Chapter 10: Estimation and hypothesis testing: two populations

What are we going to cover?

- 10.1 Comparing Two Population Means Using Large Independent Samples
- 10.2 Comparing Two Population Means Using Small Independent Samples: Equal standard deviations
- 10.3 Comparing Two Population Means Using Small Independent Samples: Unequal standard deviations
- 10.4 Paired Difference Experiments
- 10.5 Comparing Two Population Proportions Using Large Independent Samples
Chapter 10: Estimation and hypothesis testing: two populations

Why do we need to compare two populations?

- You have just received a job offer in two different countries with the same salary.
- You want to invest your money into two different stock markets.
- A pharmaceutical company just announced a new drug that is better than Panadol.
- A bank manager introduced a new serving policy that reduces the waiting time.
- You want to decide between two cars.
10.1 Inferences about the difference between two population means for large and independent samples

- Let $\mu_1$ be the mean of the first population and $\mu_2$ be the mean of the second population.
- Suppose we want to make a confidence interval and test a hypothesis about the difference between these two population means, that is, $\mu_1 - \mu_2$.
- Let $\bar{x}_1$ be the mean of a sample taken from the first population and $\bar{x}_2$ be the mean of a sample taken from the second population.
- Then, $\bar{x}_1 - \bar{x}_2$ is the sample statistic that is used to make an interval estimate and to test a hypothesis about $\mu_1 - \mu_2$. 
10.1 Inferences about the difference between two population means for large and independent samples 10.1.1 Independent versus Dependent Samples

- Two samples are *independent* if they are drawn from two different populations and the elements of one sample have no relationship to the elements of the second sample.
- If the elements of the two samples are somehow related, then the samples are said to be dependent.
- Thus, in two independent samples, the selection of one sample has no effect on the selection of the second sample.
Chapter 10: Estimation and hypothesis testing: two populations

10.1 Inferences about the difference between two population means for large and independent samples

Example: Suppose we want to estimate the difference between the mean salaries of all male and all female executives. To do so, we draw two samples, one from the population of male executives and another from the population of female executives. These two samples are independent because they are drawn from two different populations, and the samples have no effect on each other.

Example: Suppose we want to estimate the difference between the mean weights of all participants before and after a weight loss program. To accomplish this, suppose we take a sample of 40 participants and measure their weights before and after the completion of this program. Note that these two samples include the same 40 participants. This is an example of two dependent samples. Such samples are also called paired or matched samples.
10.1 Inferences about the difference between two population means for large and independent samples

10.1.2 Mean, standard deviation, and sampling distribution of $\bar{x}_2 - \bar{x}_1$

- Suppose we select two (independent) large samples from two different populations that are referred to as population 1 and population 2.

- $\mu_1$ = the mean of population 1
- $\sigma_1$ = the standard deviation of population 1
- $n_1$ = the size of the sample drawn from pop 1
- $\bar{x}_1$ = the mean of sample 1

- $\mu_2$ = the mean of population 2
- $\sigma_2$ = the standard deviation of population 2
- $n_2$ = the size of the sample drawn from pop 2
- $\bar{x}_2$ = the mean of sample 2
10.1 Inferences about the difference between two population means for large and independent samples

If independent random samples are taken from two population, then the sampling distribution of the sample difference in means $\bar{x}_1 - \bar{x}_2$ is

- Normal, if each of the sampled populations is normal and approximately normal if the sample sizes $n_1$ and $n_2$ are large

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

- Has standard deviation:

$$\sigma_{\bar{x}_1-\bar{x}_2} = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}, \text{ or } S_{\bar{x}_1-\bar{x}_2} = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$
10.1 Inferences about the difference between two population means for large and independent samples

10.1.3 Interval estimate of \( \mu_1 - \mu_2 \)

If two independent samples are from populations that are normal or each of the sample sizes is large, \(100(1 - \alpha)\)% confidence interval for \( \mu_1 - \mu_2 \) is

\[
(\bar{X}_1 - \bar{X}_2) \pm Z_{\alpha/2} \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}
\]
10.1 Inferences about the difference between two population means for large and independent samples

- If \( \sigma_1 \) and \( \sigma_2 \) are unknown and each of the sample sizes is large (\( n_1, n_2 \geq 30 \)), estimate the sample standard deviations by \( s_1 \) and \( s_2 \) and a 100(1 - \( a \)% confidence interval for \( \mu_1 - \mu_2 \) is

\[
(\bar{X}_1 - \bar{X}_2) \pm Z_{\alpha/2} \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}
\]
10.1 Inferences about the difference between two population means for large and independent samples

Example: According to the U.S. Bureau of the Census, the average annual salary of full-time state employees was $49,056 in New York and $46,800 in Massachusetts in 2001. Suppose that these mean salaries are based on random samples of 500 full-time state employees from New York and 400 full-time state employees from Massachusetts and that the population standard deviations of the 2001 salaries of all full-time state employees in these two states were $9000 and $8500, respectively.

(a) What is the point estimate of $\mu_1 - \mu_2$? What is the margin of error?

(b) Construct a 97% confidence interval for the difference between the 2001 mean salaries of all full-time state employees in these two states.
10.1 Inferences about the difference between two population means for large and independent samples

**Example:** According to the National Association of Colleges and Employers, the average salary offered to college students who graduated in 2002 was $43,732 to MIS (Management Information Systems) majors and $40,293 to accounting majors. Assume that these means are based on samples of 900 MIS and 1200 accounting majors and that the sample standard deviations for the two samples are $2200 and $1950, respectively. Find a 99% confidence interval for the difference between the corresponding population means.
10.1 Inferences about the difference between two population means for large and independent samples

10.1.4 Hypothesis testing of $\mu_1 - \mu_2$

- The three situations of the alternative hypothesis are:
  1. $\mu_1 \neq \mu_2$ is same as $\mu_1 - \mu_2 \neq 0$
  2. $\mu_1 > \mu_2$ is same as $\mu_1 - \mu_2 > 0$
  3. $\mu_1 < \mu_2$ is same as $\mu_1 - \mu_2 < 0$
10.1 Inferences about the difference between two population means for large and independent samples

Let \( H_0: \mu_1 - \mu_2 = D_0 \), the value of the test statistic \( z \) for \( \bar{X}_1 - \bar{X}_2 \) is computed as

\[
z = \frac{(\bar{X}_1 - \bar{X}_2) - D_0}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}
\]

**Alternative** | **Reject \( H_0 \) if:** | **p-Value**
--- | --- | ---
\( H_a: \mu_1 - \mu_2 > D_0 \) | \( z > z_\alpha \) | Area under std normal curve right of \( z \)
\( H_a: \mu_1 - \mu_2 < D_0 \) | \( z < -z_\alpha \) | Area under std normal curve left of \( z \)
\( H_a: \mu_1 - \mu_2 \neq D_0 \) | \( |z| > z_{\alpha/2} \), that is \( z > z_{\alpha/2} \) or \( z < -z_{\alpha/2} \) | Twice area under std normal curve right of |\( z \)|
10.1 Inferences about the difference between two population means for large and independent samples

**Example:** Refer to the 2001 average salaries of full-time state employees in New York and Massachusetts. Test at the 1% significance level if the 2001 mean salaries of full-time state employees in New York and Massachusetts are different.
10.1 Inferences about the difference between two population means for large and independent samples

**Example:** Refer to the mean salaries offered to college students who graduated in 2002 with MIS and accounting majors. Test at 2.5% significance level if the mean salary offered to college students who graduated in 2002 with the MIS major is higher than that for accounting major.
Chapter 10: Estimation and hypothesis testing: two populations

**F distribution**

- The $F$ distribution is continuous and skewed to the right.
- The $F$ distribution has two degrees of freedom: $df_1$ for the numerator and $df_2$ for denominator.
- The units of an $F$ distribution, denoted by $F_{df_1, df_2, \alpha}$ are nonnegative.

![Graph of F distribution with different degrees of freedom](image-url)
Chapter 10: Estimation and hypothesis testing: two populations

\textit{F distribution:}

\textbf{Example:} Find the \( F \) value for 8 degrees of freedom for the numerator, 14 degrees of freedom for the denominator, and 0.05 area in the right tail of \( F \) distribution curve. \((F_{.05, 8,14})\)
Chapter 10: Estimation and hypothesis testing: two populations

**F distribution:**

**Example:** Find the F value for 10 degrees of freedom for the numerator, 12 degrees of freedom for the denominator, and 0.01 area in the right tail of F distribution curve.

**Example:** Find the F value for 15 degrees of freedom for the numerator, 15 degrees of freedom for the denominator, and 0.05 area in the right tail of F distribution curve.
Comparing Two Population Variances Using Independent Samples (One-Tailed)

- If both sampled populations are normal
- We test $\text{Ho: } \sigma_1^2 = \sigma_2^2 \ (\sigma_1^2/\sigma_2^2 = 1)$
  $\text{H1: } \sigma_1^2 > \sigma_2^2 \ (\sigma_1^2/\sigma_2^2 > 1)$
- Test statistic

  $$F = \frac{S_1^2}{S_2^2}$$

- Reject Ho in favor of H1 if:
  $$F > F_{\alpha, df1, df2}$$ or if
  p-value $< \alpha$

$F_{\alpha, df1, df2}$ is based on $(n1 - 1)$ and $(n2 - 1)$ df
Comparing Two Population Variances Using Independent Samples (One-Tailed)

- If both sampled populations are normal
- We test \( H_0: \sigma_1^2 = \sigma_2^2 \) \((\sigma_1^2/ \sigma_2^2 = 1)\)
- \( H_1: \sigma_1^2 < \sigma_2^2 \) \((\sigma_1^2/ \sigma_2^2 < 1)\)

- Test statistic

\[
F = \frac{S_2^2}{S_1^2}
\]

- Reject \( H_0 \) in favor of \( H_1 \) if:
  \[
  F > F_{\alpha, df1, df2} \text{ or if } p\text{-value} < \alpha
  \]

\( F_{\alpha, df1, df2} \) is based on \((n2 – 1)\) and \((n1 – 1)\) df
Comparing Two Population Variances Using Independent Samples (Two-Tailed)

- If both sampled populations are normal
- We test $H_0: \sigma_1^2 = \sigma_2^2 \ (\sigma_1^2 / \sigma_2^2 = 1)$
  $H_1: \sigma_1^2 \neq \sigma_2^2 \ (\sigma_1^2 / \sigma_2^2 \neq 1)$

- Test statistic
  $$F = \frac{\text{larger of } s_1^2 \text{ and } s_2^2}{\text{smaller of } s_1^2 \text{ and } s_2^2}$$

- Reject $H_0$ in favor of $H_1$ if:
  - $F > F_{\alpha/2, \ df_1, \ df_2}$ or if
  - $p$-value < $\alpha$

$F_{\alpha/2, \ df_1, \ df_2}$ is based on
$df_1 = \{\text{size of sample with larger variance}\} - 1$
$df_2 = \{\text{size of sample with smaller variance}\} - 1$
10.2 Inferences about the difference between two population means for small and indep. samples: Equal stand. dev.

The $t$ distribution is used to make inferences about $\mu_1 - \mu_2$ when the following assumptions hold true:

1. The two populations from which the two samples are drawn are (approximately) normally distributed.
2. The samples are small ($n_1 < 30$ and $n_2 < 30$) and independent.
3. The standard deviations ($\sigma_1$ and $\sigma_2$) of the two populations are unknown but they are assumed to be equal that is, $\sigma_1 = \sigma_2$. 
10.2 Inferences about the difference between two population means for small and indep. samples: Equal stand. dev.

- Since \( \sigma_1 = \sigma_2 \) and they are unknown, we replace them by \( s_p \), which is called the pooled sample variance:

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

- \( n_1 - 1 \) and \( n_2 - 1 \) are the degrees of freedom for samples 1 and 2, respectively, and \( n_1 + n_2 - 2 \) are the degrees of freedom for the two samples taken together.
10.2 Inferences about the difference between two population means for small and indep. samples: Equal stand. dev.

- The estimator of the standard deviation \( \bar{x}_1 - \bar{x}_2 \) is

\[
S_{\bar{x}_1-\bar{x}_2} = \sqrt{S^2_p \left( \frac{1}{n_1} + \frac{1}{n_2} \right)}
\]

- So, If two independent samples are drawn from populations that are normal with equal variances, 100(1 - \( \alpha \))\% confidence interval for \( \mu_1 - \mu_2 \) is

\[
(\bar{x}_1 - \bar{x}_2) \pm t_{\alpha/2} \sqrt{S^2_p \left( \frac{1}{n_1} + \frac{1}{n_2} \right)}
\]

Where \( t \) is based on \( \alpha /2 \) and \( n_1 + n_2 - 2 \) degrees of freedom.
10.2 Inferences about the difference between two population means for small and indep. samples: Equal stand. dev.

The value of the test statistic $t$ for $\overline{x}_1 - \overline{x}_2$ is

$$t = \frac{(\overline{x}_1 - \overline{x}_2) - D_0}{\sqrt{s_p^2 \left( \frac{1}{n_1} + \frac{1}{n_2} \right)}}$$

<table>
<thead>
<tr>
<th>Alternative</th>
<th>Reject $H_0$ if:</th>
<th>p-Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>$H_1: \mu_1 - \mu_2 &gt; D_0$</td>
<td>$t &gt; t_\alpha$</td>
<td>Area under t distribution right of $t$</td>
</tr>
<tr>
<td>$H_1: \mu_1 - \mu_2 &lt; D_0$</td>
<td>$t &lt; -t_\alpha$</td>
<td>Area under t distribution left of $t$</td>
</tr>
<tr>
<td>$H_1: \mu_1 - \mu_2 \neq D_0$</td>
<td>$</td>
<td>t</td>
</tr>
</tbody>
</table>
10.2 Inferences about the difference between two population means for small and indep. samples: Equal stand. dev.

Example: A chemical engineer wants to test which of two catalysts maximizes the hourly yield of a chemical process. The following table gives the data collected from that experiment.

<table>
<thead>
<tr>
<th>Catalyst A</th>
<th>Catalyst B</th>
</tr>
</thead>
<tbody>
<tr>
<td>801</td>
<td>752</td>
</tr>
<tr>
<td>814</td>
<td>718</td>
</tr>
<tr>
<td>784</td>
<td>776</td>
</tr>
<tr>
<td>836</td>
<td>742</td>
</tr>
<tr>
<td>820</td>
<td>763</td>
</tr>
</tbody>
</table>

Assume the data above are approximately normal.
1- Find a 99% confidence interval for the difference between the corresponding population means.
2- Test if $\mu_1 = \mu_2$ at 5% significance level.
Chapter 10: Estimation and hypothesis testing: two populations

10.3 Inferences about the difference between two population means for small and indep. samples: Unequal stand. dev.

- In the previous section (10.2) we learned how to make inferences about \( \mu_1 - \mu_2 \).
- What if the population standard deviations are not only unknown but also unequal?
- All the procedures (confidence interval and test of hypothesis) will remain the same except for two thing:
  - The degrees of freedom will no longer be \( n_1 + n_2 - 2 \)
  - The standard deviation of \( \bar{X}_1 - \bar{X}_2 \) is not calculated using the pooled standard deviation \( s_p \).
Chapter 10: Estimation and hypothesis testing: two populations

10.3 Inferences about the difference between two population means for small and indep. samples: Unequal stand. dev.

If:

1. The two populations from which the two samples are drawn are (approximately) normally distributed.
2. The samples are small \((n_1 < 30 \text{ and } n_2 < 30)\) and independent.
3. The standard deviations \((\sigma_1 \text{ and } \sigma_2)\) are unknown and not equal, that is, \(\sigma_1 \neq \sigma_2\)

The standard deviation of \(\bar{x}_1 - \bar{x}_2\) is

\[
S_{\bar{x}_1 - \bar{x}_2} = \sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}
\]
Chapter 10: Estimation and hypothesis testing: two populations

10.3 Inferences about the difference between two population means for small and indep. samples: Unequal stand. dev.

The degrees of freedom are given by

$$\text{df} = \frac{\left( \frac{s_1^2}{n_1} + \frac{s_2^2}{n_2} \right)^2}{\frac{(s_1^2/n_1)^2}{n_1-1} + \frac{(s_2^2/n_2)^2}{n_2-1}}$$

we always round down the df to the nearest integer.
10.3 Inferences about the difference between two population means for small and indep. samples: Unequal stand. dev.

Example: A chemical engineer wants to test which of two catalysts maximizes the hourly yield of a chemical process. The following table gives the data collected from that experiment.

<table>
<thead>
<tr>
<th>Catalyst A</th>
<th>Catalyst C</th>
</tr>
</thead>
<tbody>
<tr>
<td>801</td>
<td>880</td>
</tr>
<tr>
<td>814</td>
<td>850</td>
</tr>
<tr>
<td>784</td>
<td>690</td>
</tr>
<tr>
<td>836</td>
<td>755</td>
</tr>
<tr>
<td>820</td>
<td>880</td>
</tr>
</tbody>
</table>

Assume the data above are approximately normal.
1- Find a 99% confidence interval for the difference between the corresponding population means.
2- Test if $\mu_1 = \mu_2$ at 5% significance level.
Chapter 10: Estimation and hypothesis testing: two populations

10.4 Inferences About the Difference Between Two Population Means for Paired Samples

- In Sections 10.1, 10.2, and 10.3 we were concerned with estimation and hypothesis testing about the difference between two population means when the two samples were drawn independently from two different populations.
- This section describes estimation and hypothesis-testing procedures for the difference between two population means when the samples are dependent.
- In a case of two dependent samples, two data values—one for each sample—are collected from the same source (or element) and, hence, these are also called paired or matched samples.
10.4 Inferences About the Difference Between Two Population Means for Paired Samples

- For example, we may want to make inferences about the mean weight loss for members of a health club after they have gone through an exercise program for a certain period of time.
- Suppose we select a sample of 15 members of this health club and record their weights before and after the program.
- In this example, both sets of data are collected from the same 15 persons, once before and once after the program. Thus, although there are two samples, they contain the same 15 persons.
10.4 Inferences About the Difference Between Two Population Means for Paired Samples

- Paired or Matched Samples: Two samples are said to be *paired or matched samples* when for each data value collected from one sample there is a corresponding data value collected from the second sample, and both these data values are collected from the same source.

- The procedures to make confidence intervals and test hypotheses in the case of paired samples are different from the ones for independent samples discussed in earlier sections of this chapter.
10.4 Inferences About the Difference Between Two Population Means for Paired Samples

- In paired samples, the difference between the two data values for each element of the two samples is denoted by $d$. This value of $d$ is called the paired difference.
- We then treat all the values of $d$ as one sample and make inferences applying procedures similar to the ones used for one-sample cases in Chapters 8 and 9.
- Note that because each source (or element) gives a pair of values (one for each of the two data sets), each sample contains the same number of values. That is, both samples are the same size.
10.4 Inferences About the Difference Between Two Population Means for Paired Samples

- We denote the (common) sample size by \( n \), which gives the number of paired difference values denoted by \( d \). The degrees of freedom for the paired samples are \( n - 1 \).
- Let:

\[
\begin{align*}
\mu_d &= \text{the mean of the paired differences for the population} \\
\sigma_d &= \text{the standard deviation of the paired differences for the population, which is usually never known} \\
\bar{d} &= \text{the mean of the paired differences for the sample} \\
s_d &= \text{the standard deviation of the paired differences for the sample} \\
n &= \text{the number of paired difference values} \\
df &= n - 1
\end{align*}
\]
Chapter 10: Estimation and hypothesis testing: two populations

10.4 Inferences About the Difference Between Two Population Means for Paired Samples

➢ The values of the mean and standard deviation, $\bar{d}$ and $s_d$, of paired differences for two samples are calculated as:

\[
\bar{d} = \frac{\sum d}{n}
\]

\[
s_d = \sqrt{\frac{\sum d^2 - \left(\frac{\sum d}{n}\right)^2}{n - 1}}
\]
10.4 Inferences About the Difference Between Two Population Means for Paired Samples

In paired samples, instead of using $\bar{x}_1 - \bar{x}_2$ as the sample statistic to make inferences about $\mu_1 - \mu_2$, we use the sample statistic $\bar{d}$ to make inferences about $\mu_d$. Actually the value of $\bar{d}$ is always equal to $\bar{x}_1 - \bar{x}_2$, and the value of $\mu_d$ is always equal to $\mu_1 - \mu_2$.

Sampling Distribution, Mean, and Standard Deviation of $\bar{d}$

If the sample size is large ($n \geq 30$), then the sampling distribution of $\bar{d}$ is approximately normal with its mean and standard deviation given as

$$\mu_{\bar{d}} = \mu_d \text{ and } \sigma_{\bar{d}} = \frac{\sigma_d}{\sqrt{n}}$$
10.4 Inferences About the Difference Between Two Population Means for Paired Samples

- In paired samples, most of the times, the samples sizes are small and $\sigma_d$ is unknown.
- So, if
  - $n$ is less than 30
  - $\sigma_d$ is not known.
  - the population of paired differences is (approximately) normal
then the $t$ distribution is used to make inferences about $\mu_d$ and

$$S_{\bar{d}} = \frac{S_d}{\sqrt{n}}$$
10.4 Inferences About the Difference Between Two Population Means for Paired Samples

- The 100(1 - α)% confidence interval for \( \mu_d \) is

\[
\overline{d} \pm t \cdot s_{\overline{d}}
\]

- The value of the test statistic \( t \) for the mean of differences is

\[
t = \frac{\overline{d} - \mu_d}{s_{\overline{d}}}
\]
10.4 Inferences About the Difference Between Two Population Means for Paired Samples

Example: A researcher wanted to find the effect of a special diet on systolic blood pressure. She selected a sample of seven adults and put them on this dietary plan for three months. The following table gives the systolic blood pressures of these seven adults before and after the completion of this plan. Construct a 95% confidence interval for μₜ.

Using the 5% significance level, can we conclude that the mean of the paired differences μₜ is different from zero? Assume that the population of paired differences is (approximately) normally distributed.
### 10.4 Inferences About the Difference Between Two Population Means for Paired Samples

The table below shows the differences in paired samples before and after some treatment.

<table>
<thead>
<tr>
<th>Before</th>
<th>After</th>
<th>$d$</th>
<th>$d^2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>210</td>
<td>193</td>
<td>17</td>
<td>289</td>
</tr>
<tr>
<td>180</td>
<td>186</td>
<td>-6</td>
<td>36</td>
</tr>
<tr>
<td>195</td>
<td>186</td>
<td>9</td>
<td>81</td>
</tr>
<tr>
<td>220</td>
<td>223</td>
<td>-3</td>
<td>9</td>
</tr>
<tr>
<td>231</td>
<td>220</td>
<td>11</td>
<td>121</td>
</tr>
<tr>
<td>199</td>
<td>183</td>
<td>16</td>
<td>256</td>
</tr>
<tr>
<td>224</td>
<td>233</td>
<td>-9</td>
<td>81</td>
</tr>
</tbody>
</table>

Summarizing:

- $\Sigma d = 35$
- $\Sigma d^2 = 873$
Chapter 10: Estimation and hypothesis testing: two populations

10.5 Inferences About the Difference Between Two Population Proportions for Large and Indep. Samples

- As we learned in the previous chapters, many times in our life we need to deal with proportions.
- In this section we will learn how to construct a confidence interval and test a hypothesis about the difference between two population proportions.
- We may want to estimate the difference between the proportions of defective items produced on two different machines.
- We may want to test the hypothesis that the proportion of defective items produced on Machine I is different from the proportion of defective items produced on Machine II.
10.5 Inferences About the Difference Between Two Population Proportions for Large and Indep. Samples

In this case, we are to test the null hypothesis \( p_1 - p_2 = 0 \) against the alternative hypothesis \( p_1 - p_2 \neq 0 \).

The sample statistic that is used to make inferences about \( p_1 - p_2 \) is \( \hat{p}_1 - \hat{p}_2 \) where \( \hat{p}_1 \) and \( \hat{p}_2 \) are the proportions for two large and independent samples.

For two large and independent samples, the sampling distribution of \( \hat{p}_1 - \hat{p}_2 \) is (approximately) normal with its mean and standard deviation given as

\[
\mu_{\hat{p}_1 - \hat{p}_2} = p_1 - p_2 \quad \text{and} \quad \sigma_{\hat{p}_1 - \hat{p}_2} = \sqrt{\frac{p_1 q_1}{n_1} + \frac{p_2 q_2}{n_2}}
\]
10.5 Inferences About the Difference Between Two Population Proportions for Large and Indep. Samples

- If \( np \) and \( nq \) for both samples are greater than 5, then the 100(1 - \( \alpha \))% confidence interval for \( p_1 - p_2 \) is

\[
(\hat{p}_1 - \hat{p}_2) \pm zs_{\hat{p}_1-\hat{p}_2}
\]

where

\[
s_{\hat{p}_1-\hat{p}_2} = \sqrt{\frac{\hat{p}_1\hat{q}_1}{n_1} + \frac{\hat{p}_2\hat{q}_2}{n_2}}
\]

- When testing about \( p_1 = p_2 \), we assume it is true. So we need to find a common value for \( p_1 \) and \( p_2 \) (pooled).

\[
p = \frac{x_1 - x_2}{n_1 - n_2}
\]
10.5 Inferences About the Difference Between Two Population Proportions for Large and Indep. Samples

- The estimate of the standard deviation of for \( \hat{p}_1 - \hat{p}_2 \) is

\[
S_{\hat{p}_1 - \hat{p}_2} = \sqrt{pq \left( \frac{1}{n_1} + \frac{1}{n_2} \right)}
\]

- The value of the test statistic for \( \hat{p}_1 - \hat{p}_2 \) is

\[
Z = \frac{(\hat{p}_1 - \hat{p}_2) - (p_1 - p_2)}{S_{\hat{p}_1 - \hat{p}_2}}
\]
10.5 Inferences About the Difference Between Two Population Proportions for Large and Indep. Samples

Example: researcher wanted to estimate the difference between the percentages of users of two toothpastes who will never switch to another toothpaste. In a sample of 500 users of Toothpaste A taken by this researcher, 100 said that they will never switch to another toothpaste. In another sample of 400 users of Toothpaste B taken by the same researcher, 68 said that they will never switch to another toothpaste.

a. Let \( p_1 \) and \( p_2 \) be the proportions of all users of Toothpastes A and B, respectively, who will never switch to another toothpaste. What is the point estimate of \( p_1 - p_2 \)?

b. Construct a 97% confidence interval for the difference between the proportions of all users of the two toothpastes who will never switch.

c. At the 1% significance level, can we conclude that \( p_1 \) is higher than \( p_2 \)?