Comparing two proportions

Recall: In a *two-independent sample* problem, we want to compare two populations or the responses to two different treatments using data from two independent samples.

When we are interested in comparing the *proportions of successes* in two groups, the notation is:

<table>
<thead>
<tr>
<th>Population</th>
<th>Sample 1</th>
<th>Sample 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>$p_1$</td>
<td>$n_1$</td>
<td>$\hat{p}_1$</td>
</tr>
<tr>
<td>$p_2$</td>
<td>$n_2$</td>
<td>$\hat{p}_2$</td>
</tr>
</tbody>
</table>

Example: Do seatbelts protect children during car accidents?

- study of deaths among children involved in car accidents during an 18-month period
- two simple random samples
  - one sample from population of children who were wearing seatbelts at the time of car accident
  - one sample from population of children who were not wearing seatbelts at the time of car accident
- parameters of interest: proportions of children who die in car accidents from each of these populations
<table>
<thead>
<tr>
<th>Population</th>
<th>Sample proportion</th>
<th>Sample size</th>
<th>Sample proportion</th>
</tr>
</thead>
<tbody>
<tr>
<td>seatbelts</td>
<td>$p_1$</td>
<td>123</td>
<td>$\frac{3}{123} = 0.024$</td>
</tr>
<tr>
<td>no seatbelts</td>
<td>$p_2$</td>
<td>290</td>
<td>$\frac{13}{290} = 0.045$</td>
</tr>
</tbody>
</table>

To determine whether the study provides significant evidence that seatbelts affect the proportion of kids who die if they are involved in a car accident, we test the hypotheses:

\[
H_0 : p_1 - p_2 = 0 \quad \text{or} \quad H_0 : p_1 = p_2 \\
H_a : p_1 - p_2 \neq 0 \quad \text{or} \quad H_a : p_1 \neq p_2
\]

To estimate how large the difference is, we compute a confidence interval for the difference $p_1 - p_2$.

### Confidence intervals for comparing two proportions

To compute a c.i., we estimate the population proportions $p_1$ and $p_2$ by their corresponding sample proportions $\hat{p}_1$ and $\hat{p}_2$.

The resulting standard error of $\hat{p}_1 - \hat{p}_2$ is

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

The approximate level-C two-sided confidence interval is

\[
(\hat{p}_1 - \hat{p}_2) \pm z^* SE
\]

where $z^*$ is the upper $\frac{1-C}{2}$ standard normal cut-off.

### The sampling distribution of $\hat{p}_1 - \hat{p}_2$

- When both samples are large, the distribution of $\hat{p}_1 - \hat{p}_2$ is approximately normal.
- The mean of this normal distribution is $p_1 - p_2$.
- The standard deviation of the difference is

\[
\sqrt{\frac{p_1(1 - p_1)}{n_1} + \frac{p_2(1 - p_2)}{n_2}}
\]

Because we don’t know $p_1$ and $p_2$, we must replace them with estimates. These estimates will be different for confidence intervals versus hypothesis tests.

### Rules of thumb for using this confidence interval:

1. Both populations are at least 10 times as large as the samples.
2. The counts of successes and failures are 5 or more in each sample.
Car accident example

<table>
<thead>
<tr>
<th>Population proportion</th>
<th>Sample size</th>
<th>Sample proportion</th>
</tr>
</thead>
<tbody>
<tr>
<td>Seatbelts $p_1$</td>
<td>123</td>
<td>$\frac{3}{123} = 0.024$</td>
</tr>
<tr>
<td>No seatbelts $p_2$</td>
<td>290</td>
<td>$\frac{13}{290} = 0.045$</td>
</tr>
</tbody>
</table>

The $95\%$ two-sided confidence interval is

\[
(\hat{p}_1 - \hat{p}_2) \pm z*SE = (0.024 - 0.045) \pm (1.96)(0.0184) = -0.021 \pm 0.036 = (-0.057, 0.015)
\]

We are $95\%$ confident that this interval covers the true difference between the proportions of kids who die from car accidents in the population who were wearing seatbelts at the time of the accident vs. the population who were not.

The interval includes the value $0$, so it is plausible based on this data that there is no difference!

The hypothesis test

For the formal hypothesis test, the hypotheses are:

\[
H_0 : p_1 - p_2 = 0 \quad \text{or} \quad H_0 : p_1 = p_2 \\
H_a : p_1 - p_2 \neq 0 \quad \text{or} \quad H_a : p_1 \neq p_2
\]

Suppose we had set $\alpha = .05$ when we were designing the study.

- We must standardize $\hat{p}_1 - \hat{p}_2$ to get a $z$ statistic.
- We do this under the assumption that $H_0$ is true, that is that $p_1$ and $p_2$ have the same value $p$.
  - Instead of estimating $p_1$ and $p_2$ separately in the standard deviation of the difference, we pool the two samples and use the overall sample proportion to estimate the single population parameter $p$.
  - The pooled sample proportion is
    \[
    \hat{p} = \frac{\text{total count of successes in both samples}}{n_1 + n_2}
    \]

The test statistic is

\[
z = \frac{(\hat{p}_1 - \hat{p}_2) - 0}{\sqrt{\hat{p}(1 - \hat{p}) \left( \frac{1}{n_1} + \frac{1}{n_2} \right)}}
\]
Car accident example

The pooled sample proportion is:

\[ \hat{p} = \frac{3 + 13}{123 + 290} = \frac{16}{413} = 0.039 \]

The z statistic is:

\[ z = \frac{(0.024 - 0.045) - 0}{\sqrt{(0.039)(0.961)(\frac{1}{123} + \frac{1}{290})}} = -1.01 \]

To get the p-value for the two sided test, we look for the area under a standard normal curve that is farther away from 0 than -1.01 in either direction.

Table A gives .156 as the area to the left of -1.01.

\[ p-value = 2 \times (0.156) = 0.312 \]

We cannot reject the null hypothesis. This particular set of sample data does not provide evidence that the proportion of children dying in car accidents differs between the population of those wearing seatbelts at the time of the accident and the population of those not.

Contingency Tables and the Chi-square test

An equivalent way of comparing two population proportions, that generalizes to more than two populations.

Begin by presenting the data as a two-way table, with rows representing levels of one variable and columns representing levels of the other.

Seatbelt example:

<table>
<thead>
<tr>
<th>Seatbelts</th>
<th>Died</th>
<th>Did not die</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Yes</td>
<td>3</td>
<td>120</td>
<td>123</td>
</tr>
<tr>
<td>No</td>
<td>13</td>
<td>277</td>
<td>290</td>
</tr>
<tr>
<td>Total</td>
<td>16</td>
<td>397</td>
<td>413</td>
</tr>
</tbody>
</table>

To test the hypotheses

\[ H_0 : p_1 - p_2 = 0 \text{ or } H_0 : p_1 = p_2 \]
\[ H_a : p_1 - p_2 \neq 0 \text{ or } H_a : p_1 \neq p_2 \]

using the two-way table, we must compute the expected counts. These are the counts we would expect (except for random variation) if \( H_0 \) were true.

\[ \text{expected count} = \frac{\text{row total} \times \text{column total}}{\text{table total}} \]

If \( H_0 \) were true, there would be just one \( p \) shared by both populations.

Our best estimate is again the pooled sample proportion \( \hat{p} = 0.039 \).
The Chi-square statistic

The statistic that we use for this test is the sum over all the cells in the table of \( \frac{(\text{observed count} - \text{expected count})^2}{\text{expected count}} \).

The formula in mathematical notation is

\[
X^2 = \frac{\sum_{i=1}^{rc} (O_i - E_i)^2}{E_i}
\]

where \( rc \) is the total number of cells in the table.

- \( r \) is the number of rows
- \( c \) is the number of columns

The Chi-Square Test

- Recall that the expected counts were computed under the assumption that the null hypothesis was true.
- We can test the null hypothesis by determining whether the differences between the observed and expected counts are too large to be likely to be due to chance.
- Notation
  - \( O_i \) is the observed count in cell \( i \)
  - \( E_i \) is the expected count in cell \( i \)

- Think of \( X^2 \) as a measure of the distance of the observed counts from the expected counts.
- Like any distance, \( X^2 \)
  - is always zero or positive
  - is zero only when the observed counts are exactly equal to the expected counts
- Large values of \( X^2 \) are evidence against \( H_0 \).
  - indicate that observed counts are far away from what we would expect if \( H_0 \) were true.
- The Chi-square statistic \( X^2 \) follows a Chi-square distribution (\( \chi^2 \) distribution) with \( (r - 1)(c - 1) \) degrees of freedom.
The Chi-Square test for the car-accidents example

$$X^2 = \frac{(3 - 4.8)^2}{4.8} + \frac{(120 - 118.2)^2}{118.2} + \frac{(13 - 11.3)^2}{11.3} + \frac{(277 - 278.7)^2}{278.7}$$

$$= 0.969$$

Since we have 2 rows and 2 columns in our table, the degrees of freedom is

$$(r - 1)(c - 1) = 1(1) = 1$$

We will carry out our hypothesis test at $\alpha = .05$. According to Table E, the .05 cutoff under the Chi-square distribution with 1 degree of freedom is 3.84.

The Chi-square test is always 2-sided. For the Chi-square test, we always reject if the test statistic is larger than the cutoff value.

Our computed value, 0.969, is smaller than this cutoff. Therefore we cannot reject $H_0$. This result is consistent with what we got with both the confidence interval and the z test.

When does Chi-square test give accurate enough inference?

- rule of thumb: when expected counts in all cells are $\geq 5$
- not quite satisfied in this example