Class 18: Two Sample Inference for Proportions (Text: Section 8.2)

Comparing Proportions from Two Populations: Notation and Assumptions

If the samples are independent, we can use the standard normal distribution to compare proportions.

Population	Proportion	Sample Size	Sample Count	Sample Proportion
1	p_1	n_1	X_1	$\hat{p}_1 = X_1/n_1$
2	p_2	n_2	X_2	$\hat{p}_2 = X_2/n_2$

We look at the difference in proportions, $D = \hat{p}_1 - \hat{p}_2$

For large samples (bigger than 30), by the Central Limit Theorem, ***D* is distributed**

- Approximately normally
- Mean is $\mu_D = \mu_{\hat{p}_1} - \mu_{\hat{p}_2} = p_1 - p_2$
- Since the samples are independent, the variances add

$$\sigma_D^2 = \sigma_{\hat{p}_1}^2 + \sigma_{\hat{p}_1}^2 = \frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}$$

so the standard deviation of the difference is given by

$$\sigma_D = \sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}$$

In practice, we usually don't know p_1 or p_2, so we use

$$\text{Standard error of the difference} = \text{SE}_\text{D} = \sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}$$

CONFIDENCE INTERVALS

The confidence interval for the difference in population proportions is

$$(\hat{p}_1 - \hat{p}_2 - z\,SE_D, \quad \hat{p}_1 - \hat{p}_2 - z\,SE_D)$$

Margin of error = $z\,SE_D$

HYPOTHESIS TESTS

For a null hypothesis of $p_1 = p_2$, the tests statistic is given by

$$z = \frac{\hat{p}_1 - \hat{p}_2 - (p_1 - p_2)}{SE_D},$$

Where $SE_D = \sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}$ and z has the standard normal distribution.

This assumes that we do not know p_1 or p_2.

Ex: In a 1998 Study at Columbia University called the "Back to School Teen Survey", 1000 teenagers (12-17 years old) were interviewed. Of those 870 did not smoke; 130 did smoke. 68% of the non smokers got good grades (As & Bs) and 41% of the smokers got good grades.

(a) Find a 95% confidence interval for the difference in percentages of students who get good grades in the two groups.
(b) Interpret the confidence interval.
(c) Use the confidence interval to decide if there is there a significant difference between the grades of the smokers and the non-smokers. What significance level are you using?
(d) Can you conclude smoking lowers grades?

(a) The two populations are smoker and non-smokers. The categorical variable we are interesting in is having good grades. Thus the sample statistics that we are given are $\hat{p}_1 = 0.68$, $\hat{p}_2 = 0.41$ and $n_1 = 870, n_2 = 130$. The standard deviation of the difference in sample proportions is given by

$$SE_D = \sqrt{\frac{0.68(1-0.68)}{870} + \frac{0.41(1-0.41)}{130}} = 0.045944.$$

Thus the 95% confidence interval is

$$(0.68 - 0.41 - 1.96(0.04594), 0.68 - 0.41 + 1.96(0.045944)$$
$$= (0.1799, 0.360) = (18\%, 36\%)$$

(b) The confidence interval tells us that the population proportion is very likely (95% chance) to lie between 18% and 36%. More precisely, we should say that there is a 95% chance that intervals generated this way will contain the population proportion.

(c) If there was ***no*** significant difference, 0 would be in the confidence interval, because a possible value for the difference in proportions must be 0. Since 0 is not in the interval, there ***is*** a significant difference. We are 95% confident, so we are working at the 5% significance level.

(d) No. This was an observational study, not an experiment. The study also found smokers drank and did more drugs, so this may have contributed to their lower grades. (These may be confounding variables.)

Distinction between Significance and Causation

Significance: Means unlikely to have occurred by chance if null hypothesis were true; does not imply causation. (Some confounding variable may have caused the effect if the treatment was not randomized.)

Causation: Means that the treatment caused the observed effect. The effect must also have been significant also.

HYPOTHESIS TESTS

In the previous class, we tested whether the proposed malaria drug reduced the number of infections. We compared the infection rate with the drug—11 out of 745, a proportion of 0.0148—with a ***fixed*** infection rate of 0.0349. But in practice, we often do not know the baseline infection rate; that must be determined from a sample—the **control group**, which does not get the drug.

Another Look at Number of Malaria Cases: Two Sample Case

Ex: Of 745 children treated with the malaria drug, 11 got severe malaria. During the same period, 26 of the control group of 745 got sick.[1] Does this data suggest that the drug reduces the rate of severe malaria infections?

How do you expect the results of this test to compare to the result of one-sample test in the previous lecture? Explain. In the two sample test here, the standard error is larger, so the z-value is smaller, so the null hypothesis is less likely to be rejected.

Step 1: Null hypothesis: Proportions sick in the two populations are the same. $H_0: p_1 = p_2$
Alternate hypothesis: Proportion sick in treated population is lower. $H_a: p_1 < p_2$

Step 2: The proportions that got sick are $\hat{p}_1 = 11/745 = 0.0148$ and $\hat{p}_2 = 26/745 = 0.0349$.
For one sample, the standard error is

$$\text{SE} = \sqrt{(0.0349(1-0.0349)/745)} = 0.00672.$$

For two samples, the standard error is

$$\text{SE}_\text{D} = \sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}} = \sqrt{\frac{0.0148(1-0.0148)}{745} + \frac{0.0349(1-0.0349)}{745}} = 0.008.$$

Calculate the z-value:

$$z = \frac{\hat{p}_1 - \hat{p}_2}{\text{SE}_\text{D}} = \frac{0.0148 - 0.0349}{0.008} = -2.5.$$

Step 3: Since the z-value is large, the p-value is small; 0.62%.

Step 4: We reject the null hypothesis: thus a significantly smaller proportion of the treatment group gets sick than of the control group.
Note: The z-value is smaller than before, but result is still significant.

[1] Derived from "Efficacy of the RTS,S/AS202A vaccine against *Plasmodium falciparum* infection and disease in young African children: randomized controlled trial" by P. Alonso et al, *The Lancet*, Oct 16, 2004.

For Hypothesis Tests: Standard Error of Difference in Proportions Using Shared Proportions

The formula we used to approximate the standard error,

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

did not take into account the assumption made by the null hypothesis that the two populations have the same proportion. This ***shared proportion*** or ***pooled proportion,*** $\hat{p}$, should be used in the formula for the standard error.

To estimate the shared proportion, $\hat{p}$, we find the total number of people, $X_1 + X_2$, sharing the characteristic in the two samples. (Note that X_1 and X_2 are the number with the characteristic in the each of the two samples.) Then

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

so putting $\hat{p}_1 = \hat{p}_2 = \hat{p}$ into the formula for the SE and simplifying, we have

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

The formula we used gives results that are close to this new formula if the proportions in the two populations are close to each other.

For a hypothesis test with a pooled proportion

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

Ex: Calculate the standard error of the difference in sample proportions using the shared proportion for the malaria example.

Since $\hat{p}_1 = 11/745 = 0.0148$ and $\hat{p}_2 = 26/745 = 0.0349$ and we previously used

$$\mathrm{SE_D} = \sqrt{\frac{0.0148(1-0.0148)}{745} + \frac{0.0349(1-0.0349)}{745}} = 0.00804.$$

The shared proportion is

$$\hat{p} = \frac{11 + 26}{745 + 745} = 0.0248,$$

so, although the change is small, it would have been better to use

$$\mathrm{SE_D} = \sqrt{0.00248(1-0.00248)\left(\frac{1}{745} + \frac{1}{745}\right)} = 0.00806.$$

Ex: Calculate the z-score for the difference in sample proportions using the shared proportion for the malaria example.

With the pooled proportion:

$$z = \frac{\hat{p}_1 - \hat{p}_2}{\mathrm{SE_D}} = \frac{0.0148 - 0.0349}{0.00806} = -2.4937$$

Previously, we got

$$z = \frac{\hat{p}_1 - \hat{p}_2}{\mathrm{SE_D}} = \frac{0.0148 - 0.0349}{0.00804} = -2.5,$$

so the difference is small.

Testing HIV-AIDs Vaccines

In 2007, the drug company Merck tested a HIV-AIDs vaccine, the first of a new class of drugs. High risk volunteers in the US and Latin America were randomly assigned the drug or a placebo. Both groups were given safe sex counseling. After 13 months, 24 of the 741 people who received the vaccine were infected, compared to 21 of the 762 people who received the placebo.[2]

Ex: What conclusion can you draw from this data? Did the vaccine have a significant effect?
Since the proportion infected in the treated group, $\hat{p}_1 = 24/741 = 0.032$, is ***greater*** than the proportion in the untreated group, $\hat{p}_2 = 21/762 = 0.028$, the experiment suggests the vaccine does ***not*** work.
We could test the hypothesis that significantly *more* people get infected with the vaccine.
Step 1: H_0: Equal proportions infected in each group: $p_1 = p_2$.
H_a: More infected in treated group: $p_1 > p_2$.
Step 2: Standard error is

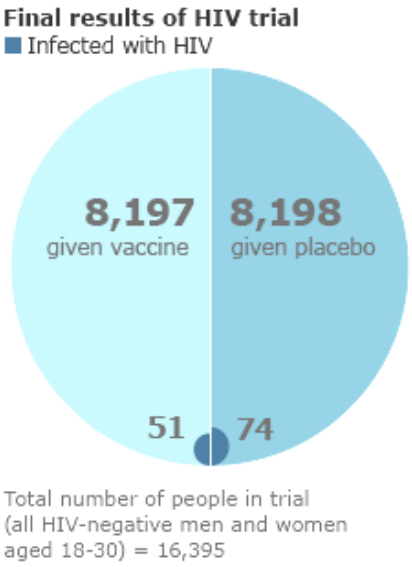
$$SE = \sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}} = \sqrt{\frac{0.032(1-0.032)}{741} + \frac{0.028(1-0.028)}{762}} = 0.0088.$$

Then

$$z = \frac{\hat{p}_1 - \hat{p}_2}{\text{SE}} = \frac{0.032 - 0.028}{0.0088} = 0.5.$$

Step 3: From the table, the percentile is 69.15, so the *p*-value is $p = 0.3085 = 30.85\%$
Step 4: Since the p-value is not small, we cannot reject the null hypothesis. The vaccine does not appear to have any effect.

In September 2009, trials on another potential HIV vaccine, RV-144, given to volunteers in Thailand, reported the results in the diagram.[3] As before, all volunteers were given safe sex counseling.



Ex: Does this vaccine have a significant effect on infection rates?
Population 1 is the treated group; population 2 is the control group.
Then $\hat{p}_1 = 51/8197 = 0.0062$ and $\hat{p}_2 = 74/8198 = 0.009$
Step 1: H_0: Equal proportions infected in each group: $p_1 = p_2$.
H_a: Smaller proportion infected in treated group: $p_1 < p_2$.
Step 2: Standard error is

$$\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}} = \sqrt{\frac{0.0062(1-0.062)}{8197} + \frac{0.009(1-0.009)}{8198}} = 0.00136.$$

Then

$$z = \frac{\hat{p}_1 - \hat{p}_2}{\text{SE}} = \frac{0.0062 - 0.009}{0.00136} = -2.1.$$

Step 3: From the table, the *p*-value is $p = 1.79\%$
Step 4: Since the *p*-value is small, we reject the null hypothesis. The vaccine appears to have an effect.

[2] "Failure of Vaccine Test is Setback in AIDS Fight", by L. Altman, A. Pollack, *New York Times,* Sept 22, 2007.
[3] http://www.nytimes.com/2009/09/25/health/research/25aids.html?_r=1&scp=1&sq=HIV%20vaccine&st=cse and http://news.bbc.co.uk/2/hi/health/8272113.stm