Introduction to Statistical Ideas and Methods

Comparing Two Groups

Comparing Two Proportions

In previous lectures we have learned how to study the results of a single sample. Most recently, we talked about having paired samples where we can just look at the difference between the result in the second sample minus the first sample, after which it reduces to just having one sample again. Now we are going to consider the case where we have two independent samples and the question is, what inferential procedures can we use in this case? Suppose we have two independent samples and we are interested in the proportion of success in each. These could be two different polls, performed independently on two different groups of people, for example. In this case, the first sample might have been of size n_1 and the second one of size n_2, since they could potentially be of different sizes. And we would probably observe different fractions of successes, denoted by $\hat{p}_1$ and $\hat{p}_2$.

We are interested in the difference in proportions $p_1 - p_2$. Using our sample estimates $\hat{p}_1$ and $\hat{p}_2$ it is not hard to show that

$$E(\hat{p}_1 - \hat{p}_2) = p_1 - p_2$$

and

$$\begin{aligned} Var(\hat{p}_1 - \hat{p}_2) &= Var(\hat{p}_1) + (-1)^2 Var(\hat{p}_2) \qquad \text{(by independence)} \\ &= \frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2} \end{aligned}$$

If the two samples are reasonably large, then the Central Limit Theorem approximately applies and we can say

$$\hat{p}_1 - \hat{p}_2 \approx N\left(p_1 - p_2, \frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}\right),$$

or equivalently

$$\frac{(\hat{p}_1 - \hat{p}_2) - (p_1 - p_2)}{\sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}} \approx N(0,1)$$

To create a confidence interval, we can substitute for the variance in the denominator our estimates of the unknown parameters p_1 and p_2, which are $\hat{p}_1$ and $\hat{p}_2$. Then a 95% confidence interval for $p_1 - p_2$ is given by

$$(\hat{p}_1 - \hat{p}_2) \pm 1.96\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}$$

Example 1

Let's consider two polls that were taken to gauge the voting support of U.S. President Obama just before the 2012 election. Here are the summary statistics which we will use

Poll #1, August 2012: $n_1 = 1010, \hat{p}_1 = 0.52$

Poll #2, October 2012: $n_2 = 563, \hat{p}_2 = 0.48$

It seems as though his support went down, but does it mean that there was a significant drop in his support or was it just the lack of polls? How much did his *true* support decrease by?

We can compute a 95% confidence interval for the true difference in support using the formula above to obtain

$$\begin{aligned}
&(\hat{p}_1 - \hat{p}_2) \pm 1.96\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}} \\
&= (0.52 - 0.48) \pm 1.96 \ \sqrt{\frac{0.52(1-0.52)}{1,010} + \frac{0.48(1-0.48)}{563}} \\
&= 0.04 \pm 0.052 \\
&\doteq [-0.012, 0.092]
\end{aligned}$$

We cannot really conclude much of interest from this confidence interval since it contains the value 0. We can say that it is possible that his true support increased by 1.2% or it might have decreased by as much as 9.2%. But we are not really sure and the situation is a bit unclear and this leads to the question of, whether we can test for whether these differences like $p_1 - p_2$ really are positive or negative or actually could be 0.

In terms of hypothesis testing, we are interested in whether the proportions of success remain constant from sample to sample or not. For example, we might be interested to know whether a politician's support has changed between one poll and the next. Thus, our null and alternative hypothesis would be

$$H_0 : p_1 = p_2 \text{ vs } H_a : p_1 \neq p_2$$

Recall from the theory discussed above that

$$\frac{(\hat{p}_1 - \hat{p}_2) - (p_1 - p_2)}{\sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}} \approx N(0,1)$$

Now, as with confidence intervals, p_1 and p_2 are unknown population parameters. But in this case, if we are testing hypotheses and we want to compute probabilities under the null hypothesis, we can use the fact that we are assuming that $p_1 = p_2$. Thus, we can approximate both p_1 and p_2 by their **pooled estimate** $\hat{p}$, which we define to be

$$\hat{p} \equiv \frac{n_1\hat{p}_1 + n_2\hat{p}_2}{n_1 + n_2}$$

Using $\hat{p}$ now as our estimate for p_1 and p_2 in the formula above, we obtain

$$\frac{(\hat{p}_1 - \hat{p}_2) - (p_1 - p_2)}{\sqrt{\hat{p}(1-\hat{p})\left(\dfrac{1}{n_1} + \dfrac{1}{n_2}\right)}} \approx N(0,1)$$

and we can use this fact to compute p-values for hypothesis tests.

Example 2

Recall from previous sections the poll results for Toronto's mayor, Rob Ford

Poll #1, June 2011: $n_1 = 1,050, \hat{p}_1 = 0.57$

Poll #2, September 2011: $n_2 = 1,046, \hat{p}_2 = 0.42$

The question we would like to investigate is whether there is a significant drop in his support from the first poll to the second. Thus, we would like to test

$$H_0 : p_1 = p_2 \text{ vs } H_a : p_1 \neq p_2$$

The observed drop in support from our samples is $\hat{p}_1 - \hat{p}_2 = 0.57 - 0.42 = 0.15$ and the pooled estimate for p is $\hat{p} = \dfrac{n_1\hat{p}_1 + n_2\hat{p}_2}{n_1 + n_2} \doteq 0.495$.

The p-value is the probability under the null hypothesis that we would observe a difference in $\hat{p}_1 - \hat{p}_2$ which was at least as big as the observed difference of 0.15. Therefore, we have

$$\begin{aligned} p\text{-value} &= P(|\hat{p}_1 - \hat{p}_2| \geq 0.15) \text{ (under } H_0) \\ &= P\left(\frac{|(\hat{p}_1 - \hat{p}_2) - (p_1 - p_2)|}{\sqrt{\hat{p}(1-\hat{p})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} \geq \frac{0.15}{\sqrt{0.495(1-0.495)\left(\frac{1}{1,050} + \frac{1}{1,046}\right)}}\right) \\ &= P(|N(0,1)| \geq 6.87) \\ &\doteq 6 \times 10^{-12} \end{aligned}$$

Since the p-value is extremely small, we can reject H_0 and conclude that p_1 is not equal to p_2. It is not true that Mr. Ford's support stayed the same between the two polls, in particular his support dropped.

We now have a method to study two different independent samples of proportions. We can compute confidence intervals and hypothesis tests for the difference $p_1 - p_2$, even when the two samples are independent. Next we will consider how to apply the same methods when we are considering means instead of proportions.