Section 10.2
Comparing Two Means
Section 10.2
Comparing Two Means

After this section, you should be able to...

✓ DESCRIBE the characteristics of the sampling distribution of the difference between two sample means

✓ CALCULATE probabilities using the sampling distribution of the difference between two sample means

✓ DETERMINE whether the conditions for performing inference are met

✓ USE two-sample \( t \) procedures to compare two means based on summary statistics or raw data

✓ INTERPRET computer output for two-sample \( t \) procedures

✓ PERFORM a significance test to compare two means

✓ INTERPRET the results of inference procedures
Theory: The Sampling Distribution of a Difference Between Two Means

The sampling distribution of a sample mean has the following properties:

**Shape** Approximately Normal if the population distribution is Normal or \( n \geq 30 \) (by the central limit theorem).

**Center** \( \mu_x = \mu \)

**Spread** \( \sigma_x = \frac{\sigma}{\sqrt{n}} \) if the sample is no more than 10% of the population.
Theory: The Sampling Distribution of a Difference Between Two Means

Using software, we generated an SRS of 12 girls and a separate SRS of 8 boys and calculated the sample mean heights. The difference in sample means was then calculated and plotted. We repeated this process 1000 times. The results are below:

What do you notice about the shape, center, and spread of the sampling distribution of $\bar{x}_f - \bar{x}_m$?
Sampling distribution of $\bar{x}_1 - \bar{x}_2$

Standard deviation

$$\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

Mean $\mu_1 - \mu_2$

Values of $\bar{x}_1 - \bar{x}_2$
Formula: **Two Mean T-Interval**

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

The degrees of freedom is determined by smaller of \(n_1 - 1\) and \(n_2 - 1\).
Conditions: Two Mean T-Interval

1) **Random**: Both sets of data should come from a well-designed random samples or randomized experiments.

2) **Normal**: Both sets of data must meet the Central Limit Theorem (CLT) with sample sizes greater than 30 or graph values that are less than 30 to check normality.

3) **Independent**: Both sets of data must be independent. When sampling without replacement, the sample size $n$ should be no more than 10% of the population size $N$ (the 10% condition). Must check the condition for each separate sample.
Big Trees, Small Trees, Short Trees, Tall Trees

The Wade Tract Preserve in Georgia is an old-growth forest of longleaf pines that has survived in a relatively undisturbed state for hundreds of years. One question of interest to foresters who study the area is “How do the sizes of longleaf pine trees in the northern and southern halves of the forest compare?” To find out, researchers took random samples of 30 trees from each half and measured the diameter at breast height (DBH) in centimeters. Comparative boxplots of the data and summary statistics from Minitab are shown below. Construct and interpret a 90% confidence interval for the difference in the mean DBH for longleaf pines in the northern and southern halves of the Wade Tract Preserve.

### Descriptive Statistics: North, South

<table>
<thead>
<tr>
<th>Variable</th>
<th>N</th>
<th>Mean</th>
<th>StDev</th>
</tr>
</thead>
<tbody>
<tr>
<td>North</td>
<td>30</td>
<td>23.70</td>
<td>17.50</td>
</tr>
<tr>
<td>South</td>
<td>30</td>
<td>34.53</td>
<td>14.25</td>
</tr>
</tbody>
</table>
Parameters:
\[ \mu_1 = \text{the true mean DBH of all trees in the northern half of the forest} \]
\[ \mu_2 = \text{the true mean DBH of all trees in the southern half of the forest.} \]

Assess Conditions:
✓ Random: Random samples of 30 trees each from the northern and southern halves of the forest.
✓ Normal: Reasonable to assume normality, since the samples sizes are each 30.
✓ Independent Researchers took independent samples from the northern and southern halves of the forest.
✓ 10 % Condition: Since we are sampling without replacement, there have to be at least 10(30) = 300 trees in each half of the forest. This is pretty safe to assume.
Name Test: Two-sample $t$ interval for the difference $\mu_1 - \mu_2$

$df = 30 - 1 = 29 \quad OR \quad df = 55.72$ (calculator)

Interval:
(-17.7238 to -3.93617)

We are 90% confident that the interval from -17.7238 to -3.93617 centimeters captures the difference in the actual mean DBH of the southern trees and the actual mean DBH of the northern trees.

Conclude: This interval suggests that the mean diameter of the southern trees is between 3.93 and 17.83 cm larger than the mean diameter of the northern trees.
Significance Tests for $\mu_1 - \mu_2$

• An observed difference between two sample means can reflect an actual difference in the parameters, or it may just be due to chance variation in random sampling or random assignment.

• Significance tests help us decide which explanation makes more sense.

• The null hypothesis has the general form: $H_0: \mu_1 = \mu_2$

• The alternative hypothesis says what kind of difference we expect:
  
  $H_a: \mu_1 > \mu_2 \quad OR \quad H_a: \mu_1 < \mu_2 \quad OR \quad H_a: \mu_1 \neq \mu_2$
Formula: **Significance Tests for** \( \mu_1 - \mu_2 \)

\[
t = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{s_1^2 \frac{1}{n_1} + s_2^2 \frac{1}{n_2}}}
\]

The degrees of freedom is determined by smaller of \( n_1 - 1 \) and \( n_2 - 1 \).
Conditions: Two Mean T-Significance Test

1) **Random**: Both sets of data should come from a well-designed random samples or randomized experiments.

2) **Normal**: Both sets of data must meet the Central Limit Theorem* with sample sizes greater than 30 or graph values that are less than 30 to check normality. No crazy outliers!

3) **Independent**: Both sets of data must be independent. When sampling without replacement, the sample size $n$ should be no more than 10% of the population size $N$ (the 10% condition). Must check the condition for each separate sample.
Using the Two-Sample $t$ Procedures: The Normal Condition

- **Sample size less than 15:** Use two-sample $t$ procedures if the data in both samples/groups appear close to Normal (roughly symmetric, single peak, no outliers). If the data are clearly skewed or if outliers are present, do not use $t$.

- **Sample size at least 15:** Two-sample $t$ procedures can be used except in the presence of outliers or **STRONG** skewness.

- **Large samples:** The two-sample $t$ procedures can be used even for clearly skewed distributions when both samples/groups are large, roughly $n \geq 30$. 
Tardy Policies

Mr. Medina and Mr. Hart are trying two new, different tardy policy systems with a randomly selected group of habitually tardy students.

Mr. Medina wants to know if his method resulted in a greater decrease in the number of tardies. Assume there are at least 210 habitually tardy students at ATM.

A negative value represents a net decrease in number of tardies from quarter 2 to quarter 3. (Fewer tardies are better.)

<table>
<thead>
<tr>
<th>Hart</th>
<th>7</th>
<th>-4</th>
<th>18</th>
<th>17</th>
<th>-3</th>
<th>-5</th>
<th>1</th>
<th>10</th>
<th>11</th>
<th>-2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Medina</td>
<td>-1</td>
<td>12</td>
<td>-1</td>
<td>-3</td>
<td>3</td>
<td>-5</td>
<td>5</td>
<td>2</td>
<td>-11</td>
<td>-1</td>
</tr>
</tbody>
</table>
Parameters & Hypotheses:

\[ H_0: \mu_1 = \mu_2 \]
\[ H_a: \mu_1 > \mu_2 \]

\( \mu_1 \) = the true mean decrease in tardies using Hart’s method

\( \mu_2 \) = the true mean decrease in tardies using Medina’s method

We will use \( \alpha = 0.05 \).
Assess Conditions:

• **Random**: The 21 students were randomly assigned to the two treatments.

• **Normal**: Since the sample sizes are less than 15, we must check and draw graphs.

The boxplots show no clear evidence of skewness and no outliers, therefore we can use $t$ procedures.

• **Independent**: Due to the random assignment, these two groups of students can be viewed as independent.
Name the Test: Two-sample $t$ test for the difference $\mu_1 - \mu_2$.

Test Statistic: \[ t = 1.634 \text{ or } 1.60372 \text{ (calc)} \]

Obtain p-value: p-value: 0.059348 or 0.06442 \text{ (calc)}

df = 9 \text{ or } 15.5905 \text{ (calc)}

\[
\begin{align*}
\bar{x}_1 - \bar{x}_2 &= \mu_1 - \mu_2 \\
\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2} &= \text{variance} \\
t &= \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} \\
&= \frac{[5.000 - (-0.273)] - 0}{\sqrt{8.743^2/10 + 5.901^2/11}} = 1.604
\end{align*}
\]
Make Decision: Because the $P$-value, 0.06442, is greater than $\alpha = 0.05$, we fail to reject the null hypothesis.

State Conclusion: There is not convincing evidence that Mr. Medina’s method yielded a statistically significant decrease in the number of tardies.
Confidence Interval vs. Significance Test

To get results that are consistent with the one-tailed test at $\alpha = 0.05$ from the example, we’ll use a 90% confidence level.

We are 90% confident that the interval from 0.4766 to -11.022 captures the difference in true mean decrease in tardies.

Because the 90% confidence interval includes 0 as a plausible value for the difference, we fail to reject the null hypothesis.
Using Two-Sample $t$ Procedures Wisely

✓ In planning a two-sample study, choose equal sample sizes if you can.

✓ Do not use “pooled” two-sample $t$ procedures!

✓ We are safe using two-sample $t$ procedures for comparing two means in a randomized experiment.

✓ Do not use two-sample $t$ procedures on paired data!

✓ Beware of making inferences in the absence of randomization. The results may not be generalized to the larger population of interest.
Formula Derivations
The Sampling Distribution of a Difference Between Two Means

Both $\bar{x}_1$ and $\bar{x}_2$ are random variables. The statistic $\bar{x}_1 - \bar{x}_2$ is the difference of these two random variables. In Chapter 6, we learned that for any two independent random variables $X$ and $Y$,

$$\mu_{X-Y} = \mu_X - \mu_Y \quad \text{and} \quad \sigma^2_{X-Y} = \sigma^2_X + \sigma^2_Y$$

Therefore,

$$\mu_{\bar{x}_1 - \bar{x}_2} = \mu_{\bar{x}_1} - \mu_{\bar{x}_2} = \mu_1 - \mu_2$$

$$\sigma^2_{\bar{x}_1 - \bar{x}_2} = \sigma^2_{\bar{x}_1} + \sigma^2_{\bar{x}_2} = \left( \frac{\sigma_1}{\sqrt{n_1}} \right)^2 + \left( \frac{\sigma_2}{\sqrt{n_2}} \right)^2 = \frac{\sigma_1^2}{n_1} + \frac{\sigma_1^2}{n_2}$$

$$\sigma_{\bar{x}_1 - \bar{x}_2} = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_1^2}{n_2}}$$
The Two-Sample \( t \) Statistic

When data come from two random samples or two groups in a randomized experiment, the statistic \( \bar{x}_1 - \bar{x}_2 \) is our best guess for the value of \( \mu_1 - \mu_2 \).

When the Independent condition is met, the standard deviation of the statistic \( \bar{x}_1 - \bar{x}_2 \) is:

\[
\sigma_{\bar{x}_1 - \bar{x}_2} = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}
\]

Since we don't know the values of the parameters \( \sigma_1 \) and \( \sigma_2 \), we replace them in the standard deviation formula with the sample standard deviations. The result is the standard error of the statistic \( \bar{x}_1 - \bar{x}_2 \):

\[
\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}
\]

If the Normal condition is met, we standardize the observed difference to obtain a \( t \) statistic that tells us how far the observed difference is from its mean in standard deviation units:

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

The two-sample \( t \) statistic has approximately a \( t \) distribution. We can use technology to determine degrees of freedom OR we can use a conservative approach, using the smaller of \( n_1 - 1 \) and \( n_2 - 1 \) for the degrees of freedom.