STAT 260 Lecture Notes
Set 30 - Hypothesis Testing With Two Proportions

Sample 1: n_1 observations, population proportion p_1, sample proportion $\hat{p}_1$
Sample 2: n_2 observations, population proportion p_2, sample proportion $\hat{p}_2$

We want to estimate $p_1 - p_2$ to see if there is a difference between the two population proportions.

$$\widehat{p_1 - p_2} = \hat{p}_1 - \hat{p}_2 = \frac{X_1}{n_1} - \frac{X_2}{n_2}$$

Theorem: For large sample sizes we have that

$$Z = \frac{(\hat{p}_1 - \hat{p}_2) - \overbrace{(p_1 - p_2)}^{\text{from } H_0}}{\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}}$$

$\hat{p}$'s are in the setup of the question

is approximately standard normal.

A $(1-\alpha)100\%$ confidence interval for $p_1 - p_2$ then is

$$(\hat{p}_1 - \hat{p}_2) \pm z_{\alpha/2}\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}$$

Example 1

In a greenhouse 50 tomato seeds are planted and 90% germinate, whereas 80 tomato seeds are planted outdoors and 95% germinate. Test if there is a difference in the germination proportion to planting in the greenhouse versus planting outdoors.

- We are testing the parameter

$p_1 - p_2$ = true difference in proportion of germination between greenhouse (p_1) and outdoors (p_2)

- The null and alternative hypotheses are

$H_0: p_1 = p_2$ $\Longrightarrow$ $H_0: p_1 - p_2 = 0$ $\leftarrow$ Write in form of $p_1 - p_2$ because we need this when we standardize.

$H_1: p_1 \neq p_2$ $\qquad$ $H_1: p_1 - p_2 \neq 0$

- The test statistic is

$$z = \frac{(\hat{p}_1 - \hat{p}_2) - (p_1 - p_2)}{\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}}$$

- The observed value of the test statistic is

$\leftarrow p_1 - p_2$ value from H_0.

$$z_{obs} = \frac{(.90 - .95) - (0)}{\sqrt{\frac{.90(.10)}{50} + \frac{(.95)(.05)}{80}}} = -1.02$$

- The p-value is

$$\begin{aligned} p\text{-value} &= P(Z < -1.02) + P(Z > 1.02) \\ &= 2 \cdot P(Z < -1.02) \\ &= 2(0.1539) \\ &= 0.3078 \end{aligned}$$

- Our conclusion is p-value $= 0.3078$ is bigger than any reasonable α value (or we can use strength of evidence and say there is little to no evidence against H_0 since p-value > 0.10) so our p-value is big and we keep H_0.

There is not enough evidence to say there is a difference in proportions of germination between the greenhouse and the outdoors.

Example 2

Using the setup from Example 1, construct a 95% confidence interval for $p_1 - p_2$.

$$(\hat{p}_1 - \hat{p}_2) \pm z_{\alpha/2} \sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}$$

$$= (.90 - .95) \pm 1.96 \sqrt{\frac{(.90)(.10)}{50} + \frac{(.95)(.05)}{80}} = [-0.1459, 0.0459]$$

Interpretation: 0 is in the CI so it is reasonable to say that $p_1 - p_2 = 0$, or $p_1 = p_2$.

The CI we found of [-0.1459, 0.0459] estimates

$p_1 - p_2$.

greenhouse outdoors

The choice to put the greenhouse first was arbitrary.
If we swap the order of the groups we get a CI for

$p_1 - p_2$ of [-0.0459, 0.1459].

outdoors greenhouse

Swapping order of groups swaps signs and order of the upper and lower bounds.

Note that 0 is still in this CI though, so the interpretation of saying $p_1 - p_2 = 0$, so $p_1 = p_2$ is the same!

In other words: The order of your groups does not matter, we get the same conclusion either way.