Problems with Two Proportions

Alan B. Gelder
06E:071, The University of Iowa[1]

1 Theoretical Introduction

1.1 Going from One to Two Proportions

Two-proportion problems are the natural extension of one-proportion problems. The major difference is that in one-proportion problems, we are comparing a proportion with a number between 0 and 1; in two-proportion problems, we are comparing two proportions with each other. Typically, we want to know if the proportion from one population differs (or is greater or less than) the proportion from another population. For instance, is the proportion of men who graduate with a four-year college degree greater than the corresponding proportion of women? Is the proportion of Chinese workers who are engineers different from the proportion of Americans workers who are engineers?

A New Distribution: Each of the two proportions that we are comparing comes from a distinct population with its own distribution. The data that we have are samples from these two different populations. In order to appropriately compare the two proportions, we need to create a new distribution which combines these two populations. We do this by subtracting one distribution from the other. **(Draw two distributions of proportions and the resultant third distribution formed by subtracting one from the other.)**

The magic number 0: For hypothesis testing and evaluating confidence intervals, zero becomes the magic number in our newly constructed distribution. Why is that? Suppose that the parameters for the two initial distributions are equal (i.e. $p_1 = p_2$). Then in our new distribution we have that $p_1 - p_2 = 0$, or rather, zero is the center of the new distribution. In hypothesis testing, zero is always the number that we are testing against. This can be seen as follows:

$H_A : p_1 > p_2 \Rightarrow p_1 - p_2 > 0$
$H_0 : p_1 \leq p_2 \Rightarrow p_1 - p_2 \leq 0$

$H_A : p_1 \neq p_2 \Rightarrow p_1 - p_2 \neq 0$

[1]The notes for this class follow the general format of Blake Whitten's lecture notes for the same course.

$H_0 : p_1 = p_2 \Rightarrow p_1 - p_2 = 0$

$H_A : p_1 < p_2 \Rightarrow p_1 - p_2 < 0$
$H_0 : p_1 \geq p_2 \Rightarrow p_1 - p_2 \geq 0$

1.2 The Standard Deviation for the New Distribution

The standard deviation in a one-proportion problem is

$$\sqrt{\frac{p(1-p)}{n}}$$

In practice, we need to substitute the parameter p with the statistic $\hat{p}$.

Recall that if two populations are independent, then their variances can simply be added together (if the two populations are not independent, then the covariance also needs to be accounted for).

$$\sigma^2_{p_1-p_2} = \frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_1}$$

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

We will approximate the standard deviation $\sigma_{\hat{p_1}-\hat{p_2}}$ as follows:

$$\sigma_{\hat{p_1}-\hat{p_2}} \approx \sqrt{\frac{\hat{p_1}(1-\hat{p_1})}{n_1} + \frac{\hat{p_2}(1-\hat{p_2})}{n_2}}$$

Confidence Interval Formula: This leads to the formula for two-proportion confidence intervals:

$$(\hat{p_1} - \hat{p_2}) \pm z^* \sigma_{\hat{p_1}-\hat{p_2}}$$

$$\approx (\hat{p_1} - \hat{p_2}) \pm z^* \sqrt{\frac{\hat{p_1}(1-\hat{p_1})}{n_1} + \frac{\hat{p_2}(1-\hat{p_2})}{n_2}}$$

1.3 Deriving the Test Statistic

Implications of the Null Hypothesis: In the two-proportion setting, the null hypothesis H_0: $p_1 = p_2$ implies that the two populations have the same underlying proportion. Following the format of hypothesis testing, we assume that H_0 is true until there is sufficient evidence to indicate otherwise. If H_0 is indeed true, then we essentially just have one population with one underlying proportion (instead of two proportions for two separate populations). Mathematically, this implies that

$$p = p_1 = p_2$$

Pooling Data: Large samples give us more information about a population than small samples. If our data is indeed all coming from one population instead of two, then we can pool our data together to make one large sample. We will do this for the hypothesis test, because, after all, we are assuming that we only have one proportion for one population. This can be seen as follows:

$$\hat{p} = \frac{x_1 + x_2}{n_1 + n_2}$$

Note that $\hat{p}_1 = \frac{x_1}{n_1}$ and $\hat{p}_2 = \frac{x_2}{n_2}$.

The Standard Deviation: As was shown previously, the standard deviation in a two proportion setting is

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

However, under the null hypothesis we assume that $p = p_1 = p_2$. This implies that the standard deviation can be written as follows:

$$\sigma_{p_1 - p_2} = \sqrt{\left(\frac{1}{n_1} + \frac{1}{n_2}\right) p(1-p)}$$

Since we do not know p, we will approximate it with $\hat{p}$, which is formed by pooling our data from the two samples.

Two-Proportion Test Statistic: Finally, we can state the test-statistic:

$$z = \frac{\hat{p_1} - \hat{p_2}}{\sqrt{\left(\frac{1}{n_1} + \frac{1}{n_2}\right) \hat{p}(1-\hat{p})}}$$

2 Example Problems

2.1 Scholarships

A college committee is concerned that the proportion of male and female students who are awarded scholarships is not the same. In order to investigate this issue, the committee conducted a survey of 300 male and 300 female students. The survey revealed that 92 of the male students and 104 of the female students had received a scholarship. Should the committee be concerned? Use $\alpha = 0.05$.

Four-Step Hypothesis Test:

1. p_1 = The proportion of male students who have received a scholarship
p_2 = The proportion of female students who have received a scholarship
$H_A : p_1 \neq p_2$
$H_0 : p_1 = p_2$

2. $\hat{p_1} = 92/300 = 0.3067$
$\hat{p_2} = 104/300 = 0.3467$

$$\hat{p} = \frac{92 + 104}{600} = \frac{196}{600} = 0.3267$$

$$z = \frac{\hat{p_1} - \hat{p_2}}{\sqrt{\left(\frac{1}{n_1} + \frac{1}{n_2}\right)\hat{p}(1-\hat{p})}} = \frac{0.3067 - 0.3467}{\sqrt{\left(\frac{1}{300} + \frac{1}{300}\right)(0.3267)(1-0.3267)}}$$

$$= \frac{-0.0400}{0.0383} = -1.0444$$

$$p - value = 2(0.1492) = 0.2984$$

3. Fail to Reject H_0 since p-value $= 0.2984 > 0.05 = \alpha$.

4. There is not enough evidence to show that the proportion of male and female students who receive scholarships differs.

Confidence Interval: Find and interpret a 95% confidence interval for the difference in the proportion of male and female students who receive scholarships.

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

$$= \sqrt{\frac{0.3067(1-0.3067)}{300} + \frac{0.3467(1-0.3467)}{300}} = 0.03826$$

$$(\hat{p_1} - \hat{p_2}) \pm z^* \sigma_{\hat{p_1}-\hat{p_2}} = (0.3067 - 0.3467) \pm 1.96(0.03826)$$
$$= -0.040 \pm 0.07499 = (-0.11499, 0.03499)$$

Interpretation of CI: With 95% confidence, the proportion of male students who receive scholarships is between 11.5% less than and 3.5% more than the proportion of female students who receive scholarships.

What can we infer since 0 is within the confidence interval?

2.2 Aspirin Use in Stroke Patients

[This example is taken from p. 55 of Dr. Whitten's Spring 2011 lecture notes.]

A clinical study examined the effectiveness of aspirin in the treatment of stroke. Stroke patients were randomly divided into two groups: a *treatment* group whose patients received a daily aspirin and a *control* group whose patients received a placebo (a sugar pill which looks and tastes like an aspirin). To prevent psychological factors from affecting the study, none of the patients knew if they were taking an aspirin or a placebo each day.

Of the 155 patients in the study, 78 patients were placed in the treatment group. Records show that 43 patients in the control group had favorable outcomes while only 15 patients in the treatment group had unfavorable outcomes.

Based on this study, is the use of aspirin recommended for stroke patients? Use $\alpha = 0.01$.

Four-Step Hypothesis Test:

1. p_1 = The proportion of aspirin users whose condition improves
p_2 = The proportion of placebo users whose condition improves

$H_A : p_1 > p_2$ (or $p_1 - p_2 > 0$)
$H_0 : p_1 \leq p_2$ (or $p_1 - p_2 \leq 0$)

2. $\hat{p_1} = (78 - 15)/78 = 63/78 = 0.8077$
$\hat{p_2} = 43/(155 - 78) = 43/77 = 0.5584$

$$\hat{p} = \frac{63 + 43}{78 + 77} = \frac{106}{155} = 0.6839$$

$$z = \frac{0.8077 - 0.5584}{\sqrt{\left(\frac{1}{78} + \frac{1}{77}\right)(0.6839)(1 - 0.6839)}} = 3.34$$

$$p - value = 0.0004$$

3. Reject H_0 since p-value $= 0.0004 < 0.01 = \alpha$.

4. There is sufficient evidence that stroke patients who use aspirin have a higher recovery rate that stroke patients who do not use aspirin.

Confidence Interval: Quantify and interpret the benefit to stroke patients from using aspirin, using 95% confidence.

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

$$= \sqrt{\frac{.8077(1 - .8077)}{78} + \frac{.5584}{1 - .5584}} \approx \sqrt{0.0051938} = 0.0721$$

$$(\hat{p_1} - \hat{p_2}) \pm z^* \sigma_{\hat{p_1} - \hat{p_2}}$$

$$= (.8077 - .5584) \pm (1.96)(0.0721) = .2493 \pm .1413 = (.1080, .3906)$$

Interpretation: With 95% confidence, stroke patients who use aspirin are between 10.80% and 39.06% more likely to have a favorable outcome than stroke patients who do not use aspirin.

Misreported Data: Suppose now that 63 patients in the placebo group had favorable outcomes (not 43). Based on the revised data, is the use of aspirin recommended for stroke patients? Use a four-step hypothesis test with $\alpha = 0.05$. Also, quantify and interpret the benefit to stroke patients from using aspirin with 95% confidence.

Quick Minitab note: When doing two-proportion problems in Minitab, *always* click the box labeled "Use pooled estimate of p for test." See the Minitab Commands document for further Minitab instructions on two-proportion problems.

3 Formulas

Two-Proportion Hypothesis Test:

$$z = \frac{\hat{p_1} - \hat{p_2}}{\sqrt{\left(\frac{1}{n_1} + \frac{1}{n_2}\right)\hat{p}(1-\hat{p})}}$$

Two-Proportion Confidence Interval:

$$(\hat{p_1} - \hat{p_2}) \pm z^* \sqrt{\frac{\hat{p_1}(1-\hat{p_1})}{n_1} + \frac{\hat{p_2}(1-\hat{p_2})}{n_2}}$$