clemson are different.

Example 1 Hypothesis Test

- What if we want 99% confidence?
- Since p-value=0.0278 > 0.01= α, we fail to reject H_o and we do not have sufficient evidence that the alternative hypothesis might be true
- This means we are 99% confident that the proportion of students who think themselves are attractive in USC and Clemson are the same.

Example 1

- We can also make the number of observations of two groups different, such as randomly sample 70 USC students and 40 Clemson students.
- We set $n_1 = 70, n_2 = 40$ and follow the same procedure to find the related confidence interval and do the hypotheses tests

Difference of Means

- We're often interested in comparing means of two groups of data.

- Let's assume two groups are **independent**!!

Sampling Distribution for Mean Difference

- **The sample mean difference** is the sample mean of group 1 minus the sample mean of group 2
  - $\overline{x_d} = \bar{x}_1 - \bar{x}_2$
- **The population mean of the sample mean differences** is the population mean of group 1 minus the population mean of group 2
  - $\mu_d = \mu_1 - \mu_2$

Sampling Distribution for Mean Difference

- **The standard error, or the standard deviation of the sample mean difference,** is seen below:
- $s_d = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$
  - where n_1 & n_2 are the number of people in each group and s_1^2 & s_2^2 are the sample variance for each group

Confidence Intervals for Mean Difference

- $\bar{x}_{\mathrm{d}} = \bar{x}_1 - \bar{x}_2;\ s_d = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$
- A confidence interval is given by:

$$\bar{x}_{\mathrm{d}} \pm t_{\frac{\alpha}{2},df} \left(\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} \right)$$

  - $df = n_1 + n_2 - 2$

Confidence Intervals for Mean Difference

- A confidence interval is given by:

$$\bar{x}_{\mathrm{d}} \pm t_{\frac{\alpha}{2},df} \left(\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} \right)$$

- If the resulting interval, (a, b), has both a and b greater than 0, this suggests that group 1 has the larger mean

Confidence Intervals for Mean Difference

- A confidence interval is given by:

$$\bar{x}_{\mathrm{d}} \pm t_{\frac{\alpha}{2},df} \left(\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} \right)$$

- If the resulting interval, (a, b), has both a and b smaller than 0, this suggests that group 2 has the larger mean

Confidence Intervals for Mean Difference

- A confidence interval is given by:

$$\bar{x}_{\mathrm{d}} \pm t_{\frac{\alpha}{2},df} \left(\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} \right)$$

- If the resulting interval, (a, b), contains 0 this suggests that there may be no difference between the two groups

Example 2 Confidence Interval

- We are interested in whether there is a difference in the mean GPA between male and female undergrad students in USC Stat department.
- We randomly selected 82 students, where 38 are boys and 44 are girls.
- They are independent and we set boys into group 1 and girls into group 2.
- We want to find the 95% confidence interval of the mean difference.

Example 2 Confidence Interval

- $\bar{x}_1 = 3.12$ (boys) ; $\bar{x}_2 = 3.26$ (girls)
- $\bar{x}_{\mathrm{d}} = \bar{x}_1 - \bar{x}_2 = 3.12 - 3.26 = -0.14$
- $s_1 = 0.19$ (boys) ; $s_2 = 0.27$ (girls)
- $\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} = \sqrt{\frac{0.19^2}{38} + \frac{0.27^2}{44}} = 0.051$
- $t_{\frac{\alpha}{2},df} = t_{\frac{0.05}{2},38+44-2} = 1.99$

Example 2 Confidence Interval

- $\bar{x}_{\mathrm{d}} = -0.14\ ;\ \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} = 0.051\ ;t_{\frac{\alpha}{2},df} = 1.99$

- $\bar{x}_{\mathrm{d}} \pm t_{\frac{\alpha}{2},df}\left(\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}\right) = -0.14 \pm 1.99 \times 0.051$

$$= (-0.24, -0.04)$$

Example 2 Confidence Interval

- $\bar{x}_{\mathrm{d}} \pm t_{\frac{\alpha}{2},df}\left(\sqrt{\frac{s_1^2}{n_1}+\frac{s_2^2}{n_2}}\right) = (-0.24, -0.04)$
- We are 95% confident that the true mean difference lies on the interval (-0.24, -0.04). This means we have 95% confidence that the mean GPA of female students is greater than the mean GPA of male students in USC stat department.

Example 2 Confidence Interval

- What if our confidence interval is (0.12, 0.29)
- We are 95% confident that the true mean difference lies on the interval (0.12, 0.29). This means we have 95% confidence that the mean GPA of male students is greater than the mean GPA of female students in USC stat department.

Example 2 Confidence Interval

- What if our confidence interval is (-0.02, 0.17)
- We are 95% confident that the true mean difference lies on the interval (-0.02, 0.17). This means we have 95% confidence that there exits no difference between the mean GPA of male and female students in USC stat department.

Hypothesis Test for Mean Difference: Step 1

- State Hypotheses:
  - **Null hypothesis**: that the population mean equals some μ_o
    - $H_o: \mu_d = \mu_1 - \mu_2 \leq 0$ (right tailed test)
    - $H_o: \mu_d = \mu_1 - \mu_2 \geq 0$ (left tailed test)
    - $H_o: \mu_d = \mu_1 - \mu_2 = 0$ (two tailed test)
  - **Alternative hypothesis:** What we're interested in
    - $H_a: \mu_1 - \mu_2 > 0$ (right tailed test)
    - $H_a: \mu_1 - \mu_2 < 0$ (left tailed test)
    - $Ha: \mu_1 - \mu_2 \neq 0$ (two tailed test)

Hypothesis Test for Mean Difference: Step 2

- Check the assumptions
  - The difference variable must be quantitative
  - The data are obtained using randomization
  - The sample size is large enough for normality assumption
    - $n_1 > 30$
    - $n_2 > 30$

Hypothesis Test for Mean Difference: Step 3

- Calculate Test Statistic
  - The test statistic measures how different the sample proportion we have is from the null hypothesis
  - We calculate the z-statistic by assuming that μ_0 is the population mean difference

$$t^* = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

Hypothesis Test for Mean Difference: Step 4

- Determine the P-value
  - The P-value describes how unusual the data would be if H_o were true.
  - We will use software or your calculator to find this, or I will give it to you.

Alternative Hypothesis	Probability	Formula for the P-value
$H_a: \mu_1 - \mu_2 > 0$	Right tail	P(T>t*)
$H_a: \mu_1 - \mu_2 < 0$	Left tail	P(T<t*)
$H_a: \mu_1 - \mu_2 \neq 0$	Two-tail	2*P(T<-|t*|)

Hypothesis Test for Mean Difference: Step 5

- Summarize the test by reporting and interpreting the P-value
  - Smaller p-values give stronger evidence against H_o
- If p-value$\leq (1 - confidence) = \alpha$
  - Reject H_o, with a p-value = _____, we have sufficient evidence that the alternative hypothesis might be true
- If p-value$> (1 - confidence) = \alpha$
  - Fail to reject H_o, with a p-value = _____, we do not have sufficient evidence that the alternative hypothesis might be true

Example 2 Hypotheses Testing

- We are interested in whether there is a difference in the mean GPA between male and female undergrad students in USC Stat department.
- We randomly selected 82 students, where 38 are boys and 44 are girls.
- They are independent and we set boys into group 1 and girls into group 2

Example 2 Hypotheses Testing: Step 1

- Test whether or not mean GPA of female students is higher with 95% confidence level.
- $H_o: \mu_d = \mu_1 - \mu_2 \geq 0$
- $H_a: \mu_d = \mu_1 - \mu_2 < 0$

Example 2 Hypotheses Testing: Step 2

- The variable, mean GPA, is quantitative
- The data are obtained using randomization
- The sample size is large enough for normality assumption
  - $n_1 = 38 > 30$
  - $n_2 = 44 > 30$

Example 2 Hypotheses Testing: Step 3

- Calculate Test Statistic
- $t^* = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} = \frac{-0.14}{0.051} = -2.74$

Example 2 Hypotheses Testing: Step 4

- Determine the P-value

- $p - value = P(T < t^*) = P(T < -2.74) = 0.0038$
- With degree of freedom 82-2=80

- We find this value using StatCrunch

Example 2 Hypotheses Testing: Step 5

- With a p-value 0.0038 < 0.05, we reject the null hypothesis and say we are 95% confident that population mean GPA of female students is higher than the population mean GPA of male students in USC stat department

Difference of Groups: Independent vs. Dependent

- **Independent Samples** – when the observations in each group are randomly selected and randomly placed
  - Up until now we always assume this
- **Dependent Samples** – when the groups are made up of pairs. For instance you can take married couples and make one group the males and one group the females. We can use Statcrunch to do the calculation for dependent samples.