5/10/12
- Warm up by finishing example 2 from 9.2
- Continue Ch. 9

Tuesday 5/15
- Finish Ch. 9
- Extra credit
- SLO assessment
- Project DUE

Thursday 5/17
- Review for Final

Tuesday 5/22
- Final exam from 5-7 PM
- Ch. 9, 10 homework due

a. Use a 0.01 significance level to test the claim that the female survivors and male survivors have different rates of thyroid diseases.

\[ \hat{p}_1 = \frac{1397}{2739} \approx 0.510, \quad n_1 = 2739 \]
\[ \hat{p}_2 = \frac{436}{1352} \approx 0.322, \quad n_2 = 1352 \]

\[ \bar{p} = \frac{1397 + 436}{2739 + 1352} \approx 0.448 \]

\[ \bar{q} = 0.552 \]

\[ Z = \frac{0.510 - 0.322 - 0}{\sqrt{\frac{(0.448)(0.552)}{2739} + \frac{(0.448)(0.552)}{1352}}} \approx 11.37 \]

(1) \( H_0: \hat{p}_1 = \hat{p}_2 \)
(2) \( Z = \frac{0.510 - 0.322 - 0}{\sqrt{\frac{(0.448)(0.552)}{2739} + \frac{(0.448)(0.552)}{1352}}} \approx 11.37 \)
(3) \( P\)-value is:
\[ 2 \cdot P(Z > 11.37) = 2 \left[ 1 - P(Z < 11.37) \right] = 0.0002 \]
(4) There is sufficient evidence to support the claim that female and male survivors have different rates of thyroid disease.

b. Construct the confidence interval corresponding to the hypothesis test conducted with a 0.01 significance level.

\[ E = 2.575 \sqrt{\frac{(0.510)(0.490)}{2739} + \frac{(0.322)(0.678)}{1352}} \]

\[ E \approx 0.0409 \]

\[ (0.510 - 0.322) - 0.0409 \leq \hat{p}_1 - \hat{p}_2 \leq (0.510 - 0.322) + 0.0409 \]
\[ 0.147 \leq \hat{p}_1 - \hat{p}_2 \leq 0.229 \]

The CI suggests that the rates are different since zero is not included.

c. What conclusion does the confidence interval suggest?
9.3 INFERENCES ABOUT TWO MEANS: INDEPENDENT SAMPLES

Key Concept...

In this section, we present methods for using sample data from 2 independent samples to test hypotheses made about 2 population means or to construct confidence interval estimates of the difference between 2 population means.

PART I: INDEPENDENT SAMPLES WITH $\sigma_1$ AND $\sigma_2$ UNKNOWN AND NOT ASSUMED EQUAL

DEFINITION

Two Samples are independent if the Sample values from one population are not related to or somehow naturally paired or matched with the Sample values from the other population.

Two Samples are dependent if the sample values are paired.
Inferences about Means of Two Independent Populations, With $\sigma_1$ and $\sigma_2$ Unknown and Not Assumed to be Equal

**OBJECTIVES**

Test a claim about 2 independent pop. means or construct a CI estimate of the difference between 2 independent pop. means.

**NOTATION**

Population 1:

$\mu_1 =$ population mean  
$s_1 =$ sample standard deviation  
$\sigma_1 =$ population standard deviation  
$x_1 =$ sample mean  
$n_1 =$ size of the first sample

The corresponding notations for $\mu_2$, $\sigma_2$, $\bar{x}_2$, $s_2$, and $n_2$ apply to population 2.

**REQUIREMENTS**

1. $\sigma_1$ and $\sigma_2$ are unknown and it is not assumed that $\sigma_1$ and $\sigma_2$ are equal.

2. The 2 samples are independent.

3. Both samples are simple random samples.

4. Either or both of these conditions are satisfied: The two sample sizes are both large (with $n_1 > 30$ and $n_2 > 30$) or both
samples come from populations having \textbf{normal} distributions.

**HYPOTHESIS TEST STATISTIC FOR TWO MEANS: INDEPENDENT SAMPLES**

\[ t = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} \]

where \( \mu_1 - \mu_2 \) are assumed to be zero under the null

**Degrees of Freedom:** When finding \( \text{critical values} \) or \( \text{p-values} \), use the following for determining the number of degrees of freedom.

1. In this book we use the conservative estimate: \( df = \text{smaller of } n_1 - 1 \) and \( n_2 - 1 \).

2. Statistical software packages typically use the more accurate but more difficult estimate given below:

\[ df = \frac{(A + B)^2}{\frac{A^2}{n_1 - 1} + \frac{B^2}{n_2 - 1}} \]

\[ A = \frac{s_1^2}{n_1}, \quad B = \frac{s_2^2}{n_2} \]

**P-values and critical values:** Use Table A-3.

**CONFIDENCE INTERVAL ESTIMATE OF \( \mu_1 - \mu_2 \): INDEPENDENT SAMPLES**

The confidence interval estimate of the difference \( \mu_1 - \mu_2 \) is

\[ (\bar{x}_1 - \bar{x}_2) - E < \mu_1 - \mu_2 < (\bar{x}_1 - \bar{x}_2) + E \]

where \( E = t_{\alpha/2} \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} \)

and the number of degrees of freedom \( df \) is as described above for hypothesis tests.
EQUIVALENCE OF METHODS

The __________ method of hypothesis testing, the __________ method of hypothesis testing, and ______________ intervals all use the same ______________ and ______________, so they are ______________ in the sense that they result in the ______________. A null hypothesis of ______________ or ______________ can be tested using the __________ method, the ______________ method, or by determining whether the ______________ includes ______________.

Example 1: Determine whether the samples are independent or dependent.

a. To test the effectiveness of Lipitor, cholesterol levels are measured in 250 subjects before and after Lipitor treatments.

Dependent

b. On each of 40 different days, the author measured the voltage supplied to his home and he also measured the voltage produced by his gasoline powered generator.

Independent

Example 2: Assume that the two samples are independent simple random samples selected from normally distributed populations. Do not assume that the population
standard deviations are equal. A simple random sample of 13 four-cylinder cars is obtained, and the braking distances are measured. The mean braking distance is 137.5 feet and the standard deviation is 5.8 feet. A SRS of 12 six-cylinder cars is obtained and the braking distances have a mean of 136.3 feet with a standard deviation of 9.7 feet (based on Data Set 16 in Appendix B).

a. Construct a 90% CI estimate of the difference between the mean braking distance of four-cylinder cars and six-cylinder cars.

\[ \bar{x}_1 = 137.5, \quad s_1 = 5.8, \quad n_1 = 13 \]
\[ \bar{x}_2 = 136.3, \quad s_2 = 9.7, \quad n_2 = 12 \]
\[ E = 1.796 \sqrt{\frac{5.8^2}{13} + \frac{9.7^2}{12}} \]
\[ E \approx 5.8 \]

\[ (137.5 - 136.3) - 5.8 < \mu_1 - \mu_2 < (137.5 - 136.3) + 5.8 \]
\[ -4.6 < \mu_1 - \mu_2 < 7.0 \]

b. Does there appear to be a difference between the two means?

No → the CI includes zero.

c. Use a 0.05 significance level to test the claim that the mean braking distance of four-cylinder cars is greater than the mean braking distance of six-cylinder cars.

1. \( H_0: \mu_1 = \mu_2 \)
2. \( H_A: \mu_1 > \mu_2 \)
3. Critical value test (traditional)
   \[ t = \frac{137.5 - 136.3 - 0}{\sqrt{\frac{5.8^2}{13} + \frac{9.7^2}{12}}} \]
   \[ t \approx 0.372 \]
4. There is not sufficient evidence to support the claim that the mean braking distance of four-cylinder cars is greater than six-cylinder cars.
PART 2: ALTERNATIVE METHODS

Part 1 in this section dealt with situations in which the two population standard deviations are unknown and not assumed to be equal. In Part 2 we address two other situations: (1) The two population standard deviations are both known; (2) the two population standard deviations are unknown but assumed to be equal.

ALTERNATIVE METHOD WHEN $\sigma_1$ AND $\sigma_2$ ARE KNOWN

In reality, the population standard deviations are almost never known, but if they are known, the test statistic and confidence interval are based on the normal distribution instead of the $t$ distribution.

Inferences about Means of Two Independent Populations, With $\sigma_1$ and $\sigma_2$ Known

REQUIREMENTS

1. The two population standard deviations $\sigma_1$ and $\sigma_2$ are both known.

2. The two samples are independent.

3. Both samples are simple random samples.
4. Either or both of these conditions is satisfied: The two sample sizes are both large, with \( n_1 > 30 \) and \( n_2 > 30 \) or both samples come from populations having normal distributions.

**HYPOTHESIS TEST**

Test statistic: 
\[
Z = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}, \text{ where } \mu_1 - \mu_2 = 0 \text{ under the null}
\]

- **P-values** and **critical values**: Refer to Table A2.

**CONFIDENCE INTERVAL ESTIMATE OF \( \mu_1 - \mu_2 \)**

\[
(\bar{x}_1 - \bar{x}_2) - E < \mu_1 - \mu_2 < (\bar{x}_1 - \bar{x}_2) + E
\]

\[
E = Z_{\alpha/2} \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}
\]

**ALTERNATIVE METHOD: ASSUME THAT \( \sigma_1 = \sigma_2 \) AND POOL THE SAMPLE VARIANCES**

Even when the specific values of \( \sigma_1 \) and \( \sigma_2 \) are not known, if it can be assumed that they have the same value, the sample variances \( s_1^2 \) and \( s_2^2 \) can be pooled to obtain an estimate of the common population variance \( \sigma^2 \). The pooled estimate of \( \sigma^2 \) is denoted \( s_p^2 \) and
Inferences about Means of Two Independent Populations, Assuming that $\sigma_1 = \sigma_2$

**REQUIREMENTS**

1. The two population standard deviations are **not known**, but they are assumed to be **equal**. That is, $\sigma_1 = \sigma_2$.

2. The two **samples** are **independent**.

3. Both samples are **simple random samples**.

4. Either or both of these conditions is satisfied: The two sample sizes are both large, with $n_1 > 30$ and $n_2 > 30$ or both samples come from **populations** having **normal distributions**.

**HYPOTHESIS TEST**

Test statistic: 
$$t = \frac{\bar{x}_1 - \bar{x}_2 - (\mu_1 - \mu_2)}{s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

where $s_p$ is calculated as:

$$s_p^2 = \frac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{(n_1-1)+(n_2-1)}$$

(pooled variance)

and the number of degrees of freedom is given by $n_1 + n_2 - 2$. 

The null hypothesis $\mu_1 - \mu_2 = 0$ is tested under these conditions.
CONFIDENCE INTERVAL ESTIMATE OF $\mu_1 - \mu_2$

Confidence interval:

$$(\bar{x}_1 - \bar{x}_2) - E < \mu_1 - \mu_2 < (\bar{x}_1 - \bar{x}_2) + E$$

where

$$E = t_{\alpha/2} \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

and $s_p^2$ is as given in the above test statistic and the number of degrees of freedom is $n_1 + n_2 - 2$.

**TI-83/84 PLUS**

```
STAT
TESTS
2-SAMPTest (hypothesis test)

or

2-SAMPInt (CI)
```
Example 3: Assume that the two samples are independent simple random samples selected from normally distributed populations. Also assume that the population standard deviations are equal. The mean tar content of a simple random sample of 25 unfiltered king size cigarettes is 21.1 mg, with a standard deviation of 3.2 mg. The mean tar content of a SRS of 25 filtered 100 mm cigarettes is 13.2 mg with a standard deviation of 3.7 mg (based on Data Set 4 in Appendix B).

a. Construct a 90% CI estimate of the difference between mean tar content of unfiltered king size cigarettes and the mean tar content of filtered 100 mm cigarettes.

\[
\bar{x}_1 = 21.1, \ S_1 = 3.2, \ n_1 = 25 \\
\bar{x}_2 = 13.2, \ S_2 = 3.7, \ n_2 = 25 \\
\]

\[
d.f.: 25 + 25 - 2 = 48 \\
S_p^2 = \frac{24 (3.2)^2 + 24 (3.7)^2}{24 + 24} = 11.965 \\
E = 1.676 \sqrt{\frac{11.965}{25} + \frac{11.965}{25}} = 1.6 \\
(21.1 - 13.2) - 1.6 < \mu_1 - \mu_2 < (21.1 - 13.2) + 1.6 \\
6.3 < \mu_1 - \mu_2 < 9.5
\]

b. Does the result suggest that 100 mm filtered cigarettes have less tar than unfiltered king size cigarettes?

Yes - the CI estimate of the mean difference is positive.

c. Use a 0.05 significance level to test the claim that unfiltered king size cigarettes have a mean tar content greater than that of filtered 100 mm cigarettes. What does the result suggest about the effectiveness of cigarette filters?

1. \( H_0: \mu_1 = \mu_2 \) \\
2. \( H_A: \mu_1 > \mu_2 \) \\
3. \( p\text{-value} < 0.005 \rightarrow 0.005 < 0.05 \) \( \rightarrow \) reject \( H_0 \) \\
4. There's sufficient evidence to support the claim that unfiltered king size cigarettes have a mean tar content greater than that of filtered 100 mm cigarettes.
9.4 INFERENCES FROM DEPENDENT SAMPLES

Key Concept...

In this section we present methods for testing hypotheses and constructing confidence intervals involving the ______________ of the ______________ of the ______________ of two ______________ ______________.

With ______________ samples, there is some ______________ whereby each value in one sample is ______________ with a ______________ value in the other sample. Here are two typical examples of dependent samples:

π Each pair of sample values consists of two measurements from the ______________ subject

π Each pair of sample values consists of a ______________ ______________.

Because the hypothesis test and CI use the same ______________ and ______________, they are ______________ in the sense that they result in the ______________ ______________. Consequently, the ______________ hypothesis that the ______________ ______________ ______________ can be tested by determining whether the ______________ ______________ ______________ includes _____.

There are no exact procedures for dealing with ______________
samples, but the _____ __________________ serves as a reasonably good approximation, so the following methods are commonly used.

### Inferences about Means of Two Dependent Populations

**OBJECTIVES**

**NOTATION**

<table>
<thead>
<tr>
<th>Symbol</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>( d )</td>
<td>( s_d )</td>
</tr>
<tr>
<td>( \mu_d )</td>
<td></td>
</tr>
<tr>
<td>( \bar{d} )</td>
<td>( n )</td>
</tr>
</tbody>
</table>

**REQUIREMENTS**

1. The __________ data are __________________.  
2. The samples are ______________ _______________ _______________.  
3. Either or both of these conditions are satisfied: The number of __________ of ____________ is ____________ (__________) or the pairs of values have ________________ that are from a population that is approximately _____________.

**HYPOTHESIS TEST FOR DEPENDENT SAMPLES**

\[ t = \]
Example 1: Assume that the paired sample data are SRSs and that the differences have a distribution that is approximately normal.

a. Listed below are BMIs of college students.

<p>| | | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>April</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>BMI</td>
<td>20.15</td>
<td>19.24</td>
<td>20.77</td>
<td>23.85</td>
</tr>
<tr>
<td>September</td>
<td>20.68</td>
<td>19.48</td>
<td>19.59</td>
<td>24.57</td>
</tr>
<tr>
<td>BMI</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

i. Use a 0.05 significance level to test the claim that the mean change in BMI for all students is equal to 0.
ii. Construct a 95% CI estimate of the change in BMI during freshman year.

iii. Does the CI include zero, and what does that suggest about BMI during freshman year?

b. Listed below are systolic blood pressure measurements (mm Hg) taken from the right and left arms of the same woman. Use a 0.05 significance level to test for a difference between the measurements from the two arms. What do you conclude?

<table>
<thead>
<tr>
<th>Right arm</th>
<th>102</th>
<th>101</th>
<th>94</th>
<th>79</th>
<th>79</th>
</tr>
</thead>
<tbody>
<tr>
<td>Left arm</td>
<td>175</td>
<td>169</td>
<td>182</td>
<td>146</td>
<td>144</td>
</tr>
</tbody>
</table>