Ch 10.3 Hypothesis Test for 2 Proportions

Ch 10.3 Hypothesis Test for 2 proportions.

Terms:

Significant different – means the parameters are not the same statistically, taking into considering of sampling error.

Independent vs dependent samples: Samples are independent if values are not naturally paired.

Inference about proportions from two populations.

1) Test a claim about p1 and p2 where p1 and p2 are population proportion (of the same category) from population one and two.

2) Estimate the confidence interval of the difference of p1 – p2.

Notations:

$\hat{p_1} = \frac{x_1}{n_1}, \hat{p_2} = \frac{x_2}{n_2}$, These are sample proportions from the two populations.

($x_1 = n_1 \cdot \hat{p_1}, x_2 = n_2 \cdot \hat{p_2}$)

Pooled sample: (variances are pooled)

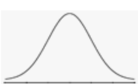
Assume p1 and p2 are the same, we can combine the two samples to create a pooled sample $\bar{p} = \frac{x_1+x_2}{n_1+n_2}$;

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

The difference of two sample proportions $\hat{p1} - \hat{p2}$ will be normally distributed with

mean = p1 – p2 , sd = $\sqrt{\frac{\bar{p}\bar{q}}{n_1} + \frac{\bar{p}\bar{q}}{n_2}}$. where

$x_1, x_2, n_1 - x_1, n_2 - x_2$ are at least 5.



Steps for Testing claim of 2 population proportions:

1) Define populations, record n1, x1, n2, x2,

Requirement is x1, n1-x1, x2, n2-x2 are at least 5.

2) Record the claim in symbolic form:

p1 (=, < , > or ≠) p2.

Write H0 (p1 = p2) and Ha (p1 <, > or ≠ p2) in symbolic form.

Note: p1 must be on the left, p2 on the right.

3) Identify significant level, type of test (left-tail, right-tail or two-tail test) and the sampling distribution.

For proportion, distribution is z-normal.

4) Use Statdisk/Hypothesis Testing/proportion two samples find test statistic and p-value.

- Select Ha: p1 not = 2 or p1 >p2 or p1<p2.
- Enter significant level: α
- Enter x1, n1, x2, n2. Evaluate.

Output:

· test statistic, z – describe how many SD the sample difference $\hat{p}_1 - \hat{p}_2$ is from the H0 assumption of p1 = p2, (p1 – p2 = 0) in a normal distribution.

· P-value: probability of getting the sample difference $\hat{p}_1 - \hat{p}_2$ or worse if p1 = p2 is true.

· Confidence interval at appropriate C-level *lower bound* < p1 – p2 < *upper bound*

5) If P-value ≤ α, reject H0, conclude there is significant difference between p1 and p2.

If p-value > α, fail to reject H0, conclude there is no significant difference between p1 and p2. The difference $\hat{p}_1 - \hat{p}_2$ are due to sample variation.

6) Write conclusion about the claim:

There is (sufficient / not sufficient) evidence to (reject / support) the claim.

Use "sufficient evidence" if sample is significant.

Use "reject" if claim is H0, use "support" if claim is Ha.

7) Make conclusion using Confidence Interval:

a) Confidence interval is included in an hypothesis test output. OR

b) Use Statdisk/Confidence Interval/Proportion 2 sample to find confidence interval.

Input appropriate C-level:

two-tail test: C-level = 1 – α.

Left-tail or right-tail test: C-level = 1 - 2α

input n1,x1, n2, x2.

Output: *lower limit* < p1 – p2 < *upper limit.*

Make conclusion below:

i) If the interval contains zero or both positive and negative values, conclude p1 and p2 has no significant difference. Sample is not significant.

(L limit is negative U limit is positive.)

ii) If the interval is all positive, conclude p1 > p2

(L. Limit and U. Limits are both positive.)

iii) If the interval is all negative, conclude p1 < p2

(L. limit and U. Limits are both negative.)

Note: appropriate C-level: Two tail test: Clevel = 1 - α

Left-tail or right-tail test: C-level = 1 - 2α

Summary:

p1 – p2 > 0 implies p1 > p2
p1 – p2 < 0 implies p1 < p2
p1 – p2 = 0 implies p1 = p2

Note: Conclusion from hypothesis test is more accurate than conclusion from Confidence Interval for Proportion test. Check p-value to confirm if sample difference is significant.

p-value ≤ α , there is significant difference.

p-value > α , there is no significant difference.

Ex1: Given the table below about number of cars with Rare and Front License plates at two states.

	Connecticut	New York
Cars with rare license plates only	239	9
cars with both front and rare license plates	1810	541
Total	2049	550

The sample proportion for Connecticut is 239/2049 (11.7%) The sample proportion for New York = 9/550 (1.6%). The samples are different, but are they statistically different?

a) Test with a significant level of 0.05 that the proportions of car with only rare license plate in Connecticut and New York are the same.

1) Define population 1 = Connecticut, , n1= 2049, x1= 239, population 2 = New York, n2 = 550, x2= 9,.

2) Claim: p1 = p2, H0: p1 = p2 , Ha: p1 ≠ p2

3) Significant level = 0.05, Ha use ≠, so two-tail test,

distribution is z.

4) Use Statdisk/Hypothesis/proportion 2 samples, Select Ha: p1 not = p2. significance = 0.05,

input n1=2049, x1=239,n2=550,x2=9,. Evaluate.

Test statistic z = 7.11, p-value = 0.0000

5) Since p-value (0.0000) < α(0.05), reject H0

6) Since H0 is the claim and we rejected H0,

There is sufficience evidence to reject the claim that the proportion of cars with rare only License plate are the same with that of New York.

b) The 95% confidence Interval of the difference is

0.083 < p1 – p2 < 0.118

Conclude the difference is between 8.3% to 11.8% at a confidence level of 95%.

Since the interval contains all positive values, the two proportions are significantly different and not the same as stated in the claim.

Note: when x is not given: x = p n

when n is not given: n = successs + failure.

Ex2: An experiment is set up to test the effectiveness of Aspirin in preventing heart disease. 11,037 adults were treated with aspirin and another 11,034 adults were given placebo. Among the subjects in the treatment group, 1.26% experienced heart attacks. Among the subjects given placebos, 2.17% experienced heart attacks.

a) Use a 0.05 significant level to test the claim that aspirin is effective in lowering heart attacks.

i) Def population 1 – treatment group,

$n1= 11037$, $x1=0.0126(11037) =139$.

population 2 – placebo group,

$n2 = 11034$, $x2=0.0217(11034) = 239$.

ii) Claim $p1< p2$,   H0: $p1 = p2$,   Ha: $p1< p2$

iii) $\alpha= 0.05$, Ha is "<", so left-tail test, use z-distribution

iv) Use Statdisk/hypothesis/Proportion 2 samples, select ha: $p1< p2$, significance = 0.05,

enter n1, x1, n2, x2, evaluate.

Output: test statistic $z = -5.19$, p-value = 0.0000

$-0.012 < p1 – p2 < -0.006$

v) p-value (0.0000) < 0.05 Reject H0, there is significant difference between p1 and p2.

vi) Since claim is H1, we use support statement,

Since we reject H0, we use "sufficient evidence".

Flow chart will be at box 3.

There is sufficient evidence that to support the claim that aspirin is effective in lowering heart attacks.

b) The 90% confidence interval is calculated because the test is a left-tail test. C-level = 1 – 2(0.05) =0.90

$-0.012 < p1 – p2 < -0.006$

Since confidence interval does not contain zero, conclude there is significant difference.

Since the whole interval is negative, conclude $p1 < p2$.

implying aspirin can lower heart attack.

c) Based on the result, would aspirin be recommended for adults to avoid heart attacks?

Yes, the result conclude that aspirin is effective.

d) Would the result be useful if the treatment group are selected from males only and the placebo group are selected from females only.

The result may not apply to general population because the two groups are not only different on the treatment. The result may be due to other confounding factors between male and female.

Ex3. In a randomly picked year from 1985 to present, there were 1840 Hispanic students at Cabrillo College out of a total of 12328 students. At Lake Tahoe College, there were 321 Hispanic students out of a total of 2441 students. Test the claim at a significant level of 1% that the percent of Hispanic student is higher in Cabrillo College.

Ans:

1) Define population 1 – Cabrillo College $n1 = 12328$, $x1 = 1840$, population 2 – Lake Tahoe College, $n2 = 2441$, $x2 = 321$.

Ex4. In a random sample of 100 forests in the United States, 56 were coniferous or contained conifers. In a random sample of 80 forests in Mexico, 40 were coniferous. Is the Is the proportion of conifers in the United States statistically more than the proportion of confiers in Mexico? At a significant level of 1%, construct an appropriate confidence interval to test the claim and estimate the difference.

Ans:

Define population1 -United States forests, n1 = 100, x1= 56.

Population 2 – Mexico's forests, n2 = 80, x2 = 40

Claim: p1 > p2. H0: p1 = p2, Ha: p1> p2 (right tail test)

We can do a confidence interval to determine if p1 > p2.

Appropriate C-level = 1 – 2 α (right-tail test) = 0.98

statdisk/Confidence Interval/proportion 2 samples/

input Clevel = 0.98 (1 – 2(0.01)), n1 = 100, x1 = 56, x2= 80, n2 = 40, evaluate

Output: - 0.114 < p1 – p2 < 0.234

Since 0 is in the interval, conclude there is no statistical difference between p1 and p2. p1 = p2,

Sample is not significant. (use "not sufficient evidence" statement). Ha is claim. (use "support")

Conclusion: There is not sufficient evidence to support the claim that the proportion of conifers in the United States statistically more than the proportion of conifers in Mexico. The difference in percentage can be between -11.4% to 23.4% at 98% confidence