Chapte Twenty Part Two: Inference for Difference in Means

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

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

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: $\mu_1 - \mu_2$
statistic: $\bar{y}_1 - \bar{y}_2$

Confidence Interval for Difference in Means

Conditions

1. **Randomization condition:** each group needs to be taken from a random sample.
2. **10% condition:** in each group, sample size needs to be less than 10% of the population for both groups.
3. **Nearly normal condition:**
   - population must be normally distributed
   - sample size must be sufficiently large...
     - ↳ symmetric but not normal population distribution: n = 10 or larger
     - ↳ skewed population distribution: n = 25 or larger
     - ↳ very skewed population distribution: n = 40 or larger
4. **Independent Groups:**
   ↳ groups need to be independent

Formula

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

$$(\bar{y}_1 - \bar{y}_2) \pm t^* \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

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

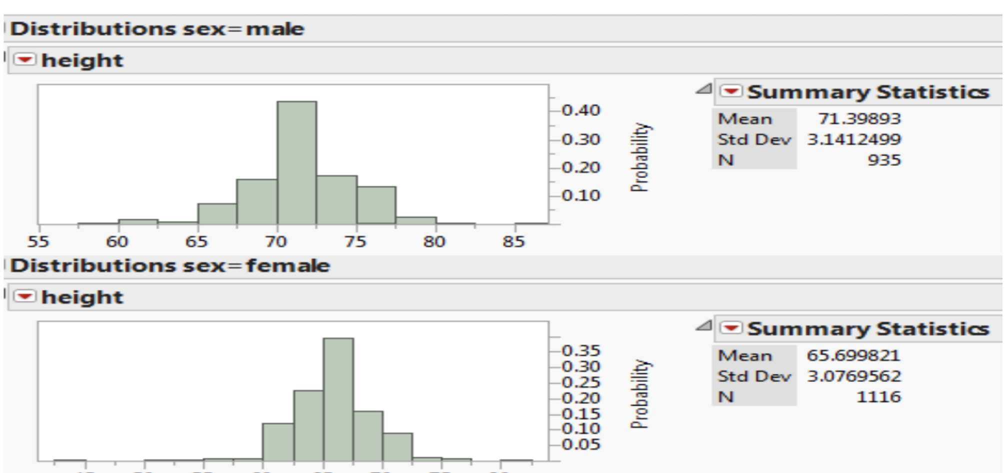
$$df = \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\left(\frac{1}{n_1-1}\right)\left(\frac{s_1^2}{n_1}\right)^2 + \left(\frac{1}{n_2-1}\right)\left(\frac{s_2^2}{n_2}\right)^2}$$

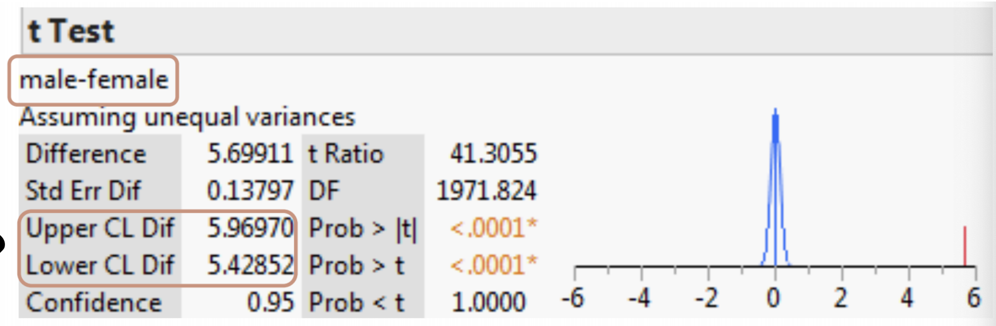
Example *At the beginning of the semester for severl years, students in Stat 101 completed a survey. In this survey, the sex and height (in inches) of the students were recorded.*

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

- *Populations*
  - All males in STAT 101 at the beginning of the semester
  - All females in STAT 101 at the beginning of the semester
- *Samples*
  - 935 male STAT 101 students
  - 1116 female STAT 101 students
- *Parameter:*

  $\mu_1 - \mu_2$ = Population mean height of all male STAT 101 students minus population mean height of all female STAT 101 students (unknown value)
- *Statistic:*
  - $\bar{Y}_1 - \bar{Y}_2$ = sample mean height of 935 male STAT 101 students minus sample mean height of 1116 female STAT 101 students
  - $\bar{Y}_1 - \bar{Y}_2 = 71.3989 - 65.6998 = 5.6991$





t Test

male-female

Assuming unequal variances

Difference	5.69911	t Ratio	41.3055
Std Err Dif	0.13797	DF	1971.824
Upper CL Dif	5.96970	Prob > |t|	<.0001*
Lower CL Dif	5.42852	Prob > t	<.0001*
Confidence	0.95	Prob < t	1.0000

Conditions:

1) can be argued this is random.
2) More than 9350 male students at ISU
   More than 11160 female students at ISU
3) Both distributions are symmetric but not normal, with $n_1, n_2 > 100$ ⇒ Nearly Normal condition met
4) Can be assumed that groups are independent from one another

$$CI: (71.39893 - 65.699821) \pm t^* \sqrt{\frac{3.14^2}{938} + \frac{3.08^2}{1116}}$$

$$5.69911 \pm t^* (0.138005)$$

$df = 19185$

$t^* \approx 1.646$

or use JMP

$CI: (5.42852, 5.96970)$

We are 95% confident that the true difference in heights between male and female Stat 101 students lies within the interval $(5.42852, 5.96970)$ inches.

Hypothesis Test for the Difference in Means

Step 1: Hypotheses

Null Hypothesis

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

$H_0: \mu_1 - \mu_2 = 0$ or $H_0: \mu_1 = \mu_2$

Alternative Hypothesis

- states that this is some type of difference between the two population means
- $H_A: \mu_1 - \mu_2 < 0$ or $H_A: \mu_1 < \mu_2$

$H_A: \mu_1 - \mu_2 > 0$ or $H_A: \mu_1 > \mu_2$

$H_A: \mu_1 - \mu_2 \neq 0$ or $H_A: \mu_1 \neq \mu_2$

Step 2: Assumptions

Check the following conditions:

1. **Randomization condition:** each group must be randomly selected.
2. **10% condition:** each groups' sample size must be less than 10% of the population
3. **Nearly Normal Condition:** see page 2
4. **Independent Groups:** groups must be independent of one another

Step 3: Test Statistic

Then our t-statistic is calculated as follows:

$$t = \frac{\bar{y}_1 - \bar{y}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

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:

$$df = \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\left(\frac{1}{n_1-1}\right)\left(\frac{s_1^2}{n_1}\right)^2 + \left(\frac{1}{n_2-1}\right)\left(\frac{s_2^2}{n_2}\right)^2}$$

We have three different options based on our alternative hypotheses:

$H_a : \mu_1 < \mu_2$

- p-value is the area less than t

$H_a : \mu_1 > \mu_2$

- p-value is greater than t

$H_a : \mu_1 \neq \mu_2$

- p-value is the area less than −|t| plus area greater than |t|

Will use JMP for this!

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:

- parameter
- context
- evidence against the null hypothesis

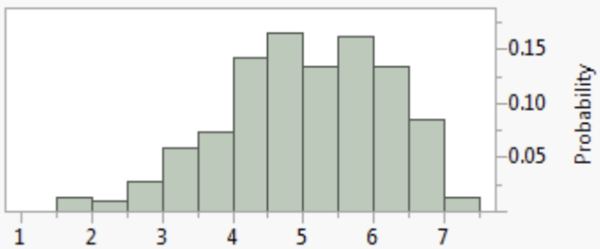
Example *At ISU, several different intro stats courses are offered. Each course is structured according to a particular audience of majors. At the beginning of the Fall 2006 semester, a "Survey of Attitudes Toward Statistics" was administered to students in Stat 101 and Stat 226. One of the components 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

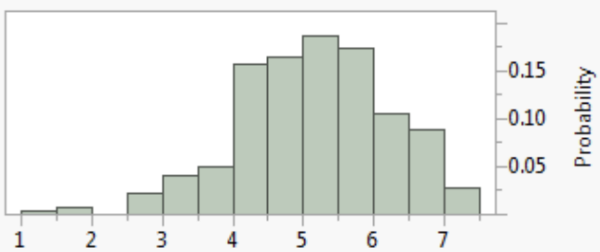


Summary Statistics

Mean	4.9556401
Std Dev	1.1362525
N	263

Distributions Course=Stat 226

Pre-CC



Summary Statistics

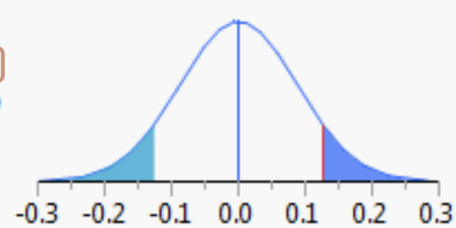
Mean	5.0808081
Std Dev	1.0491075
N	396

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



Step 1: $H_0: \mu_1 = \mu_2$

$H_A: \mu_1 > \mu_2$

Step 2: See page 4.

Step 3: $t = 1.4275$

Step 4: p-value = 0.0770

Step 5: weak evidence against null

Step 6: we have weak evidence to suggest that Stat 226 students have a higher mean cognitive competance than Stat 101 students.