Confidence interval for the difference of proportions

In the same way that we can run a hypothesis test on the difference of means, we can do hypothesis testing on the difference of proportions. Before we work through the entire hypothesis test though, let's start with simply building a confidence interval around the difference of proportions.

The point estimator and standard error

Imagine that a department store wants to know whether the proportion of walk-in customers who complete a purchase at its New York store is different than the same proportion at its San Francisco location.

They could define Population 1 as all of their New York customers and Population 2 as all of their San Francisco customers. They're trying to estimate the difference between p_1 and p_2 (the proportion of walk in customers who complete a purchase in New York and San Francisco, respectively), which means they're looking for $p_1 - p_2$. To do so, they can take a sample of size n_1 in New York and a sample of size n_2 in San Francisco, and find

$$\hat{p}_1 = \frac{x_1}{n_1} \text{ and } \hat{p}_2 = \frac{x_2}{n_2}$$

where x_1 and x_2 are the number of "successes" in samples n_1 and n_2, respectively. Then the point estimator of $p_1 - p_2$ is $\hat{p}_1 - \hat{p}_2$. Then the sampling distribution of $\hat{p}_1 - \hat{p}_2$ will have a mean of $p_1 - p_2$ and a standard error of

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

and will be normal as long as $n_1p_1 \geq 5$, $n_1(1-p_1) \geq 5$, $n_2p_2 \geq 5$, and $n_2(1-p_2) \geq 5$.

Confidence interval around the difference of proportions

The confidence interval around the point estimator $\hat{p}_1 - \hat{p}_2$ will be given by $(\hat{p}_1 - \hat{p}_2) \pm$ margin of error, where the margin of error is

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

So the confidence interval formula is

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

Let’s work through an example so that make sure we know how to calculate the confidence interval for the difference of proportions.

Example

A team of scientists wants to determine whether a new cholesterol lowering drug is more effective than a previous version. They take two random samples of 250 people. For three months, the first group gets the new drug while the second group gets the old drug. 155 people from the first group and 107 people from the second group show decreased cholesterol levels. Estimate a 99% confidence interval for the difference of proportions.

We know $n_1 = 250$ and $n_2 = 250$, that $z_{\alpha/2} = 2.58$ for a 99% confidence level, and that the sample proportions are

$$\hat{p}_1 = \frac{155}{250} = 0.620$$

$$\hat{p}_2 = \frac{107}{250} = 0.428$$

Substituting all these values into the confidence interval formula gives

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

$$(a, b) = (0.620 - 0.428) \pm 2.58\sqrt{\frac{0.620(1 - 0.620)}{250} + \frac{0.428(1 - 0.428)}{250}}$$

$$(a, b) = 0.192 \pm 2.58\sqrt{\frac{0.620(0.38)}{250} + \frac{0.428(0.572)}{250}}$$

$$(a, b) = 0.192 \pm 2.58\sqrt{\frac{0.2356}{250} + \frac{0.244816}{250}}$$

$$(a, b) = 0.192 \pm 2.58\sqrt{\frac{0.480416}{250}}$$

Simplify to find the confidence interval.

$$(a,b) \approx 0.192 \pm 0.113$$

$$(a,b) \approx (0.192 - 0.113, 0.192 + 0.113)$$

$$(a,b) \approx (0.079, 0.305)$$

$$(a,b) \approx (0.08, 0.31)$$

With 99% confidence, we can say that the new drug was associated with a greater decrease in cholesterol than the old drug, and that the difference of proportions lies between 0.08 and 0.31.

When the confidence interval contains 0

When the confidence interval we calculate contains 0, such that the lower end of the confidence interval is negative and the upper end of the confidence interval is positive, there's likely no difference in proportions.

On the other hand, when the confidence interval doesn't contain 0, like in this last example where $(a,b) \approx (0.08, 0.31)$, then it's likely that there *is* a difference in proportions.