Inference for Differences in Proportion

In previous lectures 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:

$\hat{p}_1 =$
$\hat{p}_2 =$

Structure of the Data

Grouping Variable (a categorical variable)

-
-
-

Variable of Interest

- If categorical:
- If quantitative:

Notation

Population 1

- p_1:
- n_1:
- $\hat{p_1}$:

Population 2

- p_2:
- n_2:
- $\hat{p_2}$:

Population 1

- μ_1:
- n_1:
- $\bar{y_1}$:
- s_1:

Population 2

- μ_2:
- n_2:
- $\bar{y_2}$:
- s_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*
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.*
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.*
4. *We want to determine if the proportion of students at a university that are in-state students is higher than the national average.*

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

Confidence Intervals for the difference in population proportions:

$$\text{CI for } p_1 - p_2$$

Hypothesis Test for the difference in population proportions:

$$\text{HT for } p_1 - p_2$$

In this situation, the parameter and statistic are:

parameter:

statistic:

Confidence Interval for Differences in Proportions

Conditions

1. **Randomization condition**:

2. **10% condition**:

3. **Success/Failure condition:**

4. **Independent Groups:**

Formula

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

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 over 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***
  - $n_2 = 347$
  - $y_2 = 16$
  - $\hat{p_2} = \frac{16}{347} = 0.0461$

Hypothesis Test for Differences in Proportions

Step 1: Hypotheses

Null Hypothesis

-
-

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

Alternative Hypothesis

-
-

Step 2: Assumptions

Check the following conditions:

1. **Randomization condition**:
2. **10% condition**:
3. **Success/Failure condition:**
4. **Independent Groups:**

Step 3: Test Statistic

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

Then our z-score is calculated as follows:

Step 4: Find p-value

We have three different options based on our alternative hypotheses:

$H_a : p_1 < p_2$

$H_a : p_1 > p_2$

$H_a : p_1 \neq p_2$

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:

-
-
-

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 $\alpha = 0.05$.

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

Inference for Difference in Means

Recall the notation for comparing means that we covered in part one:

Population 1

- μ_1:
- n_1:
- $\bar{y_1}$:
- s_1:

Population 2

- μ_2:
- n_2:
- $\bar{y_2}$:
- s_2:

We will consider two type of inference for difference in means:

Confidence Interval for the Differene in Population Means

$$\text{CI for } \mu_1 - \mu_2$$

Hypothesis Test for the Differene in Population Means

$$\text{HT for } \mu_1 - \mu_2$$

In these situations, the parameter and statistic are:

parameter:
statistic:

Confidence Interval for Difference in Means

Conditions

1. **Randomization condition**:

2. **10% condition**:

3. **Nearly normal condition:**

4. **Independent Groups:**

Formula

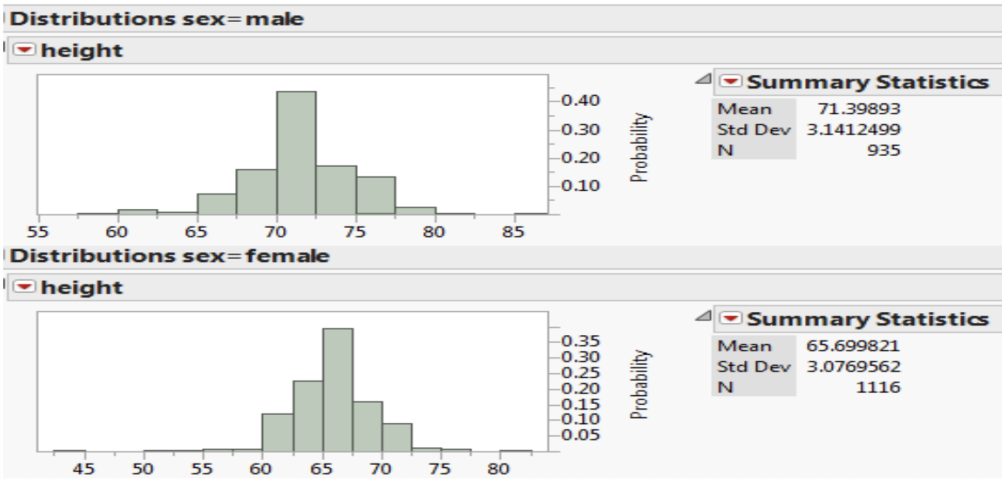
If the conditions above are met, the C% confidence interval for $\mu_1 - \mu_2$ is:

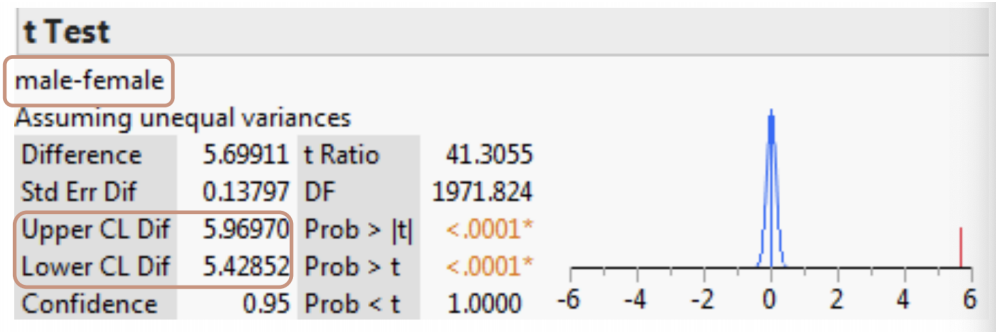
the t* value has degrees of freedom computed using the formula below:

Example *At the beginning of the semester for severl years, students in Stat 101 completed a survey. In this survey, the sex and height (in inches) of the students were recorded.*

Calculate a 95% CI for the mean difference in heights between males and females of the population of Stat 101 students.

- ***Populations***
  - _
  - _
- ***Samples***
  - _
  - _
- ***Parameter**:*
- ***Statistic**:*
  - _
  - _




t Test
male-female
Assuming unequal variances
Difference 5.69911
Std Err Dif 0.13797
Upper CL Dif 5.96970
Lower CL Dif 5.42852
Confidence 0.95
t Ratio 41.3055
DF 1971.824
Prob > |t| <.0001*
Prob > t <.0001*
Prob < t 1.0000
-6 -4 -2 0 2 4 6

Hypothesis Test for the Difference in Means

Step 1: Hypotheses

Null Hypothesis

-
-

Alternative Hypothesis

-
-

Step 2: Assumptions

Check the following conditions:

1. **Randomization condition**:
2. **10% condition**:
3. **Nearly Normal Condition:**
4. **Independent Groups:**

Step 3: Test Statistic

Then our t-statistic is calculated as follows:

Step 4: Find p-value

Remember, the p-value is found using a t-distribution with degrees of freedom. To compute the degrees of freedom, we use the following formula:

We have three different options based on our alternative hypotheses:

$H_a : \mu_1 < \mu_2$

$H_a : \mu_1 > \mu_2$

$H_a : \mu_1 \neq \mu_2$

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 μ_1 and μ_2 given the information from the hypothesis test.

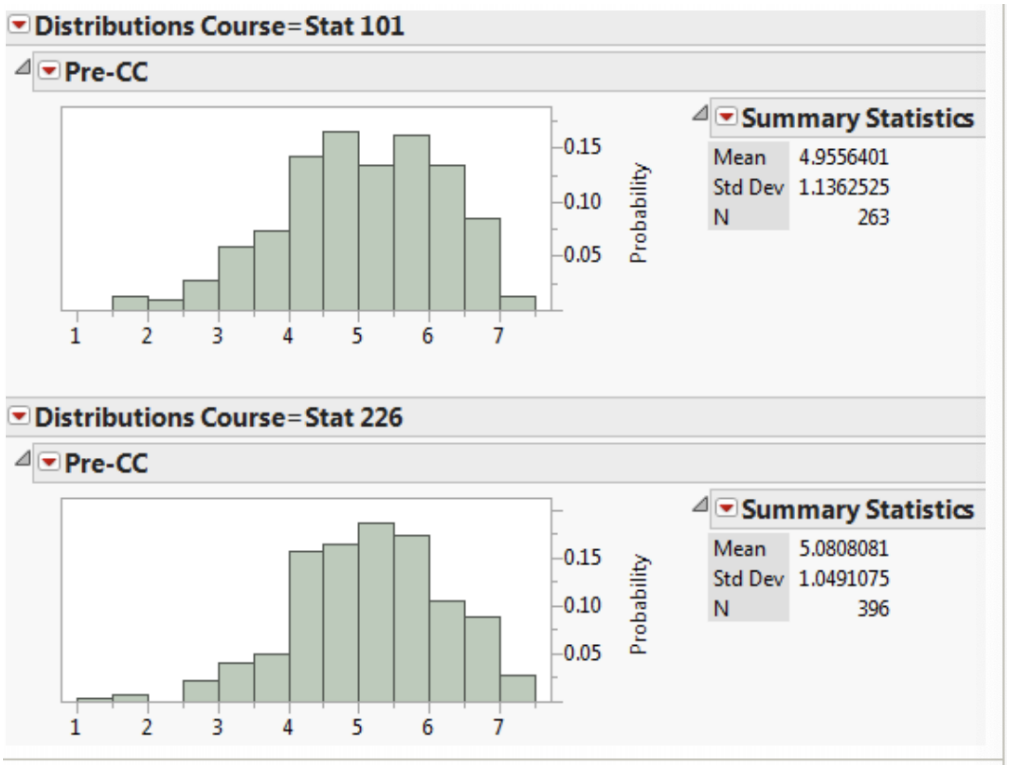
Be sure to include:

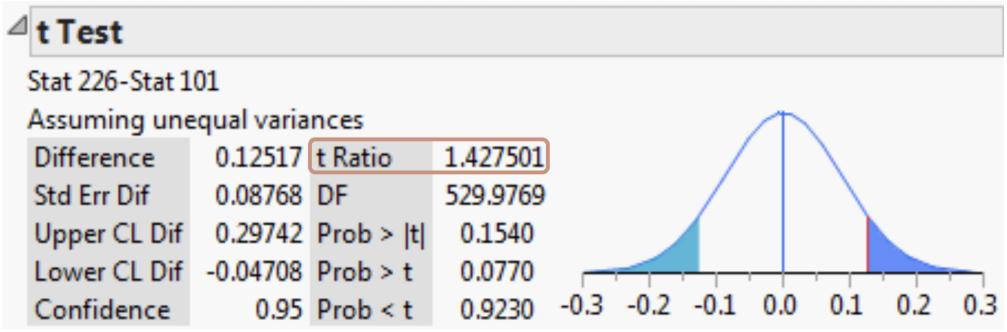
-
-
-

Example *At ISU, several different intro stats courses are offered. Each course is structured according to a particular audience of majors. At the beginning of the Fall 2006 semester, a "Survey of Attitudes Toward Statistics" was administered to students in Stat 101 and Stat 226. One of the components of this survey is called the "cognitive competence" attitude, which is rated on a scale of 1-7 where:*

- *1-3 = negative attitudes*
- *4 = neutral attitude*
- *5-7 = positive attides*

We want to determine if there is evidence that stat 226 students have a higher mean attitude towards "cognitive competence" than stat 101 students. There were 396 stat 226 students and 264 stat 101 students sampled. Our parameter of interest is $\mu_1 - \mu_2$ which means that the population mean attitude score of all stat 226 students minus the population mean attitude score of all stat 101 students.


Distributions Course=Stat 101
Pre-CC
Summary Statistics
Mean 4.9556401
Std Dev 1.1362525
N 263
0.15
0.10
0.05
Probability
1 2 3 4 5 6 7
Distributions Course=Stat 226
Pre-CC
Summary Statistics
Mean 5.0808081
Std Dev 1.0491075
N 396
0.15
0.10
0.05
Probability
1 2 3 4 5 6 7


t Test
Stat 226-Stat 101
Assuming unequal variances
Difference 0.12517 t Ratio 1.427501
Std Err Dif 0.08768 DF 529.9769
Upper CL Dif 0.29742 Prob > |t| 0.1540
Lower CL Dif -0.04708 Prob > t 0.0770
Confidence 0.95 Prob < t 0.9230
-0.3 -0.2 -0.1 0.0 0.1 0.2 0.3