

9.1: Two Proportions

There are times you want to test a claim about two population proportions or construct a confidence interval estimate of the difference between two population proportions. As with all other hypothesis tests and confidence intervals, the process is the same though the formulas and assumptions are different.

Hypothesis Test for Two Populations Proportion (2-Prop Test)

1. State the random variables and the parameters in words.

x_1 = number of successes from group 1

x_2 = number of successes from group 2

p_1 = proportion of successes in group 1

p_2 = proportion of successes in group 2

2. State the null and alternative hypotheses and the level of significance

$H_o : p_1 = p_2$ or $H_o : p_1 - p_2 = 0$

$H_A : p_1 < p_2$ $H_A : p_1 - p_2 < 0$

$H_A : p_1 > p_2$ $H_A : p_1 - p_2 > 0$

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

Also, state your α level here.

3. State and check the assumptions for a hypothesis test

a. A simple random sample of size n_1 is taken from population 1, and a simple random sample of size n_2 is taken from population 2.

b. The samples are independent.

c. The assumptions for the binomial distribution are satisfied for both populations.

d. To determine the sampling distribution of \hat{p}_1 , you need to show that $n_1 p_1 \geq 5$ and $n_1 q_1 \geq 5$, where $q_1 = 1 - p_1$. If this requirement is true, then the sampling distribution of \hat{p}_1 is well approximated by a normal curve. To determine the sampling distribution of \hat{p}_2 , you need to show that $n_2 p_2 \geq 5$ and $n_2 q_2 \geq 5$, where $q_2 = 1 - p_2$. If this requirement is true, then the sampling distribution of \hat{p}_2 is well approximated by a normal curve. However, you do not know p_1 and p_2 , so you need to use \hat{p}_1 and instead \hat{p}_2 . This is not perfect, but it is the best you can do. Since $n_1 \hat{p}_1 = n_1 \frac{x_1}{n_1} = x_1$ (and similar for the other calculations) you just need to make sure that x_1 , $n_1 - x_1$, $n_2 - x_2$, and are all more than 5.

4. Find the sample statistics, test statistic, and p-value

Sample Proportion:

n_1 = size of sample 1 n_2 = size of sample 2

$\hat{p}_1 = \frac{x_1}{n_1}$ (sample 1 proportion) $\hat{p}_2 = \frac{x_2}{n_2}$ (sample 2 proportion)

$\hat{q}_1 = 1 - \hat{p}_1$ (complement of \hat{p}_1) $\hat{q}_2 = 1 - \hat{p}_2$ (complement of \hat{p}_2)

Pooled Sample Proportion, \bar{p} :

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

$$\bar{q} = 1 - \bar{p}$$

Test Statistic:

$$z = \frac{(\hat{p}_1 - \hat{p}_2) - (p_1 - p_2)}{\sqrt{\frac{\bar{p}\bar{q}}{n_1} + \frac{\bar{p}\bar{q}}{n_2}}}$$

Usually $p_1 - p_2 = 0$, since $H_o : p_1 = p_2$

p-value: On TI-83/84: use normalcdf(lower limit, upper limit, 0, 1)

Note

If $H_A : p_1 < p_2$ then lower limit is $-1E99$ and upper limit is your test statistic. If $H_A : p_1 > p_2$, then lower limit is your test statistic and the upper limit is $1E99$. If $H_A : p_1 \neq p_2$, then find the p-value for $H_A : p_1 < p_2$, and multiply by 2.

On R: use `pnorm(z, 0, 1)`

Note

If $H_A : p_1 < p_2$, then use $\text{pnorm}(z, 0, 1)$. If $H_A : p_1 > p_2$, then use $1 - \text{pnorm}(z, 0, 1)$. If $H_A : p_1 \neq p_2$, then find the p-value for $H_A : p_1 < p_2$, and multiply by 2.

5. Conclusion This is where you write reject H_o or fail to reject H_o . The rule is: if the p-value $< \alpha$, then reject H_o . If the p-value $\geq \alpha$, then fail to reject H_o .
6. Interpretation This is where you interpret in real world terms the conclusion to the test. The conclusion for a hypothesis test is that you either have enough evidence to show H_A is true, or you do not have enough evidence to show H_A is true.

Confidence Interval for the Difference Between Two Population Proportion (2-Prop Interval)

The confidence interval for the difference in proportions has the same random variables and proportions and the same assumptions as the hypothesis test for two proportions. If you have already completed the hypothesis test, then you do not need to state them again. If you haven't completed the hypothesis test, then state the random variables and proportions and state and check the assumptions before completing the confidence interval step

1. Find the sample statistics and the confidence interval

Sample Proportion:

$$\begin{aligned} n_1 &= \text{size of sample 1} & n_2 &= \text{size of sample 2} \\ \hat{p}_1 &= \frac{x_1}{n_1} \text{ (sample 1 proportion)} & \hat{p}_2 &= \frac{x_2}{n_2} \text{ (sample 2 proportion)} \\ \hat{q}_1 &= 1 - \hat{p}_1 \text{ (complement of } \hat{p}_1) & \hat{q}_2 &= 1 - \hat{p}_2 \text{ (complement of } \hat{p}_2) \end{aligned}$$

Confidence Interval:

The confidence interval estimate of the difference $p_1 - p_2$ is

$$(\hat{p}_1 - \hat{p}_2) - E < p_1 - p_2 < (\hat{p}_1 - \hat{p}_2) + E$$

where the margin of error E is given by $E = z_c \sqrt{\frac{\hat{p}_1 \hat{q}_1}{n_1} + \frac{\hat{p}_2 \hat{q}_2}{n_2}}$

z_c = critical value

2. Statistical Interpretation: In general this looks like, "there is a C% chance that $(\hat{p}_1 - \hat{p}_2) - E < p_1 - p_2 < (\hat{p}_1 - \hat{p}_2) + E$ contains the true difference in proportions."
3. Real World Interpretation: This is where you state how much more (or less) the first proportion is from the second proportion.

The critical value is a value from the normal distribution. Since a confidence interval is found by adding and subtracting a margin of error amount from the sample proportion, and the interval has a probability of being true, then you can think of this as the statement $P((\hat{p}_1 - \hat{p}_2) - E < p_1 - p_2 < (\hat{p}_1 - \hat{p}_2) + E) = C$. So you can use the `invNorm` command on the TI-83/84 calculator or `qnorm` on R to find the critical value. These are always the same value, so it is easier to just look at the table A.1 in the Appendix.

Example 9.1.1 hypothesis test for two population proportions

Do husbands cheat on their wives more than wives cheat on their husbands ("Statistics brain," 2013)? Suppose you take a group of 1000 randomly selected husbands and find that 231 had cheated on their wives. Suppose in a group of 1200 randomly selected wives, 176 cheated on their husbands. Do the data show that the proportion of husbands who cheat on their wives are more than the proportion of wives who cheat on their husbands. Test at the 5% level.

1. State the random variables and the parameters in words.
2. State the null and alternative hypotheses and the level of significance.
3. State and check the assumptions for a hypothesis test.
4. Find the sample statistics, test statistic, and p-value.
5. Conclusion
6. Interpretation

Solution

1. x_1 = number of husbands who cheat on his wife
 x_2 = number of wives who cheat on her husband

p_1 = proportion of husbands who cheat on his wife

p_2 = proportion of wives who cheat on her husband

$$H_o : p_1 = p_2 \quad \text{or} \quad H_o : p_1 - p_2 = 0$$

$$2. H_A : p_1 > p_2 \quad H_A : p_1 - p_2 > 0$$

$$\alpha = 0.05$$

3.

- A simple random sample of 1000 responses about cheating from husbands is taken. This was stated in the problem. A simple random sample of 1200 responses about cheating from wives is taken. This was stated in the problem.
- The samples are independent. This is true since the samples involved different genders.
- The properties of the binomial distribution are satisfied in both populations. This is true since there are only two responses, there are a fixed number of trials, the probability of a success is the same, and the trials are independent.
- The sampling distributions of \hat{p}_1 and \hat{p}_2 can be approximated with a normal distribution.
 $x_1 = 231, n_1 - x_1 = 1000 - 231 = 769, x_2 = 176$, and
 $n_2 - x_2 = 1200 - 176 = 1024$ are all greater than or equal to 5. So both sampling distributions of \hat{p}_1 and \hat{p}_2 can be approximated with a normal distribution.

4. Sample Proportion:

$$n_1 = 1000$$

$$n_2 = 1200$$

$$\hat{p}_1 = \frac{231}{1000} = 0.231$$

$$\hat{p}_2 = \frac{176}{1200} \approx 0.1467$$

$$\hat{q}_1 = 1 - \frac{231}{1000} = \frac{769}{1000} = 0.769 \quad \hat{q}_2 = 1 - \frac{176}{1200} = \frac{1024}{1200} \approx 0.8533$$

Pooled Sample Proportion, \bar{p} :

$$\bar{p} = \frac{231 + 176}{1000 + 1200} = \frac{407}{2200} = 0.185$$

$$\bar{q} = 1 - \frac{407}{2200} = \frac{1793}{2200} = 0.815$$

Test Statistic:

$$z = \frac{(0.231 - 0.1467) - 0}{\sqrt{\frac{0.185 * 0.815}{1000} + \frac{0.185 * 0.815}{1200}}} = 5.0704$$

p-value:

$$\text{On TI-83/84: normalcdf}(5.0704, 1E99, 0, 1) = 1.988 \times 10^{-7}$$

$$\text{On R: } 1 - \text{pnorm}(5.0704, 0, 1) = 1.988 \times 10^{-7}$$

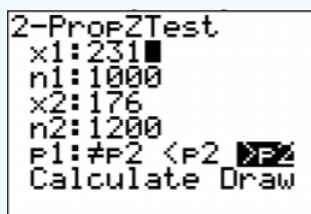


Figure 9.1.1: Setup for 2-PropZTest on TI-83/84 Calculator

```
2-PropZTest
P1>P2
↑P1=.231
P2=.1466666667
P=.185
n1=1000
n2=1200
```

Figure 9.1.2: Results for 2-PropZTest on TI-83/84 Calculator

```
2-PropZTest
P1>P2
z=5.072404516
P=1.9674319E-7
P1=.231
P2=.1466666667
↓P=.185
```

Figure 9.1.3: Results for 2-PropZTest on TI-83/84: Calculator

On R: `prop.test(c(x1, x2), c(n1, n2), alternative = "less" or "greater".` For this example, `prop.test(c(231, 176), c(1000, 1200), alternative="greater")`

2-sample test for equality of proportions with continuity correction

data: c(231, 176) out of c(1000, 1200)

X-squared = 25.173, df = 1, p-value = 2.621e-07

alternative hypothesis: greater

95 percent confidence interval:

0.05579805 1.00000000

sample estimates:

prop 1 prop 2

0.2310000 0.1466667

Note

The answer from R is the p-value. It is different from the formula or the TI-83/84 calculator due to a continuity correction that R does.

5. Conclusion

Reject H_0 , since the p-value is less than 5%.

6. Interpretation This is enough evidence to show that the proportion of husbands having affairs is more than the proportion of wives having affairs.

Example 9.1.2 confidence interval for two population properties

Do more husbands cheat on their wives more than wives cheat on the husbands ("Statistics brain," 2013)? Suppose you take a group of 1000 randomly selected husbands and find that 231 had cheated on their wives. Suppose in a group of 1200 randomly selected wives, 176 cheated on their husbands. Estimate the difference in the proportion of husbands and wives who cheat on their spouses using a 95% confidence level.

1. State the random variables and the parameters in words.
2. State and check the assumptions for the confidence interval.
3. Find the sample statistics and the confidence interval.
4. Statistical Interpretation
5. Real World Interpretation

Solution

1. These were stated in Example 9.1.1, but are reproduced here for reference.

x_1 = number of husbands who cheat on his wife

x_2 = number of wives who cheat on her husband

p_1 = proportion of husbands who cheat on his wife

p_2 = proportion of wives who cheat on her husband

2. The assumptions were stated and checked in Example 9.1.1.

3. Sample Proportion:

$$\begin{aligned} n_1 &= 1000 & n_2 &= 1200 \\ \hat{p}_1 &= \frac{231}{1000} = 0.231 & \hat{p}_2 &= \frac{176}{1200} \approx 0.1467 \\ \hat{q}_1 &= 1 - \frac{231}{1000} = \frac{769}{1000} = 0.769 & \hat{q}_2 &= 1 - \frac{176}{1200} = \frac{1024}{1200} \approx 0.8533 \end{aligned}$$

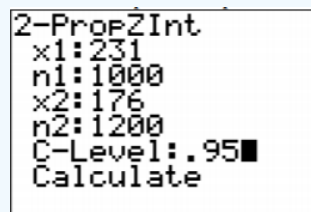
Confidence Interval:

$$z_c = 1.96$$

$$E = 1.96 \sqrt{\frac{0.231 * 0.769}{1000} + \frac{0.1467 * 0.8533}{1200}} = 0.033$$

The confidence interval estimate of the difference $p_1 - p_2$ is

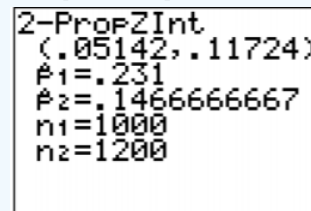
$$\begin{aligned} (\hat{p}_1 - \hat{p}_2) - E &< p_1 - p_2 < (\hat{p}_1 - \hat{p}_2) + E \\ (0.231 - 0.1467) - 0.033 &< p_1 - p_2 < (0.231 - 0.1467) + 0.033 \\ 0.0513 &< p_1 - p_2 < 0.1173 \end{aligned}$$



```

2-PropZInt
x1:231
n1:1000
x2:176
n2:1200
C-Level:.95
Calculate
    
```

Figure 9.1.4: Setup for 2-PropZInt on TI-83/84 Calculator



```

2-PropZInt
(.05142,.11724)
p1=.231
p2=.1466666667
n1=1000
n2=1200
    
```

Figure 9.1.5: Results for 2-PropZInt on TI-83/84 Calculator

On R: `prop.test(c(x1, x2), c(n1, n2), conf.level = C)`, where C is in decimal form. For this example, `prop.test(c(231,176), c(1000, 1200), conf.level=0.95)`

2-sample test for equality of proportions with continuity correction

data: c(231, 176) out of c(1000, 1200)

X-squared = 25.173, df = 1, p-value = 5.241e-07

alternative hypothesis: two.sided

95 percent confidence interval:

0.05050705 0.11815962

sample estimates:

prop 1 prop 2
0.2310000 0.1466667

Note

The answer from R is the confidence interval. It is different from the formula or the TI-83/84 calculator due to a continuity correction that R does.

4. Statistical Interpretation: There is a 95% chance that $0.0505 < p_1 - p_2 < 0.1182$ contains the true difference in proportions.
5. Real World Interpretation: The proportion of husbands who cheat is anywhere from 5.05% to 11.82% higher than the proportion of wives who cheat.

Homework

Exercise 9.1.1

In each problem show all steps of the hypothesis test or confidence interval. If some of the assumptions are not met, note that the results of the test or interval may not be correct and then continue the process of the hypothesis test or confidence interval.

1. Many high school students take the AP tests in different subject areas. In 2007, of the 144,796 students who took the biology exam 84,199 of them were female. In that same year, of the 211,693 students who took the calculus AB exam 102,598 of them were female ("AP exam scores," 2013). Is there enough evidence to show that the proportion of female students taking the biology exam is higher than the proportion of female students taking the calculus AB exam? Test at the 5% level.
2. Many high school students take the AP tests in different subject areas. In 2007, of the 144,796 students who took the biology exam 84,199 of them were female. In that same year, of the 211,693 students who took the calculus AB exam 102,598 of them were female ("AP exam scores," 2013). Estimate the difference in the proportion of female students taking the biology exam and female students taking the calculus AB exam using a 90% confidence level.
3. Many high school students take the AP tests in different subject areas. In 2007, of the 211,693 students who took the calculus AB exam 102,598 of them were female and 109,095 of them were male ("AP exam scores," 2013). Is there enough evidence to show that the proportion of female students taking the calculus AB exam is different from the proportion of male students taking the calculus AB exam? Test at the 5% level.
4. Many high school students take the AP tests in different subject areas. In 2007, of the 211,693 students who took the calculus AB exam 102,598 of them were female and 109,095 of them were male ("AP exam scores," 2013). Estimate using a 90% level the difference in proportion of female students taking the calculus AB exam versus male students taking the calculus AB exam.
5. Are there more children diagnosed with Autism Spectrum Disorder (ASD) in states that have larger urban areas over states that are mostly rural? In the state of Pennsylvania, a fairly urban state, there are 245 eight year olds diagnosed with ASD out of 18,440 eight year olds evaluated. In the state of Utah, a fairly rural state, there are 45 eight year olds diagnosed with ASD out of 2,123 eight year olds evaluated ("Autism and developmental," 2008). Is there enough evidence to show that the proportion of children diagnosed with ASD in Pennsylvania is more than the proportion in Utah? Test at the 1% level.
6. Are there more children diagnosed with Autism Spectrum Disorder (ASD) in states that have larger urban areas over states that are mostly rural? In the state of Pennsylvania, a fairly urban state, there are 245 eight year olds diagnosed with ASD out of 18,440 eight year olds evaluated. In the state of Utah, a fairly rural state, there are 45 eight year olds diagnosed with ASD out of 2,123 eight year olds evaluated ("Autism and developmental," 2008). Estimate the difference in proportion of children diagnosed with ASD between Pennsylvania and Utah. Use a 98% confidence level.
7. A child dying from an accidental poisoning is a terrible incident. Is it more likely that a male child will get into poison than a female child? To find this out, data was collected that showed that out of 1830 children between the ages one and four who pass away from poisoning, 1031 were males and 799 were females (Flanagan, Rooney & Griffiths, 2005). Do the data show that there are more male children dying of poisoning than female children? Test at the 1% level.

8. A child dying from an accidental poisoning is a terrible incident. Is it more likely that a male child will get into poison than a female child? To find this out, data was collected that showed that out of 1830 children between the ages one and four who pass away from poisoning, 1031 were males and 799 were females (Flanagan, Rooney & Griffiths, 2005). Compute a 99% confidence interval for the difference in proportions of poisoning deaths of male and female children ages one to four.

Answer

For all hypothesis tests, just the conclusion is given. For all confidence intervals, just the interval using technology (Software R) is given. See solution for the entire answer.

1. Reject H_0
2. $0.0941 < p_1 - p_2 < 0.0996$
3. Reject H_0
4. $-0.0332 < p_1 - p_2 < -0.0282$
5. Fail to reject H_0
6. $-0.01547 < p_1 - p_2 < -0.0001$
7. Reject H_0
8. $0.0840 < p_1 - p_2 < 0.1696$

This page titled [9.1: Two Proportions](#) is shared under a [CC BY-SA 4.0](#) license and was authored, remixed, and/or curated by [Kathryn Kozak](#) via [source content](#) that was edited to the style and standards of the LibreTexts platform.