Comparing two proportions: Confidence intervals and hypothesis test for differences

**Standard confidence interval:** point estimate ± margin of error.

Point estimate of mean difference, $D = \mu_{\hat{p}_1} - \mu_{\hat{p}_2} = \hat{p}_1 - \hat{p}_2$ by the addition rule for means

Margin of error: $m = z \times SE_D$

$$SE_D = \sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}$$

by the addition rule for variances

So the level C confidence interval for the difference between 2 proportions is

$$D \pm m$$

or

$$(\hat{p}_1 - \hat{p}_2) \pm \sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}$$

**Wilson estimate or “Plus Four” confidence interval:** point estimate ± margin of error.

Adjust the population proportions, again by adding four total extra observations (2 to each sample)

$$\hat{p}_1 = \frac{X_1 + 1}{n_1 + 2}$$

$$\hat{p}_2 = \frac{X_2 + 1}{n_2 + 2}$$

then it is all almost the same

Point estimate of mean difference, $D = \mu_{\hat{p}_1} - \mu_{\hat{p}_2} = \tilde{p}_1 - \tilde{p}_2$

Margin of error: $m = z \times SE_D$

$$SE_D = \sqrt{\frac{\tilde{p}_1(1-\tilde{p}_1)}{n_1 + 2} + \frac{\tilde{p}_2(1-\tilde{p}_2)}{n_2 + 2}}$$

by the addition rule for variances

So the level C “Plus Four” confidence interval for the difference between 2 proportions is

$$D \pm m$$

or

$$(\tilde{p}_1 - \tilde{p}_2) \pm \sqrt{\frac{\tilde{p}_1(1-\tilde{p}_1)}{n_1 + 2} + \frac{\tilde{p}_2(1-\tilde{p}_2)}{n_2 + 2}}$$

Again, minitab can also do the calculations based directly on the Binomial. If you use the Normal approximation, use the Wilson estimate to improve accuracy!
Hypothesis test: $H_0: p_1 = p_2$

$$z = \frac{\hat{p}_1 - \hat{p}_2}{SE_{\hat{p}_p}}$$

where the denominator is the pooled standard error of the difference

$$SE_{\hat{p}_p} = \sqrt{\hat{p}(1 - \hat{p}) \left( \frac{1}{n_1} + \frac{1}{n_2} \right)}$$

and the pooled proportion is

$$\hat{p} = \frac{X_1 + X_2}{n_1 + n_2}$$

All of the usual terms of hypothesis testing apply…

Examples:

1.) **Inference for a single proportion:** Find a 95% confidence interval for the proportion of students who saw *Harry Potter and the Goblet of Fire* over the break.

2.) **Inference for two proportions:** Find a 95% C.I. for the difference in the proportions of male and female students who went shopping on “Black Friday.”

Relationships between categorical variables – Two-way Tables

<table>
<thead>
<tr>
<th></th>
<th>Female</th>
<th>Male</th>
<th>Totals</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fresh-Soph</td>
<td>7</td>
<td>11</td>
<td>18</td>
</tr>
<tr>
<td>Jun-Senior</td>
<td>2</td>
<td>7</td>
<td>9</td>
</tr>
<tr>
<td>Total</td>
<td>9</td>
<td>18</td>
<td>27</td>
</tr>
</tbody>
</table>

Joint distributions: Cell count divided by the grand total in the sample

<table>
<thead>
<tr>
<th></th>
<th>Female</th>
<th>Male</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fresh-Soph</td>
<td>0.259</td>
<td>0.407</td>
</tr>
<tr>
<td>Jun-Senior</td>
<td>0.074</td>
<td>0.259</td>
</tr>
</tbody>
</table>
**Marginal distributions:** Row or column totals divided by the grand total

Female: 0.333  Male: 0.667
Fresh-Soph: 0.667  Jun-Senior: 0.333

**Conditional distributions:** Cell counts divided by row or column totals
For Females:  Fresh-Soph: 0.778  Jun-Senior: 0.222
For Males:  Fresh-Soph: 0.611  Jun-Senior: 0.389

Fresh-Soph:  Female: 0.389  Male: 0.611
Jun-Senior:  Female: 0.222  Male: 0.778