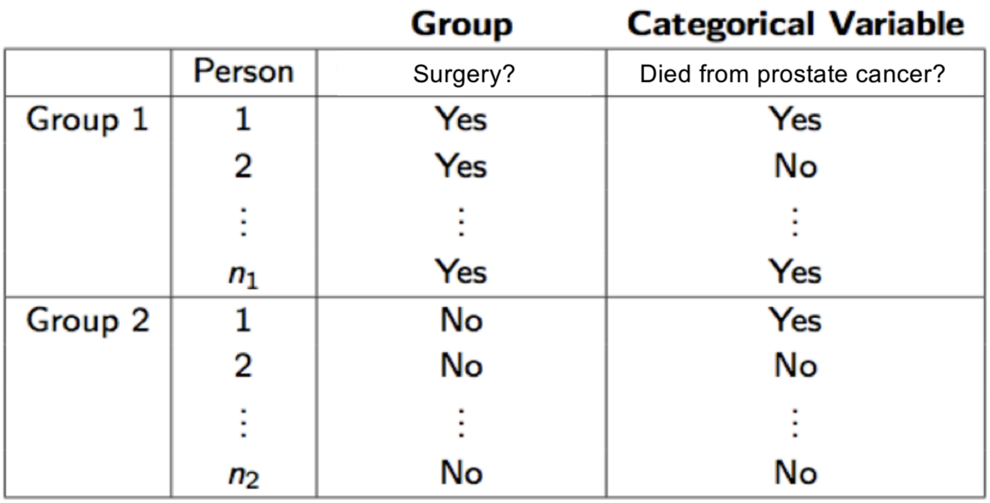
Chapter Twenty: Inference for Difference in Proportions

In previous chapters we learned about hypothesis testing and inference for one sample mean or one sample proportion. But there are a whole litany of questions that could involve **two** proportions or **two** means:

- Is there a difference in the proportion of men that die from prostate cancer for those that undergo a surgery versus those that do not?
- How much taller, on average, are adult males than adult females? ETC



		Group	Categorical Variable
	Person	Surgery?	Died from prostate cancer?
Group 1	1	Yes	Yes
	2	Yes	No
	⋮	⋮	⋮
	n_1	Yes	Yes
Group 2	1	No	Yes
	2	No	No
	⋮	⋮	⋮
	n_2	No	No

In this case the summary statistics are as follows:

$\hat{p}_1$ = proportion of males who had surgery that died from prostate cancer

$\hat{p}_2$ = proportion of males who did not have surgery that died from prostate cancer.

Structure of the Data

Grouping Variable (a categorical variable)

- There will be two groups
- often groups are numbered as 1 & 2 (order doesn't matter)
- These groups are independent from each other.

Variable of Interest

- If categorical: inference for $p_1 - p_2$
- If quantitative: inference for $\mu_1 - \mu_2$

Notation

Population 1

- p_1: proportion of category of interest in population 1
- n_1: sample size from population 1.
- $\hat{p_1}$: sample proportion of category of interest in sample 1.

Population 2

- p_2: proportion of category of interest in population 2
- n_2: sample size from population 2
- $\hat{p_2}$: sample proportion of category of interest in sample 2.

Population 1

- μ_1: mean of variable of interest in population 1
- n_1: sample size from population 1
- $\bar{y_1}$: mean of variable of interest in sample 1
- s_1: standard deviation of variable of interest in sample 1.

Population 2

- μ_2: mean of variable of interest in population 2
- n_2: sample size from population 2
- $\bar{y_2}$: mean of variable of interest in sample 2
- s_2: standard deviation of variable of interest in sample 2.

Example *Determine whether the following situations involve one group or two groups. If it involves two independent groups, identify the groups.*

1. *An educator wants to determine the average reading comprehension scores of her students* one group
2. *An educator assigns half the class to one reading activity and the other half of the class to another reading activity. She wants to determine if the average reading comprehension scores are different between the activities.* two groups; half class with reading activity 1 & half class w/ reading activity 2
3. *We want to compare the proportion of in-state students who get financial aid to the proportion of out-of-state students who get financial aid.* two groups; in-state vs. out-of-state students.
4. *We want to determine if the proportion of students at a university that are in-state students is higher than the national average.* one group.

We will consider the following two types of inferences for difference in proportions.

Confidence Intervals for the difference in population proportions:

CI for $p_1 - p_2$

Hypothesis Test for the difference in population proportions:

HT for $p_1 - p_2$

In this situation, the parameter and statistic are:

parameter: $p_1 - p_2$

statistic: $\hat{p}_1 - \hat{p}_2$

Confidence Interval for Differences in Proportions

Conditions

1. **Randomization condition**: each group comes from a random sample
2. **10% condition**: in each group, sample needs to be less than 10% of the population for both groups.
3. **Success/Failure condition:** in each group,

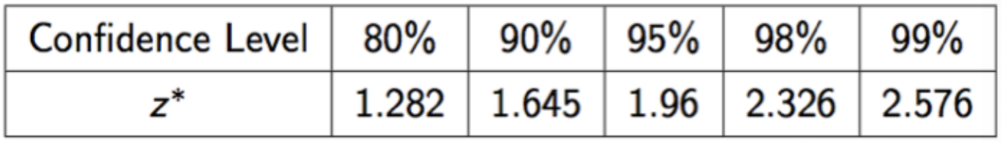
   $n_1\hat{p}_1 \geq 10$, $n_1(1-\hat{p}_1) \geq 10$ $n_2\hat{p}_2 \geq 10$, $n_2(1-\hat{p}_2) \geq 10$
4. **Independent Groups:** groups need to be independent

Formula

If the above conditions are met, the C% confidence interval for $p_1 - p_2$ is:

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

Here the z* is chosen based on the desired C% confidence level:



Confidence Level	80%	90%	95%	98%	99%
*z**	1.282	1.645	1.96	2.326	2.576

Example *There has been debate among doctors oveer whether surgery can prolong life among men suffering from prostate cancer. In a 2003 study published by the New England Journal of Medicine, men diagnosed with prostate cancer were randomly selected that either underwent surgery or not. Men were then followed to see if they died from prostate cancer.*

Find a 95% confidence interval for the difference in population proportions of men that die from prostate cancer for those that undergo surgery versus those that do not.

- ***Group 1 – no surgery***
  - $n_1 = 348$
  - $y_1 = 31$
  - $\hat{p_1} = \frac{31}{348} = 0.0891$
- ***Group 2 – surgery***
  - $n_2 = 347$
  - $y_2 = 16$
  - $\hat{p_2} = \frac{16}{347} = 0.0461$

Assumptions:

i) Both randomly selected

ii) 10% condition met for both groups

iii) $348(.08) = 27.84 \geq 10$, $348(0.92) = 320.16$

$347(.05) = 17.35 \geq 10$, $347(0.95) = 329.65$

iv) each group is independent from one another.

$$(0.0891 - 0.0461) \pm 1.96\sqrt{\frac{.0891(.9109)}{348} + \frac{(.0461)(.9539)}{347}}$$

$$0.043 \pm 0.0372$$

CI: $(0.0058, 0.0802)$

We are 95% confident that the true difference in population proportions of men that die from prostate cance for those that undergo surgery vs. those that do not lies between (0.0058, 0.0802) percent.

Hypothesis Test for Differences in Proportions

Step 1: Hypotheses

Null Hypothesis

- states that the population proportions from each group are equal
- can write the hypothesis in either of the following (equivalent) ways:

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

Note: We could instead test whether the difference is equal to a particular value, but this is rather uncommon.

Alternative Hypothesis

- States that there is some difference between the population proportions
- Choose one of the following based on circumstances of the problem (there are two ways to write each).

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

$H_A: p_1 - p_2 > 0$ vs. $HA: p_1 > p_2$

$H_A: p_1 - p_2 \neq 0$ vs. $HA: p_1 \neq p_2$

Step 2: Assumptions

Check the following conditions:

1. **Randomization condition**: experimental units randomly assigned in each group
2. 10% condition: samples need to be less than 10% of the population for both groups:
3. **Success/Failure condition:** for both groups:
   $n_1\hat{p}_1 \geq 10$, $n_1(1-\hat{p}_1) \geq 10$ $n_2\hat{p}_2 \geq 10$, $n_2(1-\hat{p}_2) \geq 10$
4. **Independent Groups:** groups need to be independent.

Step 3: Test Statistic

Because we are dealing with two sample proportions, we need to create a pooled sample proportion:

$$\hat{p}_{pooled} = \frac{n_1\hat{p}_1 + n_2\hat{p}_2}{n_1 + n_2}$$

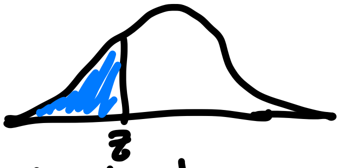
$H_0: \hat{p}_1 - \hat{p}_2 = 0$

Then our z-score is calculated as follows:

$$z = \frac{\hat{p}_1 - \hat{p}_2 - 0}{\sqrt{\frac{\hat{p}_{pooled}(1-\hat{p}_{pooled})}{n_1} + \frac{\hat{p}_{pooled}(1-\hat{p}_{pooled})}{n_2}}}$$

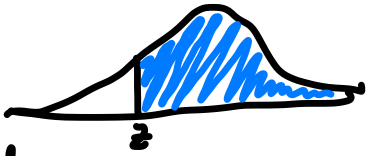
Step 4: Find p-value

We have three different options based on our alternative hypotheses:



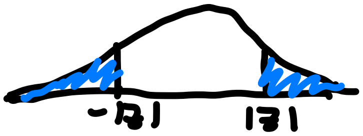
$H_a : p_1 < p_2$

↳ p-value is the area less than z in the standard normal distribution.



$H_a : p_1 > p_2$

↳ p-value is the area greater than z in the standard Normal distribution

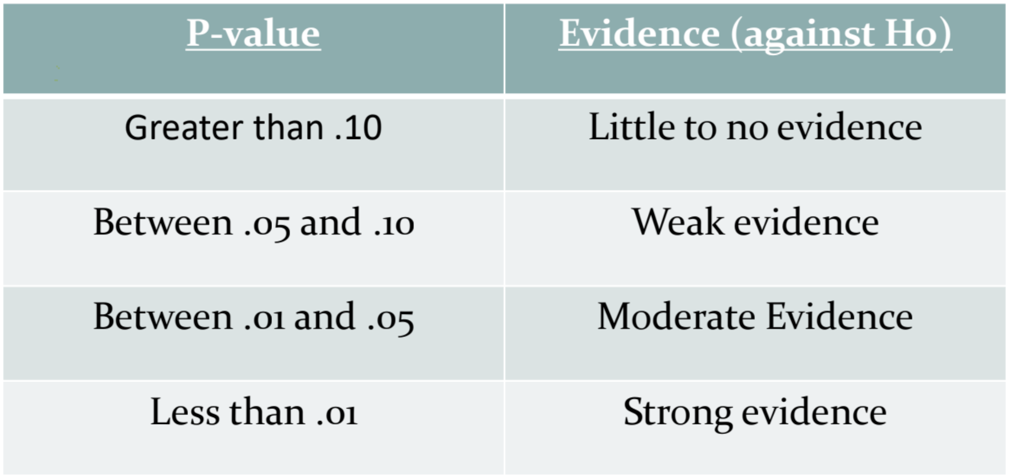


$H_a : p_1 \neq p_2$

↳ p-value is the area in the two tails of the standard normal distribution, outside of -|z| & |z|

↳ Determine whether the area below z or above z is smaller, then find the area of the smaller section & multiply by 2.

Step 5: List your decision



P-value	Evidence (against Ho)
Greater than .10	Little to no evidence
Between .05 and .10	Weak evidence
Between .01 and .05	Moderate Evidence
Less than .01	Strong evidence

Step 6: Conclusion

Make a statement about the relationship between p_1 and p_2 given the information from the hypothesis test.

Be sure to include:

- Parameter (difference in proportions for populations)
- context
- whether or not there is evidence for the alternative hypothesis

Example *There has been debate among doctors oveer whether surgery can prolong life among men suffering from prostate cancer. In a 2003 study published by the New England Journal of Medicine, men diagnosed with prostate cancer were randomly selected that either underwent surgery or not. Men were then followed to see if they died from prostate cancer.*

Perform a hypothesis test to determine if the proportion who died from cancer that received the surgery was lower than the proportion who died from cancer that did not receive the surgery. Use $\alpha = 0.05$.

Step 1: $H_0: P_1 = P_2$

$H_A: P_1 > P_2$

- ***Group 1 – no surgery***
  - $n_1 = 348$
  - $y_1 = 31$
  - $\hat{p_1} = \frac{31}{348} = 0.0891$
- ***Group 2 – surgery***
  - $n_2 = 347$
  - $y_2 = 16$
  - $\hat{p_2} = \frac{16}{347} = 0.0461$

Step 2: Same as page 4 assumptions

Step 3: $\hat{P}_{pooled}: \frac{348(.0891)+347(0.0461)}{348+347}$

$= 0.0676$

$$z = \frac{0.0891 - 0.0461}{\sqrt{\frac{.0676(1-.0676)}{348} + \frac{.0676(1-0.0676)}{347}}} = 2.26$$

Step 4: $P(Z > 2.26) =$

$1 - P(Z < 2.26) =$

z-table

$1 - 0.9881 = 0.0119$

Step 5: Moderate evidence against the null

Step 6: We have moderate evidence to suggest that the proportion who died who did not have the surgery is higher than the proportion of people who died who did have the surgery.