Chapter 9
Inferences from Two Samples

9-1 Overview
9-2 Inferences About Two Proportions
9-3 Inferences About Two Means: Independent Samples
9-4 Inferences from Matched Pairs
9-5 Comparing Variation in Two Samples

Section 9-1
Overview
Overview
There are many important and meaningful situations in which it becomes necessary to compare two sets of sample data.

This chapter extends the same methods introduced in Chapters 7 and 8 to situations involving two samples instead of only one.

Section 9-2
Inferences About Two Proportions

Key Concept
This section presents methods for using two sample proportions for constructing a confidence interval estimate of the difference between the corresponding population proportions, or testing a claim made about the two population proportions.

Requirements
1. We have proportions from two independent simple random samples.
2. For each of the two samples, the number of successes is at least 5 and the number of failures is at least 5.

Notation for Two Proportions
For population 1, we let:
\( p_1 = \text{population proportion} \)
\( n_1 = \text{size of the sample} \)
\( x_1 = \text{number of successes in the sample} \)
\( \hat{p}_1 = \frac{x_1}{n_1} \text{ (the sample proportion)} \)
\( \hat{q}_1 = 1 - \hat{p}_1 \)
The corresponding meanings are attached to \( p_2, n_2, x_2, \hat{p}_2, \) and \( \hat{q}_2, \) which come from population 2.

Pooled Sample Proportion
\( \text{The pooled sample proportion} \)
is denoted by \( \bar{p} \) and is given by:
\[
\bar{p} = \frac{x_1 + x_2}{n_1 + n_2}
\]
\( \text{We denote the complement of } \bar{p} \text{ by } \bar{q}, \)
so \( \bar{q} = 1 - \bar{p} \)
Test Statistic for Two Proportions

For \( H_0: p_1 = p_2 \)
\( H_1: p_1 \neq p_2 \), \( H_1: p_1 < p_2 \), \( H_1: p_1 > p_2 \)

\[
Z = \frac{\hat{p}_1 - \hat{p}_2 - (p_1 - p_2)}{\sqrt{\frac{\hat{p}\hat{q}}{n_1} + \frac{\hat{p}\hat{q}}{n_2}}}
\]

\( \hat{p} = \frac{x_1 + x_2}{n_1 + n_2} \) and \( \hat{q} = 1 - \hat{p} \)

Test Statistic for Two Proportions - cont

P-value: Use Table A-2. (Use the computed value of the test statistic \( z \) and find its P-value by following the procedure summarized by Figure 8-6 in the text.)

Critical values: Use Table A-2. (Based on the significance level \( \alpha \), find critical values by using the procedures introduced in Section 8-2 in the text.)

Example: For the sample data listed in the Table below, use a 0.05 significance level to test the claim that the proportion of black drivers stopped by the police is greater than the proportion of white drivers who are stopped.

<table>
<thead>
<tr>
<th>Racial Profiling Data</th>
<th>Race and Ethnicity</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Black and Non-Hispanic</td>
</tr>
<tr>
<td>Drivers stopped by police</td>
<td>24</td>
</tr>
<tr>
<td>Total number of observed drivers</td>
<td>200</td>
</tr>
<tr>
<td>Percent Stopped by Police</td>
<td>12.0%</td>
</tr>
</tbody>
</table>

Example: For the sample data listed in the previous Table, use a 0.05 significance level to test the claim that the proportion of black drivers stopped by the police is greater than the proportion of white drivers who are stopped.

\( n_1 = 200 \)
\( x_1 = 24 \)
\( \hat{p}_1 = \frac{x_1}{n_1} = \frac{24}{200} = 0.120 \)
\( \hat{q}_1 = \frac{n_1 - x_1}{n_1} = \frac{200 - 24}{200} = 0.880 \)
\( n_2 = 1400 \)
\( x_2 = 147 \)
\( \hat{p}_2 = \frac{x_2}{n_2} = \frac{147}{1400} = 0.105 \)
\( \hat{q}_2 = \frac{n_2 - x_2}{n_2} = \frac{1400 - 147}{1400} = 0.915 \)

\( \hat{p} = \frac{x_1 + x_2}{n_1 + n_2} = \frac{24 + 147}{200 + 1400} = 0.106875 \)
\( \hat{q} = 1 - \hat{p} = 0.893125 \)

\( z = \frac{\hat{p}_1 - \hat{p}_2 - (p_1 - p_2)}{\sqrt{\frac{\hat{p}\hat{q}}{n_1} + \frac{\hat{p}\hat{q}}{n_2}}} = \frac{0.120 - 0.105}{\sqrt{\frac{0.106875(0.893125)}{200} + \frac{0.106875(0.893125)}{1400}}} = 0.64 \)
**Example:** For the sample data listed in the previous Table, use a 0.05 significance level to test the claim that the proportion of black drivers stopped by the police is greater than the proportion of white drivers who are stopped.

\[ n_1 = 200 \]
\[ x_1 = 24 \]
\[ \hat{p}_1 = \frac{x_1}{n_1} = 24 = 0.120 \]
\[ n_2 = 1400 \]
\[ x_2 = 147 \]
\[ \hat{p}_2 = \frac{x_2}{n_2} = 147 = 0.105 \]

\[ z = 0.64 \]

This is a right-tailed test, so the P-value is the area to the right of the test statistic \( z = 0.64 \). The P-value is 0.2611. Because the P-value of 0.2611 is greater than the significance level of \( \alpha = 0.05 \), we fail to reject the null hypothesis.

**Example:** For the sample data listed in the previous Table, use a 0.05 significance level to test the claim that the proportion of black drivers stopped by the police is greater than the proportion of white drivers who are stopped.

\[ n_1 = 200 \]
\[ x_1 = 24 \]
\[ \hat{p}_1 = \frac{x_1}{n_1} = 24 = 0.120 \]
\[ n_2 = 1400 \]
\[ x_2 = 147 \]
\[ \hat{p}_2 = \frac{x_2}{n_2} = 147 = 0.105 \]

\[ z = 0.64 \]

Because we fail to reject the null hypothesis, we conclude that there is not sufficient evidence to support the claim that the proportion of black drivers stopped by police is greater than that for white drivers. This does not mean that racial profiling has been disproved. The evidence might be strong enough with more data.

**Confidence Interval Estimate of \( p_1 - p_2 \)**

\[
(\hat{p}_1 - \hat{p}_2) - E < (p_1 - p_2) < (\hat{p}_1 - \hat{p}_2) + E
\]

where \( E = z_{\alpha/2} \sqrt{\frac{\hat{p}_1 \hat{q}_1}{n_1} + \frac{\hat{p}_2 \hat{q}_2}{n_2}} \)

**Example:** For the sample data listed in the previous Table, find a 90% confidence interval estimate of the difference between the two population proportions.

\[ n_1 = 200 \]
\[ x_1 = 24 \]
\[ \hat{p}_1 = \frac{x_1}{n_1} = 24 = 0.120 \]
\[ n_2 = 1400 \]
\[ x_2 = 147 \]
\[ \hat{p}_2 = \frac{x_2}{n_2} = 147 = 0.105 \]

\[ E = 1.645 \sqrt{\frac{0.120(0.88)}{200} + \frac{0.105(0.895)}{1400}} = 0.040 \]

**Example:** For the sample data listed in the previous Table, use a 0.05 significance level to test the claim that the proportion of black drivers stopped by the police is greater than the proportion of white drivers who are stopped.

\[ n_1 = 200 \]
\[ x_1 = 24 \]
\[ \hat{p}_1 = \frac{x_1}{n_1} = 24 = 0.120 \]
\[ n_2 = 1400 \]
\[ x_2 = 147 \]
\[ \hat{p}_2 = \frac{x_2}{n_2} = 147 = 0.105 \]

\[ E = 0.040 \]

\[ (0.120 - 0.105) - 0.040 < (p_1 - p_2) < (0.120 - 0.105) + 0.040 \]

\[ -0.025 < (p_1 - p_2) < 0.055 \]

\[ n_2 = 1400 \]
\[ x_2 = 147 \]
\[ \hat{p}_2 = \frac{x_2}{n_2} = 147 = 0.105 \]
\[ n_2 = 1400 \]
Section 9-3
Inferences About Two Means: Independent Samples

Key Concept
This section presents methods for using sample data from two independent samples to test hypotheses made about two population means or to construct confidence interval estimates of the difference between two population means.

Definitions
Two samples are independent if the sample values selected from one population are not related to or somehow paired or matched with the sample values selected from the other population.

Two samples are dependent (or consist of matched pairs) if the members of one sample can be used to determine the members of the other sample.

Requirements
1. \( \sigma_1 \) and \( \sigma_2 \) are unknown and no assumption is made about the equality of \( \sigma_1 \) and \( \sigma_2 \).
2. The two 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 normal distributions.

Hypothesis Test for Two Means: Independent Samples

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

Hypothesis Test - cont

Test Statistic for Two Means: Independent Samples

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

P-values: Refer to Table A-3. Use the procedure summarized in Figure 8-6.

Critical values: Refer to Table A-3.
McGwire Versus Bonds

Sample statistics are shown for the distances of the home runs hit in record-setting seasons by Mark McGwire and Barry Bonds. Use a 0.05 significance level to test the claim that the distances come from populations with different means.

\[
\begin{array}{|c|c|c|}
\hline
 & McGwire & Bonds \\
\hline
n & 70 & 73 \\
x & 418.5 & 403.7 \\
\bar{s} & 45.5 & 30.6 \\
\hline
\end{array}
\]

McGwire Versus Bonds - cont

Claim: \( \mu_1 \neq \mu_2 \)
\[H_0 : \mu_1 = \mu_2 \]
\[H_1 : \mu_1 \neq \mu_2 \]
\[\alpha = 0.05 \]

\[n_1 - 1 = 69 \]
\[n_2 - 1 = 72 \]
\[df = 69 \]
\[t_{0.025} = 1.994 \]

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

\[= \frac{(418.5 - 403.7) - 0}{\sqrt{\frac{45.5^2}{70} + \frac{30.6^2}{73}}} \]
\[= 2.273 \]
McGwire Versus Bonds - cont

Claim: \( \mu_1 \neq \mu_2 \)

\( H_0 : \mu_1 = \mu_2 \)

\( H_1 : \mu_1 \neq \mu_2 \)

\( \alpha = 0.05 \)

There is significant evidence to support the claim that there is a difference between the mean home run distances of Mark McGwire and Barry Bonds.

Confidence Intervals

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

where

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

McGwire Versus Bonds Confidence Interval Method

Using the data given in the preceding example, construct a 95% confidence interval estimate of the difference between the mean home run distances of Mark McGwire and Barry Bonds.

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

\[
E = 1.994 \sqrt{\frac{45.5^2}{70} + \frac{30.6^2}{73}}
\]

\[ E = 13.0 \]

Part 2: Alternative Methods

Independent Samples with \( \sigma_1 \) and \( \sigma_2 \) Known.
Requirements
1. The two population standard deviations are both known.
2. The two 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 normal distributions.

Hypothesis Test for Two Means: Independent Samples with \( \sigma_1 \) and \( \sigma_2 \) Both Known

\[
z = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}
\]

*P*-values and critical values: Refer to Table A-2.

Confidence Interval: Independent Samples with \( \sigma_1 \) and \( \sigma_2 \) Both Known

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

where \( E = z_{\alpha/2} \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} \)

Methods for Inferences About Two Independent Means

Figure 9-3

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 are 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 Statistic for Two Means: Independent Samples and \( \sigma_1 = \sigma_2 \)

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

Where

\[
s_\sigma^2 = \frac{(n_1 - 1) s_1^2 + (n_2 - 1) s_2^2}{(n_1 - 1) + (n_2 - 1)}
\]

and the number of degrees of freedom is \( df = n_1 + n_2 - 2 \)

Confidence Interval Estimate of \( \mu_1 - \mu_2 \): Independent Samples with \( \sigma_1 = \sigma_2 \)

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

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

and number of degrees of freedom is \( df = n_1 + n_2 - 2 \)

Strategy

Unless instructed otherwise, use the following strategy:

Assume that \( \sigma_1 \) and \( \sigma_2 \) are unknown, do not assume that \( \sigma_1 = \sigma_2 \), and use the test statistic and confidence interval given in Part 1 of this section. (See Figure 9-3.)