Section 4C – Proportion Relationships: Two-population Proportion Test

Sometimes we wish to determine if a specific percentage from categorical data is related to various groups (populations). If we only have two populations, we can use a two-population proportion hypothesis test with a Z-score test statistic. If we have three or more populations, we will need to use a more advanced test statistic called the chi-squared test statistic. This is sometimes called a "Goodness of Fit" test. The key idea is to ask the question if the population percentage is the same in the various groups or is it significantly different.

Two-Population Proportion Test for Proportion Relationships

There are different ways of writing the null and alternative hypothesis. A population proportion can be described with the Greek letter pi (π) or with a "p". Remember equal proportions goes with the null hypothesis of "not related" while any difference between the proportions indicates a relationship.

H_0: $p_1 = p_2$ The population % is NOT related to a categorical variable (% is not related to the groups)
H_A: $p_1 \neq p_2$ The population % is related to a categorical variable (% is related to the groups)

As we learned in the last chapter, the alternative hypothesis determines the type of test. If the alternative hypothesis is greater than (>) it is a right-tailed test. If the alternative hypothesis is less than (<) it is a left-tailed test. If the alternative hypothesis is not equal $(\neq)$ it is a two-tailed test. While some prefer to use ≥ or ≤ for the null hypothesis, I prefer not to because of relationship implications.

Two-Tailed Null and Alternative Hypothesis

H_0: $p_1 = p_2$ The population % is NOT related to a categorical variable (% is not related to the groups)
H_A: $p_1 \neq p_2$ The population % is related to a categorical variable (% is related to the groups)

OR

H_0: $\pi_1 = \pi_2$ The population % is NOT related to a categorical variable (% is not related to the groups)
H_A: $\pi_1 \neq \pi_2$ The population % is related to a categorical variable (% is related to the groups)

Right-Tailed Null and Alternative Hypothesis

H_0: $p_1 = p_2$ The population % is NOT related to a categorical variable (% is not related to the groups)
H_A: $p_1 > p_2$ The population % is related to a categorical variable (% is related to the groups)

OR

H_0: $\pi_1 = \pi_2$ The population % is NOT related to a categorical variable (% is not related to the groups)
H_A: $\pi_1 > \pi_2$ The population % is related to a categorical variable (% is related to the groups)

Left-Tailed Null and Alternative Hypothesis

H_0: $p_1 = p_2$ The population % is NOT related to a categorical variable (% is not related to the groups)
H_A: $p_1 < p_2$ The population % is related to a categorical variable (% is related to the groups)

OR

H_0: $\pi_1 = \pi_2$ The population % is NOT related to a categorical variable (% is not related to the groups)
H_A: $\pi_1 < \pi_2$ The population % is related to a categorical variable (% is related to the groups)

Assumptions

It is very important to always check the assumptions for a hypothesis test in order to make sure that our sample data is as unbiased as possible. Remember that biased data may lead to a wrong conclusion (type 1 or type 2 error). Since we are now using this test to determine relationships, we may also need to prove cause and effect. If that is the case, we will need to use random assignment instead of a random sample.

This chapter is from Introduction to Statistics for Community College Students, 1st Edition, by Matt Teachout, College of the Canyons, Santa Clarita, CA, USA, and is licensed under a "CC-By" Creative Commons Attribution 4.0 International license – 10/1/18

Assumptions for a Two-population Proportion Test for Relationship

1. Random: The sample categorical data either should be a random sample *(if proving there is relationship)* or have used random assignment *(if proving cause and effect)*.

2. Large sample size: The sample categorical data should have at least ten success $(x \geq 10)$ and at least ten failures $(n - x \geq 10)$. *For example, there should be at least 10 people with congestive heart failure (CHF) in the sample from the U.S. and at least 10 people without CHF in the sample from the U.S. There should also be at least 10 people with CHF in the sample from the Australia and at least 10 people without CHF in the sample from Australia.*

3. Data values within each sample and between the samples should be independent of each other. If the data was collected from one sample then the assumption is just that data values within the sample should be independent. If the data was collected from more than one sample, then the data values between the samples should also be checked for independence. *For example, we should not have people in our samples that are family members or the same people measured twice. The sample from the U.S. should not be connected to the sample from Australia. For example, the congestive heart failure (CHF) data should not come from a company that has hospitals in both countries. In an experiment, we should not control confounding variables by using the same group of people measured multiple times. This would fail the independent individuals' assumption. Random assignment is a better option for controlling confounding variables.*

Note: Some statisticians like to use the chi-squared test statistic even if there are only two populations of interest. If that is the case, use the assumptions for the goodness of fit test.

Z-test statistic for two-population proportion tests

The Z-test statistic measures the number of standard errors that the sample proportion from group 1 $(\hat{p}_1)$ is above or below the sample proportion from group 2 $(\hat{p}_2)$. It is "above" when the Z-test statistic is positive and "below" when the Z-test statistic is negative. It can also be thought of as the number of standard errors that the difference between the sample proportions $(\hat{p}_1 - \hat{p}_2)$ is from zero or some other claimed difference. Z-scores usually are significant around two standard errors, but it is always good to refer to the critical value or P-value when judging significance. The formula below seems daunting to calculate. Remember, no one in data science calculates this by hand with a calculator. Always use a computer program like R or Statcato. In the two-population proportion Z test, we often use pooling ($\overline{p}$). Pooling the proportions is combines the two data sets together before calculating the proportion. We need to assume that the population proportions are equal in the null hypothesis in order to pool. For this reason, pooling is usually used for a two-population proportion hypothesis test and is not used in a two-population proportion confidence interval.

p-pooled ($\overline{p}$) = $\frac{(x_1+x_2)}{(n_1+n_2)}$

Z-test statistic for two population proportion *(pooled)* = $\frac{(sample\ 1 - sample\ 2)}{standard\ error} = \frac{(\hat{p}_1 - \hat{p}_2)}{\sqrt{\left(\frac{\overline{p}(1-\overline{p})}{n_1} + \frac{\overline{p}(1-\overline{p})}{n_2}\right)}}$

While this formula looks daunting, it is only counting how many standard errors the sample proportion for group 1 is above or below group 2. The most important thing is not calculating. It is interpreting and explaining the test statistic.

Z-test statistic for two-population sentence: The sample proportion for group 1 is # of standard errors (above or below) the sample proportion for group 2.

Look at the following two-population proportion printout.

Significance Level	Critical Value	Test Statistic Z	p-Value
0.05	-1.96, 1.96	-0.412	0.6800

This chapter is from Introduction to Statistics for Community College Students, 1st Edition, by Matt Teachout, College of the Canyons, Santa Clarita, CA, USA, and is licensed under a "CC-By" Creative Commons Attribution 4.0 International license – 10/1/18

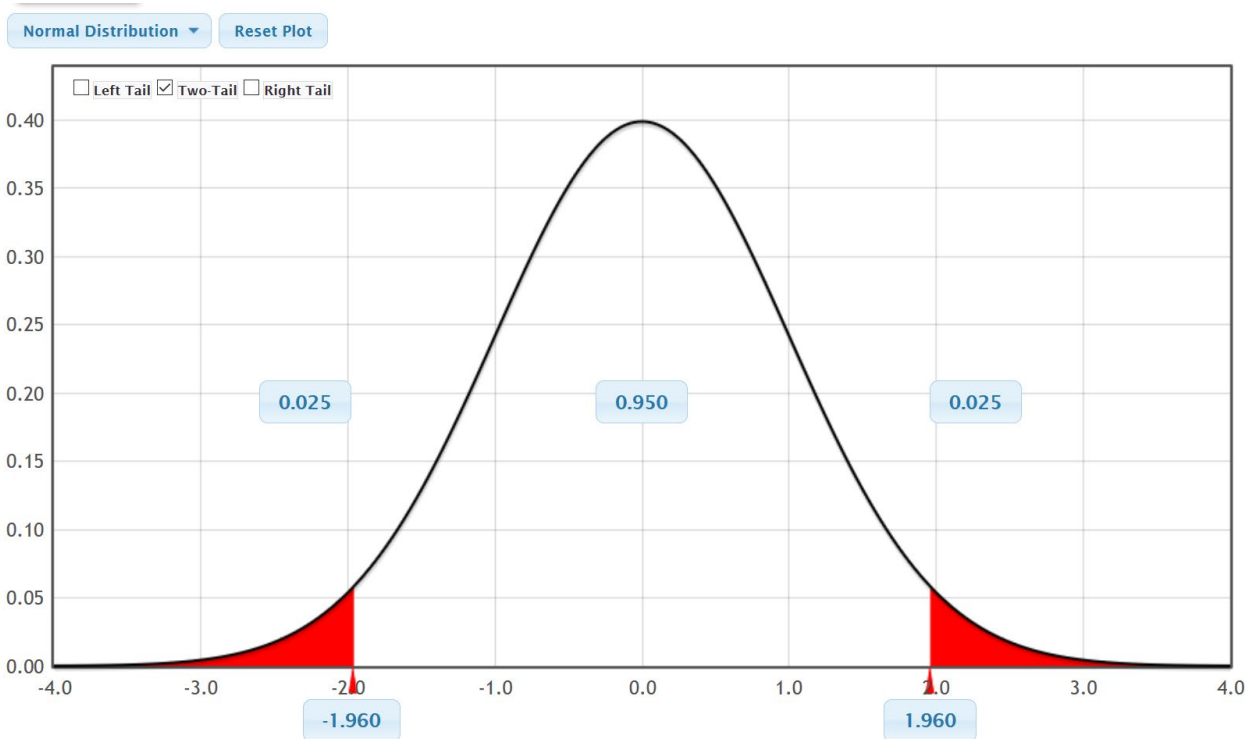
The Z-score test statistic is *−0.412*. Since it is negative, we know that the sample proportion for group 1 is lower than the sample proportion for group 2.

Z-test statistic sentence: The sample proportion for group 1 is 0.412 standard errors below the sample proportion for group 2. *(Notice we did <u>not</u> say "−0.412 standard errors". A Z-score of −0.412 means "0.412 standard errors below".)*

Test statistics tell us if the sample data significantly disagrees with the null hypothesis. Remember the following rules.

- If the test statistic falls in one of the tails determined by the critical value or values, then the sample data significantly disagrees with the null hypothesis.
- If the test statistic falls does NOT fall in one of the tails determined by the critical value or values, then the sample data does NOT significantly disagree with the null hypothesis.

In the Statcato printout, the critical values are ± 1.96. So the Z-test statistic does not fall in either tail. The sample data does not significantly disagree with the null hypothesis.



↑
Z = −0.412

<u>P-value</u>

We also learned in the last chapter, that it is vital to know if the sample data occurred because of sampling variability (random chance). Remember, sample data always disagrees with the null hypothesis to some extent. The key is to determine why it is different. There are two reasons why the sample data disagrees with the null hypothesis. Maybe the null hypothesis is correct and the sample data disagrees because of sample data is always different (random chance). Another option is that the sample data disagrees because the null hypothesis is wrong. The key is that if you determine that it was not random chance, the only other option is that the null hypothesis is wrong. P-value is the key to making this decision about whether the data occurred by random chance. If the P-value is low (close to zero) then it is unlikely to be random chance. If the P-value is high, there is a possibility of the sample data occurring because of random chance.

This chapter is from <u>Introduction to Statistics for Community College Students</u>, 1st Edition, by Matt Teachout, College of the Canyons, Santa Clarita, CA, USA, and is licensed under a "CC-By" <u>Creative Commons Attribution 4.0 International license</u> – 10/1/18

- If P-value ≤ significance level (α), then the sample data is unlikely to have occured by random chance. Since sampling variability is ruled out, the null hypothesis must be wrong. So we "reject the null hypothesis". A low P-value also indicates that the sample data significantly disagrees with the null hypothesis.
- If P-value > significance level (α), then the sample data is could have occurred by random chance. Since we do not know if sampling variability is involved or not, we also do not know if the null hypothesis is right or wrong. So we say we "fail to reject the null hypothesis" in this case. A high P-value also indicates that the sample data does not significantly disagree with the null hypothesis.

Randomized Simulation

In the last chapter, we saw that P-value could be calculated with randomized simulation or a randomization technique. This is a fabulous way for us to visualize what sampling variability (random chance) looks like if the null hypothesis is true. We have a computer create thousands of random samples under the premise that the null hypothesis is true. These simulated samples have the same sample sizes as the original sample data. If the real original sample data falls in the tail of the simulation it indicates that it is significant. The more in the tail the data is, the smaller the P-value. For a one-population proportion test, we see if the sample proportion is in the tail. For a two-population proportion test, we will see if the difference between the two sample proportions falls in the tail.

- If sample statistic falls is in a tail of the simulation, then the sample data is significant and significantly disagrees with the null hypothesis.
- If the sample statistic does not fall in a tail of the simulation, then the sample data is not significant and does not significantly disagree with the null hypothesis.

This chapter is from Introduction to Statistics for Community College Students, 1st Edition, by Matt Teachout, College of the Canyons, Santa Clarita, CA, USA, and is licensed under a "CC-By" Creative Commons Attribution 4.0 International license – 10/1/18

Here is a chart from chapter four that summarizes test statistics, P-value and simulation.

		Significant Test Statistic		Test Statistic NOT Significant
		(Test Statistic falls in tail determined by the critical value or values)		*(Test Statistic does NOT fall in tail determined by the critical value or values)*
		OR		**OR**
		Small P-value		**Large P-value**
		(P-value ≤ significance level)		*(P-value > significance level)*
		OR		**OR**
		Sample Data in Tail		**Sample Data NOT in Tail**
		(when simulating the Null Hypothesis)		*(when simulating the Null Hypothesis)*
Is the sample data significantly		Yes. Significantly different		Not Significantly different
different than H_0?				
Could the sample data happen		Unlikely		Could happen
by random chance				
(sampling variability) if H_0 is true?				
Reject H_0 or		Reject H_0		Fail to Reject H_0
Fail to Reject H_0?				
Is there significant		Yes. Is evidence		No evidence
Evidence?				

Example (Two-Population Proportion Categorical Relationship Test)

Many high school and college students love to listen to music when they study. Some like to listen to their favorite music, while others just like the background noise. Use a 5% significance level to test the claim that liking the music is related to being able to memorize a large amount of information. A randomized experiment was done to test this claim. A group of college students were randomly assigned into two groups. Both groups had to memorize the same amount of information. The number of students that were able to memorize a significant amount of the information were classified as "high retention". One group listened to their favorite music and the other group had to listen to a type of music they hated. Confounding variables like the room environment and music volume were the same in both groups.

Label your variables.

p_1: The percentage of college students that listen to liked music and can memorize a significant amount of information (high retention).

p_2: The percentage of college students that listen to hated music and can memorize a significant amount of information (high retention).

This chapter is from Introduction to Statistics for Community College Students, 1st Edition, by Matt Teachout, College of the Canyons, Santa Clarita, CA, USA, and is licensed under a "CC-By" Creative Commons Attribution 4.0 International license – 10/1/18

Here is the sample data.

Liked Music: 25 total people, 10 high retention, $\hat{p}_1 \approx 0.4$
Hated Music: 24 total people, 11 high retention, $\hat{p}_2 \approx 0.458$

Sample Difference: $\hat{p}_1 - \hat{p}_2 \approx 0.4 - 0.458 = -0.058$

H_0: $p_1 = p_2$ (The population % for high retention is NOT related to liking the music)
H_A: $p_1 \neq p_2$ (The population % for high retention is related to liking the music) CLAIM

(Notice this is a two-tailed test.)

Let us check the assumptions.

Is the sample data random or representative? Yes. The data was not a random sample of the population, but it was randomly assigned. So the sample data will not apply to all college students, but it has the capacity to prove cause and effect.

Is there at least 10 success and 10 failures in the sample data? Yes. In the liked music group, there were 10 high retention and 14 not high retention. In the hated music group, there were 11 high retention and 14 not high retention.

Are data values independent? It is difficult to know this without a detailed look at the people in the experiment. We did not have the same people measured twice, but instead used random assignment. We also should not have family members. If some of the students are friends and know each other, they may have similar taste in music. We will assume this data passes this assumption, but it might need further study.

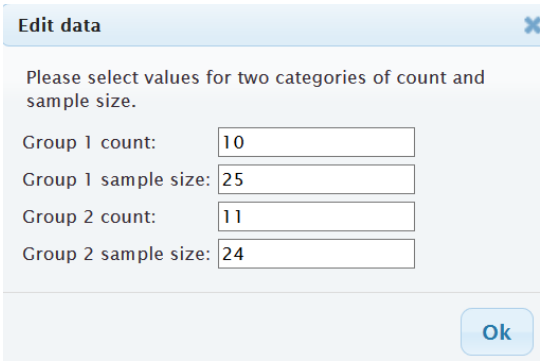
Simulation Approach

Let us use StatKey to simulate the null hypothesis. Remember, the null hypothesis is equivalent to the difference being zero, so the simulation should be centered close to zero.

StatKey Directions for Two Population Proportions (percentages)

Randomization Hypothesis Tests → Test for Difference in Proportions → Under "edit data", put in summary counts → click "generate 1000 samples" multiple times → click on tail determined by the alternative hypothesis → Enter sample proportion difference in bottom box. (If the difference is negative, put it in the left box. If the difference is positive, put it in the right box. P-value will be automatically calculated above the sample difference in the tail.)

We put the data into StatKey, simulated the null hypothesis, and then clicked on "two-tail".

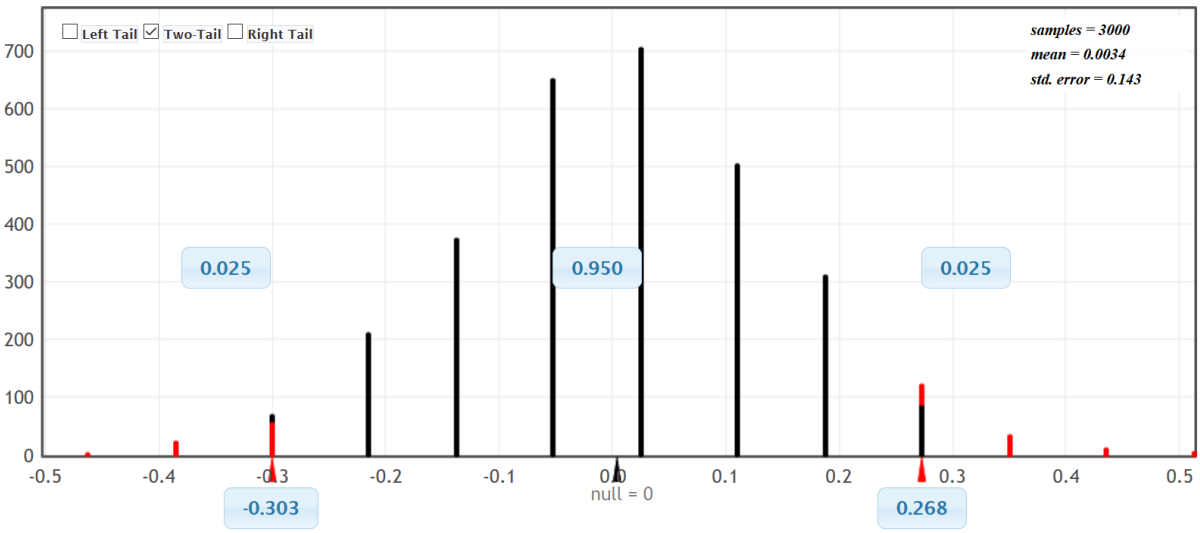


This chapter is from Introduction to Statistics for Community College Students, 1st Edition, by Matt Teachout, College of the Canyons, Santa Clarita, CA, USA, and is licensed under a "CC-By" Creative Commons Attribution 4.0 International license – 10/1/18

Original Sample

Group	Count	Sample Size	Proportion
Group 1	10	25	0.400
Group 2	11	24	0.458
Group 1-Group 2	-1	n/a	-0.058

Randomization Dotplot of $\hat{p}_1 - \hat{p}_2$ **Null Hypothesis:** $p_1 = p_2$



↑
Sample Proportion Difference = −0.058

Notice that in simulation, it is important to identify the tail. With a 5% significance level and a two-tailed test, there is 2.5% in each tail. We see from the simulation that sample differences of approximately −0.303 or less are significant. Also sample differences of approximately +0.268 or higher are significant. Our real sample difference −0.058 was not in either of the tails. The sample data does not significantly disagree with the null hypothesis. It also tells us that the sample proportions are not significantly different.

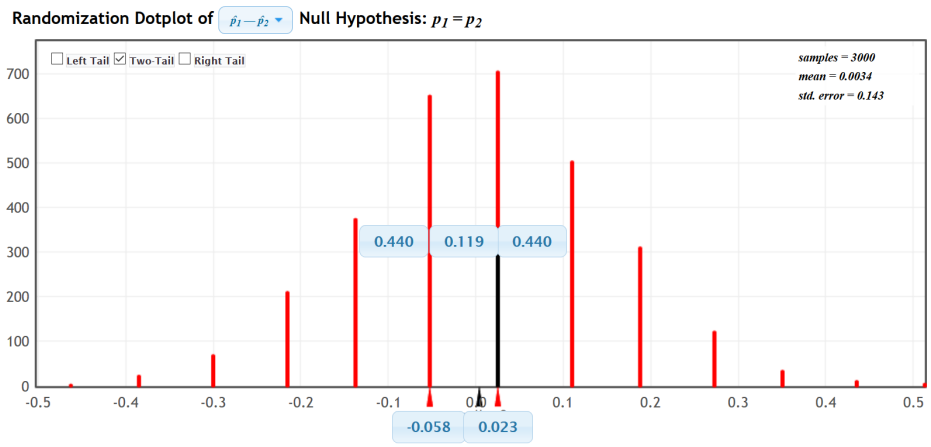
StatKey does not calculate the Z-score test statistic, but we do have the approximate standard error from the simulation of about 0.143. Using the test statistic formula, we get the following.

$$Z = \frac{(sample\ proportion\ 1 - sample\ proportion\ 2)}{standard\ error} = \frac{-0.058}{0.143} \approx -0.41$$

So the sample proportion of high retention for the liked music group was only 0.41 standard errors below the sample proportion of high retention for the hated music group. This is not significant. We do not have a critical value, yet we saw that the sample difference was not in the tail of the simulation.

Now let us use the simulation, to calculate the P-value and check whether this data could have happened because of sampling variability (random chance). If we enter the original sample difference of −0.058 in the left bottom box, we get the following.

This chapter is from Introduction to Statistics for Community College Students, 1st Edition, by Matt Teachout, College of the Canyons, Santa Clarita, CA, USA, and is licensed under a "CC-By" Creative Commons Attribution 4.0 International license – 10/1/18



Notice in a two-tailed test, you need to add the proportions in both tails (upper boxes) to get the P-value.

P-value ≈ 0.440 + 0.440 = 0.880 = 88.0%

P-value Sentence: If the null hypothesis is true, there is an 88.0% probability of getting this sample data or more extreme because of sampling variability.

Interpret the P-value: This is a very large P-value and is much larger than the 5% significance level. This indicates that the population proportions may be equal and the sample data could have happened because of sampling variability. Since sampling variability is involved, we must fail to reject the null hypothesis.

Conclusion: There is not significant evidence to support the claim that liking music is related to high retention. Notice that the alternative hypothesis (related) was the claim and we have a high P-value. Data seems to indicate they are not related, though we do not have significant evidence. This was an experiment with random assignment, so we may say the data indicates that liking the music does not cause a significant difference in the high retention percentage.

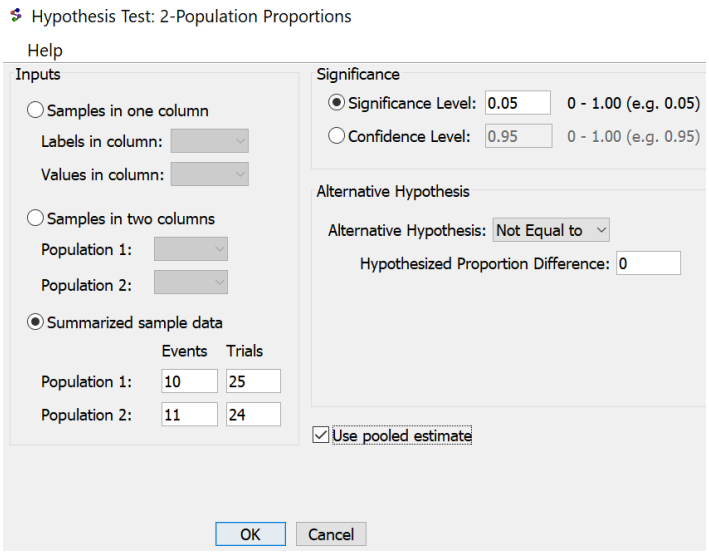
We could also use Statcato to calculate the test statistic, critical values and P-value.

Statcato Directions for Two Population Proportions (percentages)

Statistics → Hypothesis Tests → 2-Population Proportions → Samples in one column, samples in two columns or summarized sample data → put in alternative hypothesis sign (usually ≠ for relationships)
→ Hypothesized proportion difference: 0 → check "use pooled estimate" → put in significance level
→ push "OK".

Here is the Statcato printouts for the same problem.

This chapter is from Introduction to Statistics for Community College Students,
1st Edition, by Matt Teachout, College of the Canyons, Santa Clarita, CA, USA, and is licensed
under a "CC-By" Creative Commons Attribution 4.0 International license – 10/1/18



Hypothesis Test - Two population proportions: confidence level = 0.95

	Number of Events	Number of trials	Proportion
Sample 1	10	25	0.4
Sample 2	11	24	0.458

Null hypothesis: $p_1 - p_2 = 0.0$

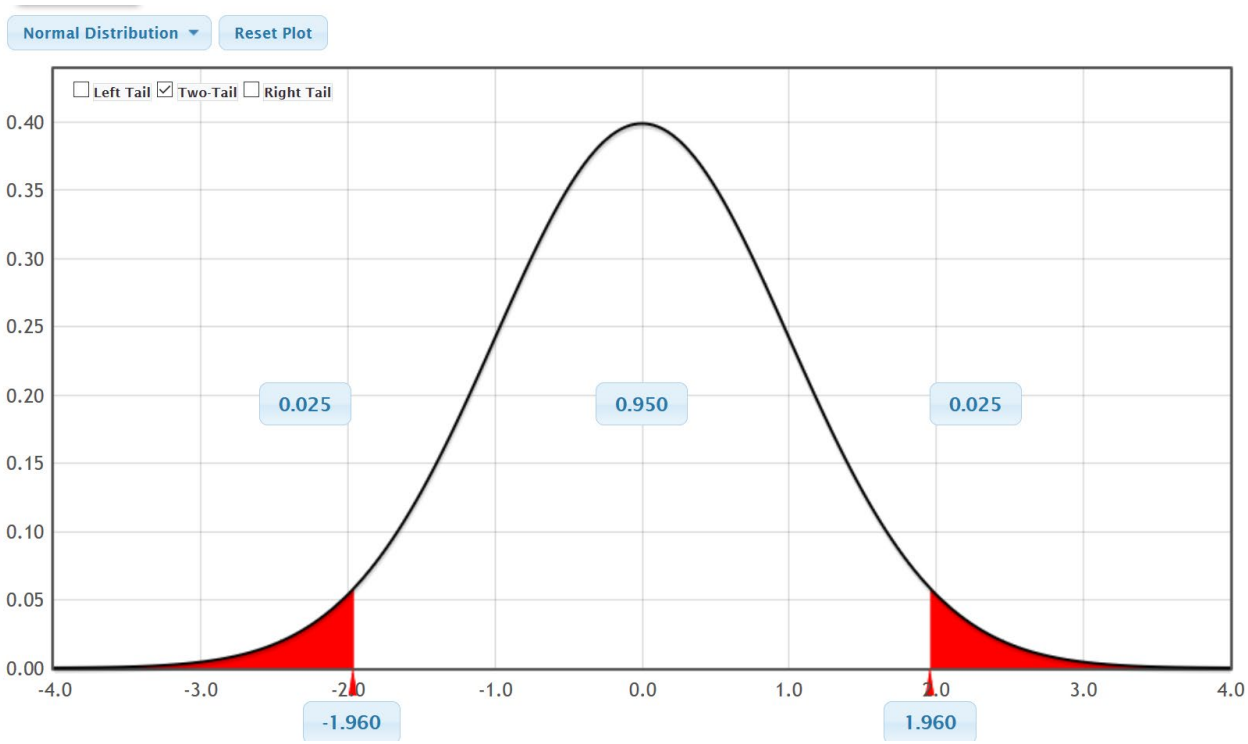
Alternative hypothesis: $p_1 - p_2 \neq 0.0$

Significance Level	Critical Value	Test Statistic Z	p-Value
0.05	-1.96, 1.96	-0.412	0.6800

Notice the Z-test statistic is similar to what we got with StatKey, though the P-value is lower. Notice Statcato gave us critical values to compare the Z-test statistic to. The test statistic does not fall in the tail determined by the critical values. The P-value is still extremely large and indicates that if the null hypothesis was true, this data or more extreme could have happened because of sampling variability (random chance).

Also, notice the way Statcato wrote the null and alternative hypothesis. Saying two parameters are equal is the same as saying the difference is zero. You may see the null and alternative hypothesis written in different ways.

This chapter is from Introduction to Statistics for Community College Students, 1st Edition, by Matt Teachout, College of the Canyons, Santa Clarita, CA, USA, and is licensed under a "CC-By" Creative Commons Attribution 4.0 International license – 10/1/18



↑

Z = −0.412

This chapter is from Introduction to Statistics for Community College Students, 1st Edition, by Matt Teachout, College of the Canyons, Santa Clarita, CA, USA, and is licensed under a "CC-By" Creative Commons Attribution 4.0 International license – 10/1/18