Economics 250
Guide to Chapter 8: Six Thing to Know about Inference for Proportions

A proportion measures the incidence of some either/or proprty and so lies between 0 and 1. A proportion may not sound fascinating but it often is reported as a rate, such as the unemployment rate, the crime rate, or the literacy rate. Chapter 8 contains many other examples. I hope you can see that drawing inferences about these is important.

1. The Key Distribution

From chapter 5 remember the binominal distribution is approximated by the normal distribution when the sample size n is large. There x is the number of successes and

$$x \sim N\big(np, \sqrt{np(1-p)}\big).$$

Then the sample success rate or proportion of successes is defined as

$$\hat{p} \equiv \frac{x}{n}.$$

And its distribution (just using the effects of a change of scale we have seen before) is:

$$\hat{p} \sim N\big(p, \sqrt{\frac{p(1-p)}{n}}\big).$$

2. Confidence Intervals

For large samples, a $100(1-\alpha)\%$ confidence interval for the population proportion p is:

$$\hat{p} \pm z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}.$$

Notice that this includes only sample information. Also notice we use z (not t) because we do not have the extra uncertainty from not knowing an added parameter.

e.g. Suppose a survey of 100 voters suggests that 20 intend to vote Green. Find a 95% CI for the population proportion.

Answer: Here $\hat{p} = 0.20$. Also $z_{\alpha/2} = 1.96$. With $n = 100$, using these values gives:

$$0.20 \pm 1.96 \times 0.04 = 0.20 \pm 0.0784 = (0.1216, 0.2784).$$

Often political polls are reported with some measure of uncertainty such as "the results are valid within plus or minus some amount 19 times out of 20." But that it just a 95% CI and its ME, as calculated here.

3. Hypothesis Tests

Suppose we are asked to test a null hypothesis that p takes a specific value, say p_0. We simply form a test statistic like this:

$$z \equiv \frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}}.$$

An important note: Be sure to use p_0—the hypothesized value—in calculating the standard deviation in the numerator of this statistic, and not $\hat{p}$, the sample version. That is because we are finding the distribution while tentatively assuming that the null is true.

Also be sure not to confuse p or $\hat{p}$ with the P-value of a test!

4. The Wilson or Plus 4 Correction

Suppose that we have a small sample like $n = 10$. We survey voters and find none of them intends to vote Green, so $\hat{p} = 0$. But also notice that the standard deviation is

$$\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} = \frac{0 \times 1}{10} = 0,$$

so we also seem to have no uncertainty at all. That seems strange and probably is just an artifact of having such a small sample. So in such cases we sometimes make a small-sample correction called the Wilson or plus 4 rule: Add 4 to the number of trials and 2 to the number of successes. Call the resulting calculation:

$$\tilde{p} = \frac{x+2}{n+4},$$

so in this case we would find $\tilde{p} = (0+2)/(10+4) = 0.1428$. We then could proceed to construct a CI using $\tilde{p}$ instead of $\hat{p}$.

e.g. Suppose $n = 10$ but now $x = 10$ also. Find $\tilde{p}$ and its standard deviation.

As the sample becomes large, the difference between $\tilde{p}$ and $\hat{p}$ will disappear which is why it is sometimes called a small-sample correction.

Finally, don't memorize the textbook's rules on when to use the Wilson rule: We'll simply ask you directly to investigate whether it affects your conclusions.

5. CI for Differences in Proportions

Suppose that we want to estimate the difference between two population proportions, labelled p_1 and p_2. The idea is to see if these rates differ acorss places, times, or groups of people. The corresponding sample difference $\hat{p}_1 - \hat{p}_2$ has standard deviation:

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

which can be used with $z_{\alpha/2}$ to form a confidence interval. Here we're using an idea we have now seen several times: the variance of a difference betwen two independent variables is the sum of their variances. Then we simply take the square root to get the standard deviation.

6. Hypothesis Testing for Differences in Proportions

We may need to compare proportions and we usually do that by testing whether they are equal or, equivalently, whether their difference is zero. For this test, use the pooled estimate of the common value of p_1 and p_2:

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

The idea is that, if the null of equal proportions were true, then this would be the best single estimate of the common value. We then use that twice, once with each sample size, to form:

$$SE_{Dp} = \sqrt{\hat{p}_{pool}(1-\hat{p}_{pool})(\frac{1}{n_1} + \frac{1}{n_2})}.$$

So our test statistic would be:

$$z = \frac{\hat{p}_1 - \hat{p}_2 - 0}{SE_{Dp}}.$$

And of course how we find the P-value depends on the alternative hypothesis as usual.