Chapter 7 Two Independent Samples

Fall 2010
7.1 Introduction
Draw Two Sample Picture on the Board
Chapter 7 describes two ways to compare two populations on the basis of independent samples: a \textit{confidence interval for the difference in population means} and a \textit{hypothesis test}.

The basic structure of the confidence interval is the same as in the previous chapter — an estimate plus or minus a multiple of a standard error.

Hypothesis testing will introduce several new concepts.
Example

- Hematocrit level is a measure of the concentration of red cells in blood.
- Researchers are interested in the gender difference.
- They collected two samples of 17 year old Americans: 489 males, and 469 females.
- Males ($\bar{y}_1 = 45.8$, $s_1 = 2.8$), Females ($\bar{y}_2 = 40.6$, $s_2 = 2.9$)
7.2 Standard Error of $\bar{y}_1 - \bar{y}_2$
Standard Error of $\bar{y}_1 - \bar{y}_2$

The **standard error** of the difference in two sample means is an empirical measure of how far the difference in sample means will typically be from the difference in the respective population means.

$$
\text{SE}(\bar{y}_1 - \bar{y}_2) = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}
$$

An alternative formula is

$$
\text{SE}(\bar{y}_1 - \bar{y}_2) = \sqrt{(\text{SE}(\bar{y}_1))^2 + (\text{SE}(\bar{y}_2))^2}
$$
Pooled Standard Error

If we wish to assume that the two population standard deviations are equal, $\sigma_1 = \sigma_2$, then it makes sense to use data from both samples to estimate the common population standard deviation. We estimate the common population variance with a weighted average of the sample variances, weighted by the degrees of freedom.

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

The pooled standard error is then as below.

$$SE_{pooled} = s_{pooled} \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$
7.3 Confidence Intervals
Sampling Distributions

The sampling distribution of the difference in sample means has these characteristics.

- **Mean:** $\mu_1 - \mu_2$
- **SD:** $\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$
- **Shape:** Exactly normal if both populations are normal, approximately normal if populations are not normal but both sample sizes are sufficiently large.
The recipe for constructing a confidence interval for a single population mean is based on facts about the sampling distribution of the statistic

\[ T = \frac{\bar{Y} - \mu}{SE(\bar{Y})}. \]
Similarly, the theory for confidence intervals for $\mu_1 - \mu_2$ is based on the sampling distribution of the statistic

$$T = \frac{(\bar{Y}_1 - \bar{Y}_2) - (\mu_1 - \mu_2)}{SE(\bar{Y}_1 - \bar{Y}_2)}$$

where we standardize by subtracting the mean and dividing by the standard deviation of the sampling distribution.
If both populations are normal and if we know the population standard deviations, then

\[
\Pr \left\{ -1.96 \leq \frac{\bar{Y}_1 - \bar{Y}_2 - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \leq 1.96 \right\} = 0.95
\]

where we can choose \( z \) other than 1.96 for different confidence levels. This statement is true because the expression in the middle has a standard normal distribution.
But in practice, we don’t know the population standard deviations. If we substitute in sample estimates instead, we get this.

\[
\Pr \left\{ -t \leq \frac{\bar{Y}_1 - \bar{Y}_2 - (\mu_1 - \mu_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} \leq t \right\} = 0.95
\]

We need to choose different end points to account for the additional randomness in the denominator.
Theory (continued)

It turns out that the sampling distribution of the statistic above is approximately a \( t \) distribution where the degrees of freedom should be estimated from the data as well.
Theory (continued)

Algebraic manipulation leads to the following expression.

\[
\Pr \left\{ (\bar{Y}_1 - \bar{Y}_2) - t \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} \leq \mu_1 - \mu_2 \leq (\bar{Y}_1 - \bar{Y}_2) + t \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} \right\} = 0.95
\]
We use a $t$ multiplier so that the area between $-t$ and $t$ under a $t$ distribution with the estimated degrees of freedom will be 0.95.
Confidence Interval for \( \mu_1 - \mu_2 \)

The confidence interval for differences in population means has the **same structure** as that for a single population mean.

\[
(\text{Estimate}) \pm (t \text{ Multiplier}) \times \text{SE}
\]

The only difference is that for this more complicated setting, we have **more complicated formulas** for the standard error and the degrees of freedom.
The formula

\[ df = \frac{(SE_1^2 + SE_2^2)^2}{SE_1^4/(n_1 - 1) + SE_2^4/(n_2 - 1)} \]

where \( SE_i = s_i/\sqrt{n_i} \) for \( i = 1, 2 \).

As a check, the value is often close to \( n_1 + n_2 - 2 \). (This will be exact if \( s_1 = s_2 \) and if \( n_1 = n_2 \).)
Degrees of Freedom—What if it is not an integer?

- The formula

\[
\text{df} = \frac{(SE_1^2 + SE_2^2)^2}{SE_1^4/(n_1 - 1) + SE_2^4/(n_2 - 1)}
\]

where \(SE_i = s_i/\sqrt{n_i}\) for \(i = 1, 2\).

- Example \(df = 12.8\)?
  - Using the table, round down (\(df = 12\))
  - Using R, use \(df = 12.8\)
Example 7.7

- Board Example
- $R$

$$qt(.975, 12.8) = 2.163805$$
7.4 Hypothesis Tests
Hypothesis Tests

- **Hypothesis tests** are an alternative approach to statistical inference.
- Unlike confidence intervals where the goal is estimation with assessment of likely precision of the estimate, the goal of hypothesis testing is to ascertain whether or not data is consistent with what we might expect to see assuming that a hypothesis is true.
- The logic of hypothesis testing is a probabilistic form of proof by contradiction.
- In logic, if we can say that a proposition $H$ leads to a contradiction, then we have proved $H$ false and have proved $\{\text{not} H\}$ to be true.
- In hypothesis testing, if observed data is highly unlikely under an assumed hypothesis $H$, then there is strong (but not definitive) evidence that the hypothesis is false.
Logic of Hypothesis Tests

All of the hypothesis tests we will see this semester fall into this general framework.

1. State a null hypothesis and an alternative hypothesis.
2. Gather data and compute a test statistic.
3. Consider the sampling distribution of the test statistic assuming that the null hypothesis is true.
4. Compute a p-value, a measure of how consistent the data is with the null hypothesis in consideration of a specific alternative hypothesis.
5. Assess the strength of the evidence against the null hypothesis in the context of the problem.

We will introduce all of these concepts in the setting of testing the equality of two population means, but the general ideas will reappear in many settings throughout the remainder of the semester.
Wisconsin Fast Plants Example

- In an experiment, seven Wisconsin Fast Plants (*Brassica campestris*) were grown with a treatment of Ancymidol (ancy) and eight control plants were given ordinary water.
- The **null hypothesis** is that the treatment has no effect on plant growth (as measured by the height of the plant after 14 days of growth).
- The **alternative hypothesis** is that the treatment has an effect which would result in different mean growth amounts.
- A summary of the sample data is as follows. The eight control plants had a mean growth of 15.9 cm and standard deviation 4.8 cm. The seven ancy plants had a mean growth of 11.0 cm and standard deviation 4.7 cm.
- The question is, is it reasonable to think that the observed difference in sample means of 4.9 cm is due to **chance variation alone**, or is there evidence that some of the difference is due to the ancy treatment?
## Data Summary

<table>
<thead>
<tr>
<th></th>
<th>Control</th>
<th>Water</th>
</tr>
</thead>
<tbody>
<tr>
<td>$n$</td>
<td>8</td>
<td>7</td>
</tr>
<tr>
<td>$\bar{y}$</td>
<td>15.9</td>
<td>11.0</td>
</tr>
<tr>
<td>$s$</td>
<td>4.8</td>
<td>4.7</td>
</tr>
<tr>
<td>SE</td>
<td>1.7</td>
<td>1.8</td>
</tr>
</tbody>
</table>
Let $\mu_1$ be the population mean growth with the control conditions and let $\mu_2$ be the population mean with ancy. The null and alternative hypotheses are expressed as

$$H_0: \mu_1 = \mu_2 \quad H_A: \mu_1 \neq \mu_2$$

We state statistical hypotheses as statements about population parameters.
Example: Calculate a Test Statistic

In the setting of a difference between two independent sample means, our test statistic is

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

(Your book adds a subscript, \(t_s\), to remind you that this is computed from the sample.)

\[
t = \frac{(15.9 - 11.0) - 0}{\frac{2.46}{\frac{2}{n_2}}} = 1.99
\]
Example: Find the Sampling Distribution

The sampling distribution of the test statistic is a $t$ distribution with degrees of freedom calculated by the messy formula.

In our example:

$$df = \frac{(1.7^2 + 1.8^2)^2}{1.7^4/7 + 1.8^4/6} = 12.8$$
Example: Compute a P-Value

- To describe how likely it is to see such a test statistic, we can ask what is the probability that chance alone would result in a test statistic at least this far from zero?
- The answer is the area below $-1.99$ and above $1.99$ under a $t$ density curve with 12.8 degrees of freedom.
- With the $t$-table, we can only calculate this p-value within a range.
- If we round down to 12 df, the $t$ statistic is bracketed between 1.912 and 2.076 in the table.
- Thus, the area to the right of 1.99 is between 0.03 and 0.04.
- The p-value in this problem is twice as large because we need to include as well the area to the left of $-1.99$.
- So, $0.06 < p < 0.08$. 
Compute a P-Value (in R)

\[ p = 2 \times pt(-1.99, 12.8) \]

\[ p = 0.06834 \]
Example: Interpreting a P-Value

1. The smaller the p-value, the more inconsistent the data is with the null hypothesis, the stronger the evidence is against the null hypothesis in favor of the alternative.

2. Traditionally, people have measured statistical significance by comparing a p-value with arbitrary significance levels such as $\alpha = 0.05$. The phrase “statistically significant at the 5% level” means that the p-value is smaller than 0.05.

3. In reporting results, it is best to report an actual p-value and not simply a statement about whether or not it is “statistically significant”.
For this example, one possible summary of the results follows:

*The data do not provide sufficient evidence* ($p = .068$, *two-sided independent sample t-test*) *at the .05 level of significance to conclude that ancy and water differ in their effects on fast plant growth (under the conditions of the experiment that was conducted).*
Example

- Example on Board
7.5 More On the t Test
Suppose that we were asked to make a decision about a hypothesis based on data.

We may decide, for example to reject the null hypothesis if the $p$-value were smaller than 0.05 and to not reject the null hypothesis if the $p$-value were larger than 0.05.

This procedure has a significance level of 0.05, which means that if we follow the rule, there is a probability of 0.05 of rejecting a true null hypothesis.

Rejecting the null hypothesis occurs precisely when the test statistic falls into a rejection region, in this case either the upper or lower 2.5% tail of the sampling distribution.
The rejection region corresponds exactly to the test statistics for which a 95% confidence interval contains 0.

We would reject the null hypothesis

\[ H_0: \mu_1 - \mu_2 = 0 \]

versus the two-sided alternative at the \( \alpha = 0.05 \) level of significance if and only if a 95% confidence interval for \( \mu_1 - \mu_2 \) does not contain 0.

We could make similar statements for general \( \alpha \) and a \( (1 - \alpha) \times 100\% \) confidence interval.
Example 7.16

- Example on Board
Type I and Type II Errors

There are two possible decision errors.

- Rejecting a true null hypothesis is a **Type I error**.
- You can interpret $\alpha = \Pr\{\text{rejecting } H_0 \mid H_0 \text{ is true}\}$, so $\alpha$ is the probability of a Type I error. (You cannot make a Type I error when the null hypothesis is false.)
- Not rejecting a false null hypothesis is a **Type II error**.
- It is convention to use $\beta$ as the probability of a Type II error, or $\beta = \Pr\{\text{not rejecting } H_0 \mid H_0 \text{ is false}\}$. If the null hypothesis is false, one of the many possible alternative hypotheses is true. It is typical to calculate $\beta$ separately for each possible alternative. (In this setting, for each value of $\mu_1 - \mu_2$.)
- **Power** is the probability of rejecting a false null hypothesis. Power $= 1 - \beta$. 
## Type I and Type II Errors - Summary

<table>
<thead>
<tr>
<th>True Situation</th>
<th>( H_0 ) true</th>
<th>( H_0 ) false</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Our Decision</strong></td>
<td>Do not reject ( H_0 )</td>
<td>Correct</td>
</tr>
<tr>
<td></td>
<td>Reject ( H_0 )</td>
<td>Type I error</td>
</tr>
</tbody>
</table>
Comparing $\alpha$ and $P$-values

- In this setting, the significance level $\alpha$ and $p$-values are both areas under $t$ curves, but they are not the same thing.
- The significance level is a prespecified, arbitrary value, that does not depend on the data.
- The $p$-value depends on the data.
- If a decision rule is to reject the null hypothesis when the test statistic is in a rejection region, this is equivalent to rejecting the null hypothesis when the $p$-value is less than the significance level $\alpha$. 
A verbal definition of a $p$-value is as follows.

*The $p$-value of the data is the probability calculated assuming that the null hypothesis is true of obtaining a test statistic that deviates from what is expected under the null (in the direction of the alternative hypothesis) at least as much as the actual data does.*

The $p$-value is not the probability that the null hypothesis is true. Interpreting the $p$-value in this way will mislead you!
7.6 One-tailed t Tests
One-tailed Tests

- Often, we are interested not only in demonstrating that two population means are different, but in demonstrating that the difference is in a particular direction.

- Instead of the two-sided alternative $\mu_1 \neq \mu_2$, we would choose one of two possible one-sided alternatives, $\mu_1 < \mu_2$ or $\mu_1 > \mu_2$.

- For the alternative hypothesis $H_A: \mu_1 < \mu_2$, the $p$-value is the area to the left of the test statistic.

- For the alternative hypothesis $H_A: \mu_1 > \mu_2$, the $p$-value is the area to the right of the test statistic.

- If the test statistic is in the direction of the alternative hypothesis, the $p$-value from a one-sided test will be half the $p$-value of a two-sided test.
Example

Researchers are interested in whether niacin increases weight gain. They conduct the following experiment on lambs. The observation $Y$ will be weight gain in a two-week trial. Ten animals will receive Diet 1 (Standard ration & niacin), and ten animals will receive Diet 2 (Standard ration).

$H_0$: Niacin is not effective in increasing weight gain ($\mu_1 = \mu_2$)

$H_A$: Niacin is effective in increasing weight gain ($\mu_1 > \mu_2$)
One-Tailed Test Procedure

Two steps:

1. Check directionality—see if the data deviate from $H_0$ in the direction specified by $H_A$:
   - If not, the p-value is greater than .50
   - If so, proceed to Step 2.

2. The p-value of the data is the one-tailed are beyond $t_s$. 
Example 7.22

- Example on Board
Computing p-values
Computing p-values (for a t-test)

1. Using R, `pt()`
2. Using the t table (p. 677)
Example 7.14 On the Board

- Demonstrate with both the t-table and R.
Hypothesis Testing - Accepting versus not Rejecting $H_0$
Accepting or Not Rejecting $H_0$?

- $p$-value $= \mathbb{P}\{\text{obtaining a value of } T \text{ as extreme as in the data, when } H_0 \text{ is true.}\}$
- It is not a probability that $H_0$ is true!!
Recall p-value

low p-value $\implies$ evidence that $H_0$ is false
high p-value $\implies$ data consistent with $H_0$.

It is **not** evidence that $H_0$ is true.
Body Weight Example (7.26)

<table>
<thead>
<tr>
<th></th>
<th>males</th>
<th>females</th>
</tr>
</thead>
<tbody>
<tr>
<td>$n$</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>$\bar{y}$</td>
<td>175</td>
<td>143 (lbs)</td>
</tr>
<tr>
<td>$s$</td>
<td>35</td>
<td>34</td>
</tr>
</tbody>
</table>

$H_0$: “average body weight is the same in males and females”,
$H_A$: “average body weight differ in males and females”.

We get $\bar{y}_1 - \bar{y}_2 = 32$ lbs and $t_s = .93$ so $p = .45$.

We fail to reject $H_0$ (rather than firmly accept it).
The data are consistent with $H_0$, but they could also be consistent with $H_A$.

**Low power** here: With samples of $n = 2$ subjects, it is hard to detect the actual difference between $\mu_{\text{male}}$ and $\mu_{\text{female}}$.

When $p$ is very high (> .50 say) we **fail to reject** $H_0$. We might say we accept $H_0$, but we always keep in mind it is possible that $\mu_1 \neq \mu_2$, with the difference being too small to be detected with the sample size we have.
Accepting or Not Rejecting \( H_0 \)?

- The data are consistent with \( H_0 \), but they could also be consistent with \( H_A \).
- **Low power** here: With samples of \( n = 2 \) subjects, it is hard to detect the actual difference between \( \mu_{\text{male}} \) and \( \mu_{\text{female}} \).
- When \( p \) is very high (> .50 say) we **fail to reject** \( H_0 \). We might say we accept \( H_0 \), but we always keep in mind it is possible that \( \mu_1 \neq \mu_2 \), with the difference being too small to be detected with the sample size we have.
The Reverse “Problem” – very large $n$

- If $n$ is too small, your test will not be very powerful (you will fail to reject $H_0$)
- If $n$ is “too big” will you always reject the null hypothesis?
- Almost any null hypothesis can be disproved with a large enough sample.
- Are there really populations out there exactly equal to each other in every way?
- We should care about the differences only if they are large enough to matter.
Statistical Significance versus Biological Importance
(Significant Difference versus Important Difference)
A significant (or statistically significant) result means that a null hypothesis has been rejected. Often at $\alpha = .05$. 
Statistical Significance versus Biological Importance

- Statistical Significance tells us how confidently we can reject a null hypothesis, but now how big or how important the effect is.
- The importance of a result depends on the value of the question and the size of the effect.
Automobile accidents increase during full moons, and this result is statistically significant.

The size of the effect is about 1%, which is too small to make it worth changing our driving habits or public policy.

Is this a scientifically important difference?
Example 7.31 - Yields of Tomatoes

A horticulturist is comparing the yields of two varieties of tomatoes; yield is measured as pounds of tomatoes per plant. On the basis of practical considerations, the horticulturist has decided that a difference between the varieties is “important” only if it exceeds 1 pound per plant, on the average. That is, the difference is important if

\[ |\mu_1 - \mu_2| > 1\text{lb} \]