

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 $\sigma_D^2 = \sigma_{\hat{p}_1}^2 + \sigma_{\hat{p}_2}^2 = \frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}$
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, of course we usually don't know p_1 or p_2 , so we use

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

CONFIDENCE INTERVALS

The confidence interval for the difference in population proportions is

Margin of error =

HYPOTHESIS TESTS

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

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?

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.

Step 1:

Step 2:

Step 3:

Step 4:

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,

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

did not take into account the assumption made by the null hypothesis that the two populations have the same proportion. This ***shared proportion*** 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 which 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.

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

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

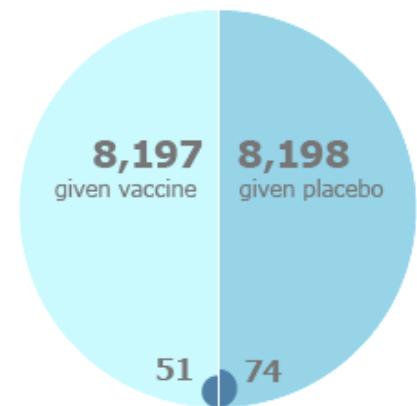
Ex: What conclusion can you draw from this data? Did the vaccine have a significant effect?

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?

Final results of HIV trial

■ Infected with HIV



Total number of people in trial
(all HIV-negative men and women
aged 18-30) = 16,395

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>