Chapter 9.4: Inferences Concerning a Difference Between Population Proportions

Instructor: Dr. Arnab Maity
We have learned about methods of comparing means of two different populations. In this chapter, we will learn about how to compare two population proportions.

**Example (from book):** The article Aspirin Use and Survival After Diagnosis of Colorectal Cancer (J. of the Amer. Med. Assoc., 2009: 649658) reported that of 549 study participants who regularly used aspirin after being diagnosed with colorectal cancer, there were 81 colorectal cancer-specific deaths, whereas among 730 similarly diagnosed individuals who did not subsequently use aspirin, there were 141 colorectal cancer-specific deaths. Does this data suggest that the regular use of aspirin after diagnosis will decrease the incidence rate of colorectal cancer-specific deaths?

<table>
<thead>
<tr>
<th></th>
<th>Total participants</th>
<th>Deaths</th>
</tr>
</thead>
<tbody>
<tr>
<td>Aspirin users</td>
<td>549</td>
<td>81</td>
</tr>
<tr>
<td>Non-users</td>
<td>730</td>
<td>141</td>
</tr>
</tbody>
</table>

(a) What are the population parameter we are interested in?

\[ p_1 = \text{true proportion of deaths among Aspirin users} \]

\[ p_2 = \text{true proportion of deaths among non-users} \]

Parameter \((p_1 - p_2)\) is of interest.

(b) Provide a point estimate of the parameter of interest.

\[ \hat{p}_1 = \frac{81}{549} \]

\[ \hat{p}_2 = \frac{141}{730} \]

\((p_1 - p_2)\) can be estimated as \(\hat{p}_1 - \hat{p}_2 = -0.046\)

(c) What hypothesis would you like to test in this case? Provide \(H_0\) and \(H_a\).

\[ H_0: p_1 - p_2 = 0 \]

\[ H_a: p_1 - p_2 < 0 \]
Setting: We have two populations. Regard an individual or object as a success $S$ if he/she/it processes some characteristic of interest. Define

\[ p_1 = \text{true proportion of } S \text{ in population 1} \]
\[ p_2 = \text{true proportion of } S \text{ in population 2}. \]

We are interested in the difference $\Delta = p_1 - p_2$.

We observe a sample of size $m$ from the first population and independently a sample of size $n$ is selected from the second population.

- Let $X$ denote the number of $S$'s in the first sample
- Let $Y$ denote the number of $S$'s in the second sample
- We can assume $X \sim \text{Binomial}(m, p_1)$ and $Y \sim \text{Binomial}(n, p_2)$ and that they are independent

Provide point estimators for the following quantities

- Point estimator for $p_1$ is $\hat{p}_1 = \frac{X}{m}$
- Point estimator for $p_2$ is $\hat{p}_2 = \frac{Y}{n}$
- Point estimator for $p_1 - p_2$ is $\hat{\Delta} = \frac{X}{m} - \frac{Y}{n}$

Show that $\hat{\Delta}$ is an unbiased estimator of $p_1 - p_2$.

\[
E(\hat{\Delta}) = E\left(\frac{X}{m} - \frac{Y}{n}\right) = \frac{E(X)}{m} - \frac{E(Y)}{n}
\]

Compute the variance of $\hat{\Delta}$.

\[
\text{Var}(\hat{\Delta}) = \text{Var}\left(\frac{X}{m} - \frac{Y}{n}\right) = \frac{\text{Var}(X)}{m^2} + \frac{\text{Var}(Y)}{n^2} = \frac{mp_1(1-p_1)}{m^2} + \frac{np_2(1-p_2)}{n^2}
\]
Recall test for one proportion

\[ H_0: \ p = p_0 \]

\[
\hat{p} = \frac{X}{m} \quad v(\hat{p}) = \frac{p(1-p)}{m}
\]

Test stat

\[
z = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{m}}}
\]

Two proportions

\[ H_0: \ p_1 - p_2 = 0 \]

\[ \hat{p}_1 - \hat{p}_2 = 0 \]

\[
z = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\frac{p_1(1-p_1)}{m} + \frac{p_2(1-p_2)}{n}}}
\]

\[ \hat{p} = \frac{X+Y}{n+m} \quad (0.174) \]
A large sample testing procedure

We consider the null hypothesis

$$H_0 : p_1 - p_2 = 0.$$  

When $H_0$ is true, this implies that the two proportions are same. Let $p$ denote this common value.

**Result:** If $H_0$ is true then

$$Z = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p}(1 - \hat{p}) \left(\frac{1}{m} + \frac{1}{n}\right)}}$$

follows (approximately) a standard normal distribution $N(0, 1)$ when sample sizes are large enough.

Note that $Z$ as defined above can not be a test statistic (explain why). How do you estimate the common proportion $p$ (when $H_0$ is true)?

$$\text{pooled estimate of } p \text{ is } \hat{p} = \frac{x+y}{m+n}$$

Null hypothesis: $H_0 : p_1 - p_2 = 0$.

Test statistic value:

$$z = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p}(1 - \hat{p}) \left(\frac{1}{m} + \frac{1}{n}\right)},}$$

where $\hat{p}_1 = x/m, \hat{p}_2 = y/n, \hat{p} = (x+y)/(m+n)$ is the pooled proportion estimate.

<table>
<thead>
<tr>
<th>Alternative Hypothesis</th>
<th>Rejection Region for Level $\alpha$ Test</th>
</tr>
</thead>
<tbody>
<tr>
<td>$H_a : p_1 - p_2 &gt; 0$</td>
<td>$z \geq z_\alpha$</td>
</tr>
<tr>
<td>$H_a : p_1 - p_2 &lt; 0$</td>
<td>$z \leq -z_\alpha$</td>
</tr>
<tr>
<td>$H_a : p_1 - p_2 \neq 0$</td>
<td>$z \leq -z_{\alpha/2}$ OR $z \geq z_{\alpha/2}$</td>
</tr>
</tbody>
</table>

A p-value is calculated in the same way as for previous z tests.

The test can be applied as long as $m\hat{p}_1, m(1 - \hat{p}_1), n\hat{p}_2$ and $n(1 - \hat{p}_2)$ are greater than 10.
1. The article Aspirin Use and Survival After Diagnosis of Colorectal Cancer (J. of the Amer. Med. Assoc., 2009: 649658) reported that of 549 study participants who regularly used aspirin after being diagnosed with colorectal cancer, there were 81 colorectal cancer-specific deaths, whereas among 730 similarly diagnosed individuals who did not subsequently use aspirin, there were 141 colorectal cancer-specific deaths.

<table>
<thead>
<tr>
<th>Total participants</th>
<th>Deaths</th>
</tr>
</thead>
<tbody>
<tr>
<td>Aspirin users</td>
<td>549</td>
</tr>
<tr>
<td>Non-users</td>
<td>730</td>
</tr>
</tbody>
</table>

Does this data suggest that the regular use of aspirin after diagnosis will decrease the incidence rate of colorectal cancer-specific deaths? Conduct a hypothesis test using $\alpha = 0.05$.

\[
\hat{p}_1 = \frac{81}{549}, \quad \hat{p}_2 = \frac{141}{730}, \quad \hat{p}_1 - \hat{p}_2 = -0.046
\]

\[
\hat{p} = \frac{81 + 141}{549 + 730} = 0.174
\]

\[
z = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p}(1-\hat{p})(\frac{1}{n_1} + \frac{1}{n_2})}} = -2.13
\]

$H_0: \hat{p}_1 - \hat{p}_2 = 0$ \qquad versus \qquad $H_a: \hat{p}_1 - \hat{p}_2 < 0$

**Rejection Region:** \( \{ z \leq -2 \times 3 \} \)

\(-1.645\)

**We reject** $H_0$ in favor of $H_a: \hat{p}_1 - \hat{p}_2 < 0$
A Large-Sample Confidence Interval

While testing for difference \( p_1 - p_2 \) is important, sometimes one needs to estimate the difference and provide appropriate confidence intervals. Such an interval can be constructed using large sample techniques.

Recall that we computed the variance of \( \hat{p}_1 - \hat{p}_2 \) as

\[
\text{var}(\hat{p}_1 - \hat{p}_2) = \frac{p_1(1 - p_1)}{m} + \frac{p_2(1 - p_2)}{n}.
\]

Thus the estimated standard deviation is

\[
\sqrt{\frac{\hat{p}_1(1 - \hat{p}_1)}{m} + \frac{\hat{p}_2(1 - \hat{p}_2)}{n}}.
\]

A CI for \( p_1 - p_2 = \) is given by

\[
(\hat{p}_1 - \hat{p}_2) \pm z_{a/2} \sqrt{\frac{\hat{p}_1(1 - \hat{p}_1)}{m} + \frac{\hat{p}_2(1 - \hat{p}_2)}{n}}
\]

The interval can be used as long as \( m\hat{p}_1, m(1 - \hat{p}_1), n\hat{p}_2 \) and \( n(1 - \hat{p}_2) \) are greater than 10.

2. (Previous example continued). Construct a 95% confidence interval for \( p_1 - p_2 \).

\[
\hat{p}_1 - \hat{p}_2 = -0.046
\]

\[
2z_{1/2} = 1.96
\]

\[
\sqrt{\frac{\hat{p}_1(1 - \hat{p}_1)}{m} + \frac{\hat{p}_2(1 - \hat{p}_2)}{n}} = 0.021
\]

\[
(-0.046 \pm (1.96)(0.021))
\]