

7.2 - Comparing Two Population Proportions

7.2 - Comparing Two Population Proportions

Introduction

When we have a categorical variable of interest measured in two populations, it is quite often that we are interested in comparing the proportions of a certain category for the two populations.

Let's consider the following example.

Application

\$100 Received by Mistake



Males and females were asked about what they would do if they received a \$100 bill by mail, addressed to their neighbor, but wrongly delivered to them. Would they return it to their neighbor? Of the 69 males sampled, 52 said "yes" and of the 131 females sampled, 120 said "yes."

Does the data indicate that the proportions that said "yes" are different for male and female? How do we begin to answer this question?

If the proportion of males who said "yes, they would return it" is denoted as p_1 and the proportion of females who said "yes, they would return it" is denoted as p_2 , then the following equations indicate that p_1 is equal to p_2 .

$$p_1 - p_2 = 0 \text{ or } \frac{p_1}{p_2} = 1$$

We would need to develop a confidence interval or perform a hypothesis test for one of these expressions.

Moving forward

There may be other ways of setting up these equations such that the proportions are equal. We choose the difference due to the theory discussed in the last section. Under certain conditions, the sampling distribution of \hat{p}_1 , for example, is approximately normal and centered around p_1 . Similarly, the sampling distribution of \hat{p}_2 is approximately normal and centered around p_2 . Their difference, $\hat{p}_1 - \hat{p}_2$, will then be approximately normal and centered around $p_1 - p_2$, which we can use to determine if there is a difference.

In the next subsections, we explain how to use this idea to develop a confidence interval and hypothesis tests for $p_1 - p_2$.

7.2.1 - Confidence Intervals

7.2.1 - Confidence Intervals

In this section, we begin by defining the point estimate and developing the confidence interval based on what we have learned so far.

Point Estimate

The point estimate for the difference between the two population proportions, $p_1 - p_2$, is the difference between the two sample proportions written as $\hat{p}_1 - \hat{p}_2$.

We know that a point estimate is probably not a good estimator of the actual population. By adding some amount of error to this point estimate, we can create a confidence interval as we did with one sample parameters.

Derivation of the Confidence Interval

Consider two populations and label them as population 1 and population 2. Take a random sample of size n_1 from population 1 and take a random sample of size n_2 from population 2. If we consider them separately,

Proportion from Sample 1:

If $n_1 p_1 \geq 5$ and $n_1(1 - p_1) \geq 5$, then \hat{p}_1 will follow a normal distribution with...

Mean:	p_1
Standard Error:	$\sqrt{\frac{p_1(1 - p_1)}{n_1}}$
Estimated Standard Error:	$\sqrt{\frac{\hat{p}_1(1 - \hat{p}_1)}{n_1}}$

Proportion from Sample 2:

If $n_2 p_2 \geq 5$ and $n_2(1 - p_2) \geq 5$, then \hat{p}_2 will follow a normal distribution with...

$$\begin{aligned}
 \text{Mean:} & \quad p_2 \\
 \text{Standard Error:} & \quad \sqrt{\frac{p_2(1-p_2)}{n_2}} \\
 \text{Estimated Standard Error:} & \quad \sqrt{\frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}
 \end{aligned}$$

Sample Proportion 1 - Sample Proportion 2:

Using the theory introduced previously, if n_1p_1 , $n_1(1-p_1)$, n_2p_2 , and $n_2(1-p_2)$ are all greater than five and we have independent samples, then the sampling distribution of $\hat{p}_1 - \hat{p}_2$ is approximately normal with...

$$\begin{aligned}
 \text{Mean:} & \quad p_1 - p_2 \\
 \text{Standard Error:} & \quad \sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}} \\
 \text{Estimated Standard Error:} & \quad \sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}
 \end{aligned}$$

Putting these pieces together, we can construct the confidence interval for $p_1 - p_2$. Since we do not know p_1 and p_2 , we need to check the conditions using $n_1\hat{p}_1$, $n_1(1-\hat{p}_1)$, $n_2\hat{p}_2$, and $n_2(1-\hat{p}_2)$. If these conditions are satisfied, then the confidence interval can be constructed for two independent proportions.

Confidence interval for two independent proportions

The $(1-\alpha)100\%$ confidence interval of $p_1 - p_2$ is given by:

$$\hat{p}_1 - \hat{p}_2 \pm z_{\alpha/2} \sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}$$

Example 7-1: \$100 Received by Mistake



Males and females were asked about what they would do if they received a \$100 bill by mail, addressed to their neighbor, but wrongly delivered to them. Would they return it to their neighbor? Of the 69 males sampled, 52 said "yes" and of the 131 females sampled, 120 said "yes."

Find a 95% confidence interval for the difference in proportions for males and females who said "yes."

Answer

Let's let sample one be males and sample two be females. Then we have:

Males:

$$n_1 = 69, \hat{p}_1 = \frac{52}{69}$$

Females:

$$n_2 = 131, \hat{p}_2 = \frac{120}{131}$$

Checking conditions we see that $n_1\hat{p}_1$, $n_1(1 - \hat{p}_1)$, $n_2\hat{p}_2$, and $n_2(1 - \hat{p}_2)$ are all greater than five so our conditions are satisfied.

Using the formula above, we get:

$$\begin{aligned} \hat{p}_1 - \hat{p}_2 &\pm z_{\alpha/2} \sqrt{\frac{\hat{p}_1(1 - \hat{p}_1)}{n_1} + \frac{\hat{p}_2(1 - \hat{p}_2)}{n_2}} \\ \frac{52}{69} - \frac{120}{131} &\pm 1.96 \sqrt{\frac{\frac{52}{69}(1 - \frac{52}{69})}{69} + \frac{\frac{120}{131}(1 - \frac{120}{131})}{131}} \\ -0.1624 &\pm 1.96(0.05725) \\ -0.1624 &\pm 0.1122 \text{ or } (-0.2746, -0.0502) \end{aligned}$$

We are 95% confident that the difference of population proportions of males who said "yes" and females who said "yes" is between -0.2746 and -0.0502.

Based on both ends of the interval being negative, it seems like the proportion of females who would return it is higher than the proportion of males who would return it.

We will discuss how to find the confidence interval using Minitab after we examine the hypothesis test for two proportion. Minitab calculates the test and the confidence interval at the same time.

Caution! What happens if we defined \hat{p}_1 to be the proportion of females and \hat{p}_2 for the proportion of males? If you follow through the calculations, you will find that the confidence interval will differ only in sign. In other words, if female was \hat{p}_1 , the interval would be 0.0502 to 0.2746. It still shows that the proportion of females is higher than the proportion of males.

7.2.2 - Hypothesis Testing

7.2.2 - Hypothesis Testing

Derivation of the Test

We are now going to develop the hypothesis test for the difference of two proportions for independent samples. The hypothesis test will follow the same six steps we learned in the previous Lesson although they are not explicitly stated.

We will use the sampling distribution of $\hat{p}_1 - \hat{p}_2$ as we did for the confidence interval. One major difference in the hypothesis test is the null hypothesis and assuming the null hypothesis is true.

For a test for two proportions, we are interested in the difference. If the difference is zero, then they are not different (i.e., they are equal). Therefore, the null hypothesis will always be:

$$H_0: p_1 - p_2 = 0$$

Another way to look at it is $H_0: p_1 = p_2$. This is worth stopping to think about. Remember, in hypothesis testing, we assume the null hypothesis is true. In this case, it means that p_1 and p_2 are equal. Under this assumption, then \hat{p}_1 and \hat{p}_2 are both estimating the same proportion. Think of this proportion as p^* . Therefore, the sampling distribution of both proportions, \hat{p}_1 and \hat{p}_2 , will, under

certain conditions, be approximately normal centered around p^* , with standard error $\sqrt{\frac{p^*(1-p^*)}{n_i}}$, for $i = 1, 2$.

We take this into account by finding an estimate for this p^* using the two sample proportions. We can calculate an estimate of p^* using the following formula:

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

This value is the total number in the desired categories ($x_1 + x_2$) from both samples over the total number of sampling units in the combined sample ($n_1 + n_2$).

Putting everything together, if we assume $p_1 = p_2$, then the sampling distribution of $\hat{p}_1 - \hat{p}_2$ will be approximately normal with mean 0 and standard error of $\sqrt{p^*(1-p^*)\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}$, under certain conditions.

Therefore,

$$z^* = \frac{(\hat{p}_1 - \hat{p}_2) - 0}{\sqrt{\hat{p}^*(1 - \hat{p}^*)\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

...will follow a standard normal distribution.

Finally, we can develop our hypothesis test for $p_1 - p_2$.

Null: $H_0: p_1 - p_2 = 0$

Possible Alternatives:

$H_a: p_1 - p_2 \neq 0$

$H_a: p_1 - p_2 > 0$

$H_a: p_1 - p_2 < 0$

Conditions:

$n_1\hat{p}_1, n_1(1 - \hat{p}_1), n_2\hat{p}_2$, and $n_2(1 - \hat{p}_2)$ are all greater than five

The test statistic is:

$$z^* = \frac{\hat{p}_1 - \hat{p}_2 - 0}{\sqrt{\hat{p}^*(1 - \hat{p}^*)\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

...where $\hat{p}^* = \frac{x_1 + x_2}{n_1 + n_2}$.

The critical values, rejection regions, p-values, and decisions will all follow the same steps as those from a hypothesis test for a one sample proportion.

Example 7-2: \$100 Received by Mistake



Let's continue with the question that was asked previously.

Males and females were asked about what they would do if they received a \$100 bill by mail, addressed to their neighbor, but wrongly delivered to them. Would they return it to their neighbor? Of the 69 males sampled, 52 said "yes" and of the 131 females sampled, 120 said "yes."

Does the data indicate that the proportions that said "yes" are different for males and females at a 5% level of significance? Conduct the test using the p-value approach.

Answer

- [By Hand](#) ^[1]
- [Using Minitab](#) ^[2]

1. Again, let's define males as sample 1.

The conditions are all satisfied as we have shown previously.

The null and alternative hypotheses are:

$$H_0: p_1 - p_2 = 0 \text{ vs } H_a: p_1 - p_2 \neq 0$$

The test statistic:

$$n_1 = 69, \hat{p}_1 = \frac{52}{69}$$

$$n_2 = 131, \hat{p}_2 = \frac{120}{131}$$

$$\hat{p}^* = \frac{x_1 + x_2}{n_1 + n_2} = \frac{52 + 120}{69 + 131} = \frac{172}{200} = 0.86$$

$$z^* = \frac{\hat{p}_1 - \hat{p}_2 - 0}{\sqrt{\hat{p}^*(1 - \hat{p}^*)\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} = \frac{\frac{52}{69} - \frac{120}{131}}{\sqrt{0.86(1 - 0.86)\left(\frac{1}{69} + \frac{1}{131}\right)}} = -3.1466$$

The p-value of the test based on the two-sided alternative is:

$$\text{p-value} = 2P(Z > |-3.1466|) = 2P(Z > 3.1466) = 2(0.0008) = 0.0016$$

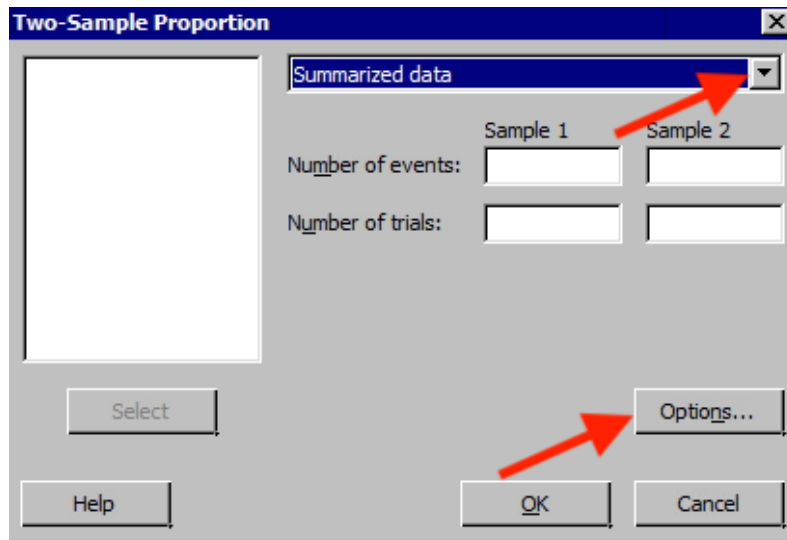
Since our p-value of 0.0016 is less than our significance level of 5%, we reject the null hypothesis. There is enough evidence to suggest that proportions of males and females who would return the money are different.

2. **Minitab: Inference for Two Proportions with Independent Samples**

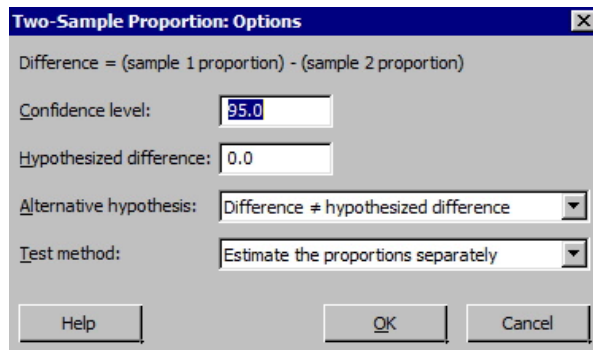
To conduct inference for two proportions with an independent sample in Minitab...

1. Choose Stat > Basic Statistics > 2 proportions

The following window will appear. In the drop-down choose 'Summarized data' and entered the number of events and trials for both samples.



2. Choose Options to display this window.



Notice how the difference is calculated. We also want to make sure that we are using the pooled estimate of the proportion as the test method. In Minitab, you need to get into options and select "Use pooled estimate of p for test." If you don't think that is reasonable to assume, then don't check the option.

You should get the following output for this example:

Test and CI for Two Proportions

Sample	X	N	Sample p
1	52	69	0.753623
2	120	131	0.916031

Difference = $p(1) - p(2)$

Estimate for difference: -0.162407

95% CI for difference: (-0.274625, -0.0501900)

Test for difference = 0 (vs $\neq 0$): Z = -3.15 P-Value = 0.002 **(Use this!)**

Fisher's exact test: P-Value = 0.003 **(Ignore the Fisher's exact test. This test uses a different method to calculate a test statistic from the Z-test we have learned in this lesson.)**

Ignore the Fisher's p -value! The p -value highlighted above is calculated using the methods we learned in this lesson. The Fisher's test uses a different method than what we explained in this lesson to calculate a test statistic and p -value. This method incorporates a log of the ratio of observed to expected values. It's just a different technique that is more complicated to do by hand. Minitab automatically includes both results in its output.

Note! Notice that Minitab also produces the 95% confidence interval for us. You should compare this confidence interval to the one found in the previous section.

Try it!

In 1980, of 750 men 20-34 years old, 130 were found to be overweight. Whereas, in 1990, of 700 men, 20-34 years old, 160 were found to be overweight.

At the 5% significance level, do the data provide sufficient evidence to conclude that, for men 20-34 years old, a higher percentage were overweight in 1990 than 10 years earlier? Conduct the test using the p -value approach.

- [By Hand](#) ^[3]
- [Using Minitab](#) ^[4]

1. Let's define 1990 as sample 1.

The null and alternative hypotheses are:

$$H_0: p_1 - p_2 = 0 \text{ vs } H_a: p_1 - p_2 > 0$$

$$n_1 = 700, \hat{p}_1 = \frac{160}{700}$$

$$n_2 = 750, \hat{p}_2 = \frac{130}{750}$$

$$\hat{p}^* = \frac{x_1 + x_2}{n_1 + n_2} = \frac{160 + 130}{700 + 750} = \frac{290}{1450} = 0.2$$

The conditions are all satisfied: $n_1\hat{p}_1$, $n_1(1 - \hat{p}_1)$, $n_2\hat{p}_2$, and $n_2(1 - \hat{p}_2)$ are all greater than 5.

The test statistic:

$$z^* = \frac{\hat{p}_1 - \hat{p}_2 - 0}{\sqrt{\hat{p}^*(1 - \hat{p}^*)\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} = \frac{\frac{160}{700} - \frac{130}{750}}{\sqrt{0.2(1 - 0.2)\left(\frac{1}{700} + \frac{1}{750}\right)}} = 2.6277$$

The p -value of the test based on the right-tailed alternative is:

$$p\text{-value} = P(Z > 2.6277) = 0.0043$$

Since our p-value of 0.0043 is less than our significance level of 5%, we reject the null hypothesis. There is enough evidence to suggest that the proportion of males overweight in 1990 is greater than the proportion in 1980.

2. Using Minitab

To conduct inference for two proportions with independent samples in Minitab...

1. Choose Stat > Basic Statistics > 2 proportions
2. Choose Options
3. Select "Difference < hypothesized difference" for 'Alternative Hypothesis.

You should get the following output.

Test and CI for Two Proportions

Sample	X	N	Sample p
1	130	750	0.173333
2	160	700	0.228571

Difference = $p(1) - p(2)$

Estimate for difference: -0.0552381

95% upper bound for difference: -0.0206200

Test for difference = 0 (vs < 0): Z = -2.63 P-Value = 0.004

Fisher's exact test: P-Value = 0.005 **(Ignore the Fisher's exact test)**