EDP308: STATISTICAL LITERACY

The University of Texas at Austin, Fall 2020

RAZ: Rebecca A. Zárate, MA

Overview

- One Sample t-test Review
- Independent Samples t-test
  - Assumptions
- Sampling Distribution of Difference in Means
  - Hypotheses Formulation
    - The Null: Zero Enough
  - Pooled Standard Deviation (S_{pooled})
  - Pooled Standard Error (SE_{pooled})
  - Degrees of Freedom (N-2)
- Confidence Intervals for Mean Differences
- Cohen's d – Effect Size Measure
- Independent Samples t-test in R
  - Summary Data
  - Full Datasets

One Sample t-test Review

- Situations when:
  - You are comparing your *one* sample mean to a **known population mean**
  - σ **unknown**
    - If it was known, you'd use a z-test

But what if you want to compare two samples means you collected to each other?

Independent Samples t-test

Independent Samples t-test

- Now, we are going to compare two sample means to each other, rather than to a known population mean
  - Perhaps there is no known population mean for the groups of interest
    - Ex. What is the average number of minutes/week UT students exercise compared to students from St. Edwards?
  - Or maybe there is a new thing we are measuring (or a new scale) that doesn't have established population means
    - Ex. Is there a difference in fidget scores on the Fidget Assessment Battery-II between people diagnosed with ADHD and those not diagnoses with ADHD?

Assumptions

- Independence of Observations
  - There is no relationship between the people in the groups or across the groups
- The Data are Normally Distributed
  - Follow a normal distribution, no skews or crazy outliers
- Homogeneity of Variance
  - Both groups have similar of variances within their groups
    - Ex. If the variance in Group A is 25, the variance in Groups B should be fairly close
    - The ratio of the larger over the smaller variance should not exceed 1.5

Sampling Distribution of Differences in Means

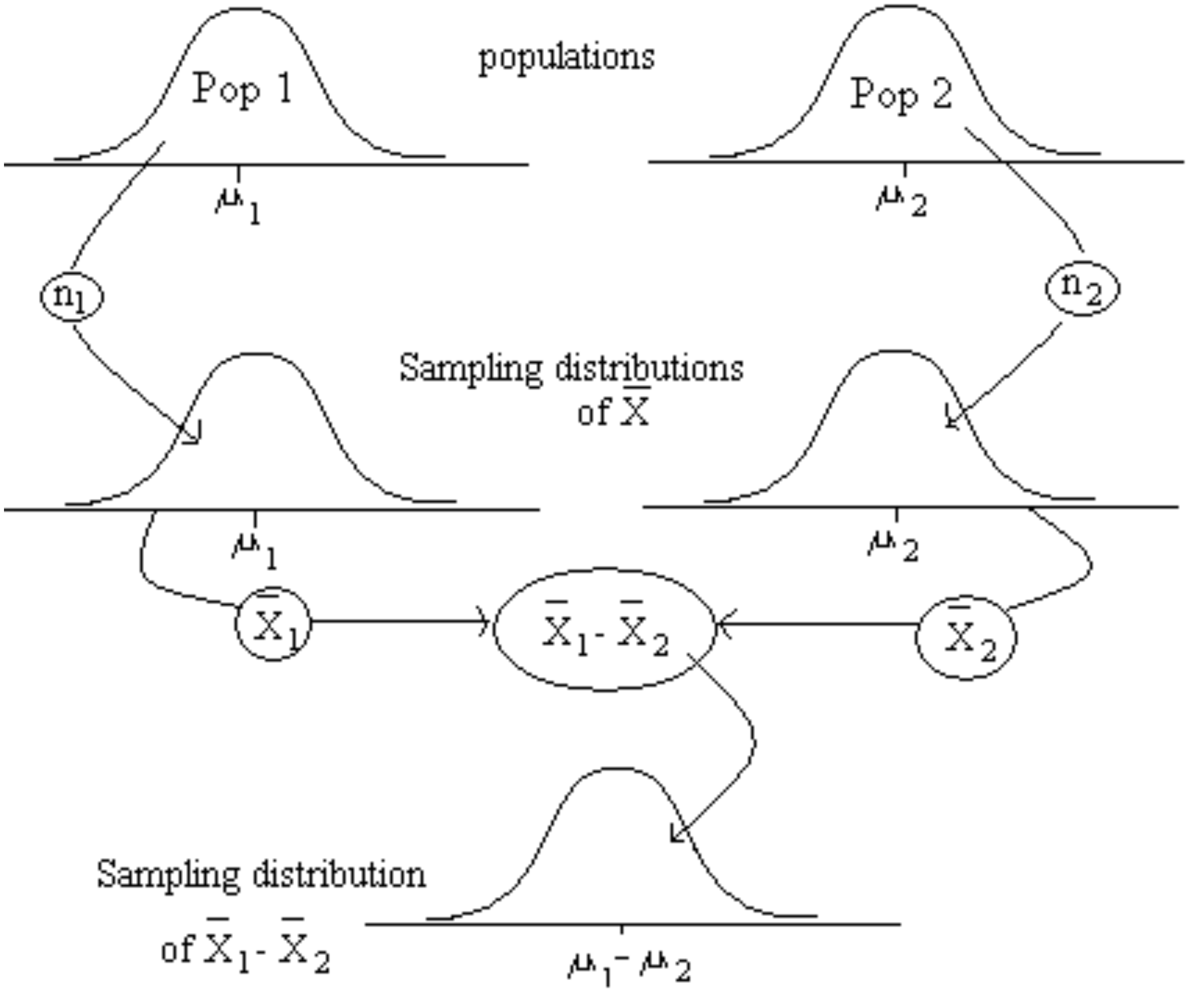
- Now, rather than comparing our one sample to the one known population, now we will test the difference between our two sample means
  - If we did this again and again, drawing samples, and taking the difference, $\bar{x}_A - \bar{x}_B$ the difference between them, over repeated samples will
    - Follow a normal distribution (thanks to the central limit theorem)
    - Have their own joint standard error
- Independent samples t-test are typically two-tailed, but one tailed are possible if there is already some evidence or theory to show that there is directionality

Independent Samples Visualized

There are some populations out there…

…that we take our samples from to then test the difference between them and…

…if we sampled and calculated those differences again and again, the average difference between the groups will be a normally distributed sampling distribution



Independent Groups

- The two groups are categorized by some sort of categorical (nominal) variable
  - Ex. Female vs. Male, Mountain People vs. Beach People, Side Sleepers vs. Back Sleepers, etc.
- The two groups you choose to compare can be anything, really, but…
- They should only differ on that one grouping quality, all else should be fairly equal
  - Ex. Female vs. Male, the participants should be of similar SES, age, etc.
    - This of course depends heavily on what the research question is and what is available to you

Hypothesis Formulation - Differences in Means

In this class, and for most comparisons between groups, the null hypothesis assumes a true difference of 0, so *no difference between the groups*. That is, we usually assume:

$$H_0:\ \mu_A - \mu_B = 0$$

$$H_0:\ \mu_{A} = \mu_{B}$$

$$H_1:\ \mu_{A} \neq \mu_{B}$$

Two sided.

$$H_0:\ \mu_{A} \geq \mu_{B}$$

$$H_1:\ \mu_{A} < \mu_{B}$$

One sided.
A is less than B

$$H_0:\ \mu_{A} \leq \mu_{B}$$

$$H_1:\ \mu_{A} > \mu_{B}$$

One sided.
A is greater than B

Independent Samples t-test

Group A Mean – Group B Mean

Null Hypothesis…

$$t_{stat} = \frac{(\bar{x}_A - \bar{x}_B) - (\mu_A - \mu_B)}{\sqrt{\frac{s_A^2}{n_A} + \frac{s_B^2}{n_B}}}$$

Standard Deviation of Each Group

Sample Size of Each Group

Pooled Standard Error

Independent Samples t-test NULL

$$t_{stat} = \frac{(\bar{x}_A - \bar{x}_B) - (\mu_A - \mu_B)}{\sqrt{\frac{s_A^2}{n_A} + \frac{s_B^2}{n_B}}}$$

What is the null hypothesis with an Independent Samples t-test?

Independent Samples t-test NULL

- The NULL hypothesis for an independent samples t-test is: “There is no difference between the means.”

$$H_0\colon \mu_A - \mu_B = 0$$

So this will always be equal to zero…

$$t_{stat} = \frac{(\bar{x}_A - \bar{x}_B) - (\mu_A - \mu_B)}{\sqrt{\frac{s_A^2}{n_A} + \frac{s_B^2}{n_B}}}$$

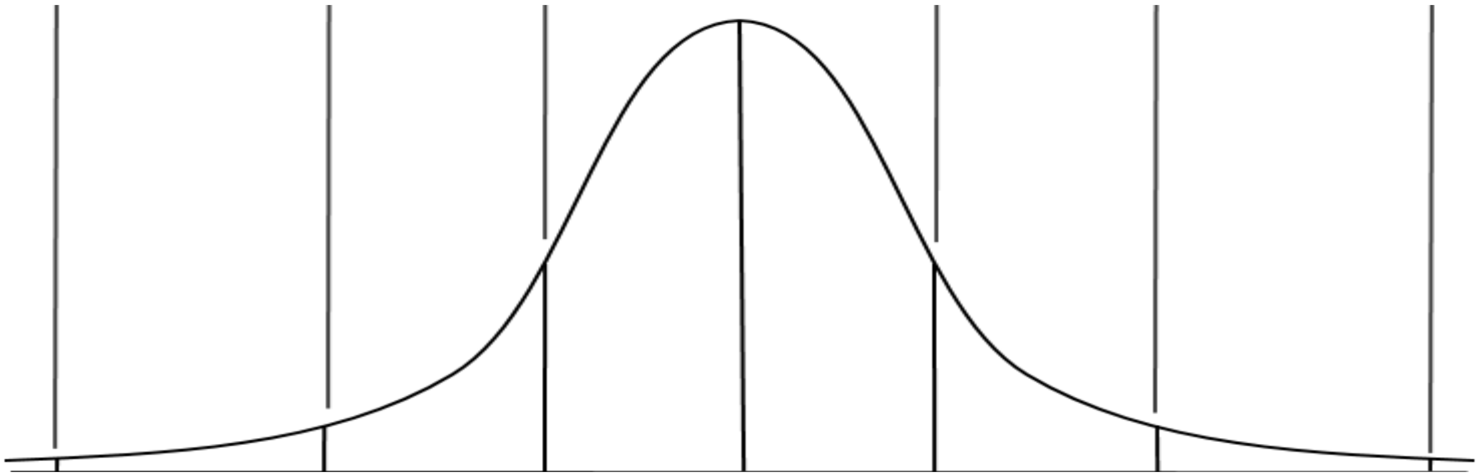
$$t_{stat} = \frac{(\bar{X}_A - \bar{X}_B) - 0}{\sqrt{\frac{s_A^2}{n_A} + \frac{s_B^2}{n_B}}}$$

And the equation is actually just this…

$$t_{stat} = \frac{(\bar{x}_A - \bar{x}_B)}{\sqrt{\frac{s_A^2}{n_A} + \frac{s_B^2}{n_B}}}$$

NULL is Zero

SAMPLING distribution of $(\bar{x}_A - \bar{x}_B)$



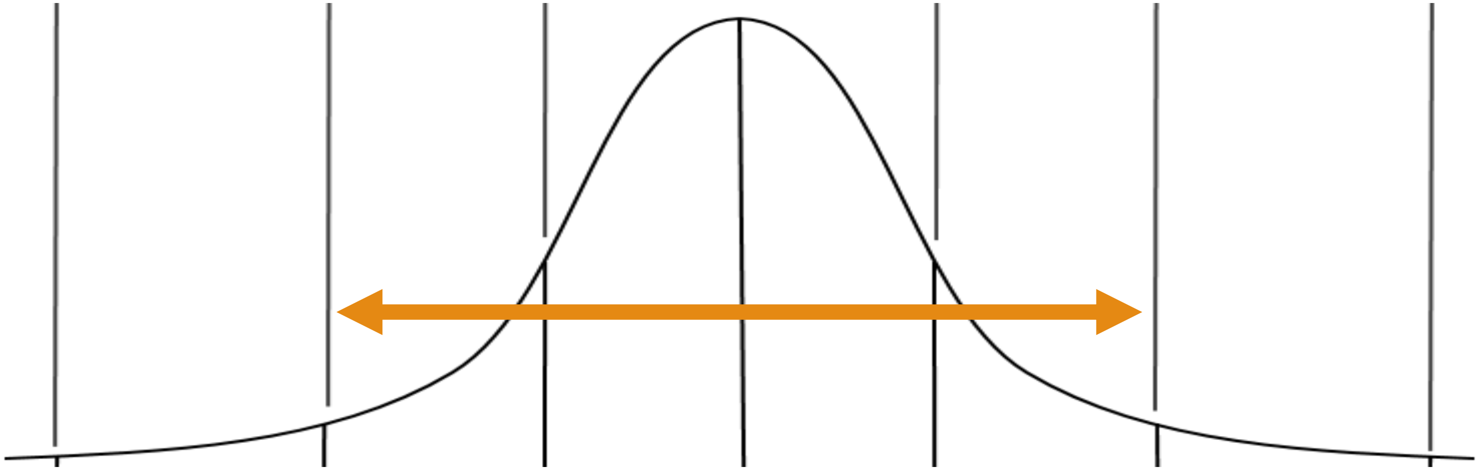
$(\mu_{\bar{x}_A} - \mu_{\bar{x}_B}) = \mu_A - \mu_B = 0$

If the NULL were true, would you expect that every sample of $(\bar{x}_A - \bar{x}_B)$ equal exactly 0?

Why or why not?

NULL is Zero

- Because of random error and chance, the difference will not always be exactly zero.
  - Just like before, we want to know a range of reasonable values that we can be 95% confident with
    - A range of "zero enough"



$$(\mu_{\bar{x}_A} - \mu_{\bar{x}_B}) = \mu_A - \mu_B = 0$$

SAMPLING distribution of $(\bar{x}_A - \bar{x}_B)$

Pooled Standard Deviation $\mathrm{S_p}$

We also rarely know σ_A^2 and σ_B^2, so we'll have to approximate these with s_A^2 and s_B^2. Then we can combine the two samples' standard deviation together to give us one measure of variation, pooled standard deviation, which we can use for calculating Cohen's d.

$$\mathrm{S_p} = Pooled\ Standard\ Deviation$$

$$= \sqrt{\frac{(n_A - 1)s_A^2 + (n_B - 1)s_B^2}{n_A + n_B - 2}}$$

Pooled Standard Error $\mathrm{SE_p}$

We also need to calculate a pooled standard error for the sampling distribution, i.e. the standard deviation for the sampling distribution.

$\mathrm{SE_p} =$

$$Pooled\ Standard\ Error = \sqrt{\frac{\sigma_A^2}{n_A} + \frac{\sigma_B^2}{n_B}} \approx \sqrt{\frac{s_A^2}{n_A} + \frac{s_B^2}{n_B}}$$

Degrees of Freedom

- □ Because we have two sample groups now, we will have to update our degrees of freedom to:

$$df = (n_A - 1) + (n_B - 1)$$
$$df = n_A + n_B - 2$$
$$df = N - 2$$

All of these equations will give you the same thing

$$\mathrm{N} = (n_A + n_B)$$
$$\mathrm{N} = \text{total sample size}$$

Practice

Try it. Diet.

- Let's say we want to find out which of two diets helps you lose more weight: a low fat diet (LF) or a low carb diet (LC)
    - We form two groups (the independent variables: LF vs LC) by randomly assigning dieters to either the LF or LC condition
  - There were 8 LF dieters and 14 LC dieters
- After 3 months, the LF dieters lost an average of 8.2 lbs with a standard deviation of 2.1 lbs. The LC dieters lost an average of 5.3 lbs with a standard deviation of 3.2 lbs
  - Test at the .05 level of significance.

Step 1: Hypotheses

Step 1:

$$H_0\colon\ \mu_{LF} - \mu_{LC} = 0$$
$$H_1\colon \mu_{LF} - \mu_{LC} \neq 0$$

H_0: Low fat and low carb dieters have the same average amount of weight loss. (There is no difference in the means).

H_1: Low fat and low carb dieters have different average amounts of weight loss. (There is a difference in the means).

This is a two-tailed test because we make no assertions as to which diet may be more effective.

Step 2 and 3: Significance and Test

Step 2:

$$\alpha = .05$$

Step 3:

$$t_{stat} = \frac{(\bar{x}_{LF} - \bar{x}_{LC}) - (\mu_{LF} - \mu_{LC})}{\sqrt{\frac{s_{LF}^2}{n_{LF}} + \frac{s_{LC}^2}{n_{LC}}}} \rightarrow t_{stat} = \frac{\bar{x}_{LF} - \bar{x}_{LC}}{\sqrt{\frac{s_{LF}^2}{n_{LF}} + \frac{s_{LC}^2}{n_{LC}}}}$$

Remember this just equals 0

Step 4: Get Critical Value(s)

Step 4:

$$\alpha = .05$$

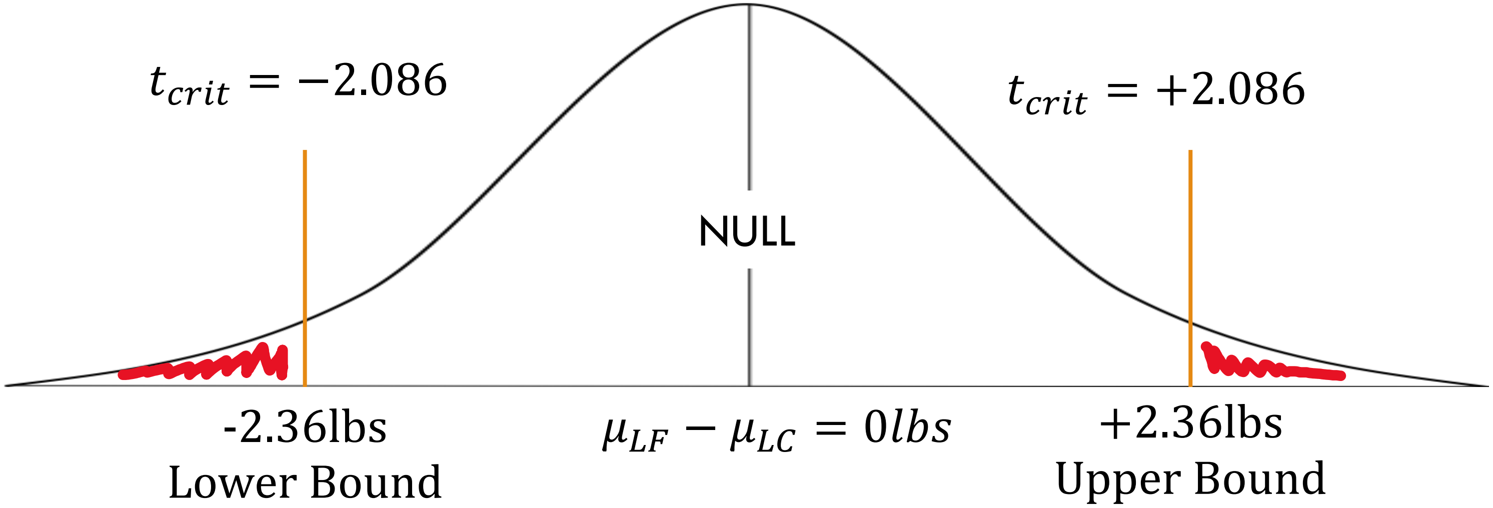
$$n_{LF} = 8 \text{ and } n_{LC} = 14 \text{ ,}$$
$$df = (8 + 14) - 2 = 20$$

Our test is a two-tailed test, meaning that our two critical values should cut off tails with probabilities of .025 each.

The t-critical values that satisfies $df = 20$ and $t_{.025}$ are: $t_{crit} = \pm 2.086$.

Visualizing Critical Cut Offs

- □ If our test statistic is greater or less than the critical value of ± 2.086, or more than 2.36 pounds of difference, then we can be confident that the true difference between the samples is not equal to zero



$$\text{Upper and Lower Bounds} = 0 \pm 2.086 \ * \ (1.13)$$

Step 5: Compute the Test Statistic

Step 5:

$$t_{stat} = \frac{\bar{x}_{LF} - \bar{x}_{LC}}{\sqrt{\frac{s_{LF}^2}{n_{LF}} + \frac{s_{LC}^2}{n_{LC}}}} = \frac{8.2 - 5.3}{\sqrt{\frac{2.1^2}{8} + \frac{3.2^2}{14}}} \approx \frac{2.9}{1.133} \approx 2.56$$

$$t_{stat} = 2.56$$

Step 5: Compute the Test Statistic

Step 5:

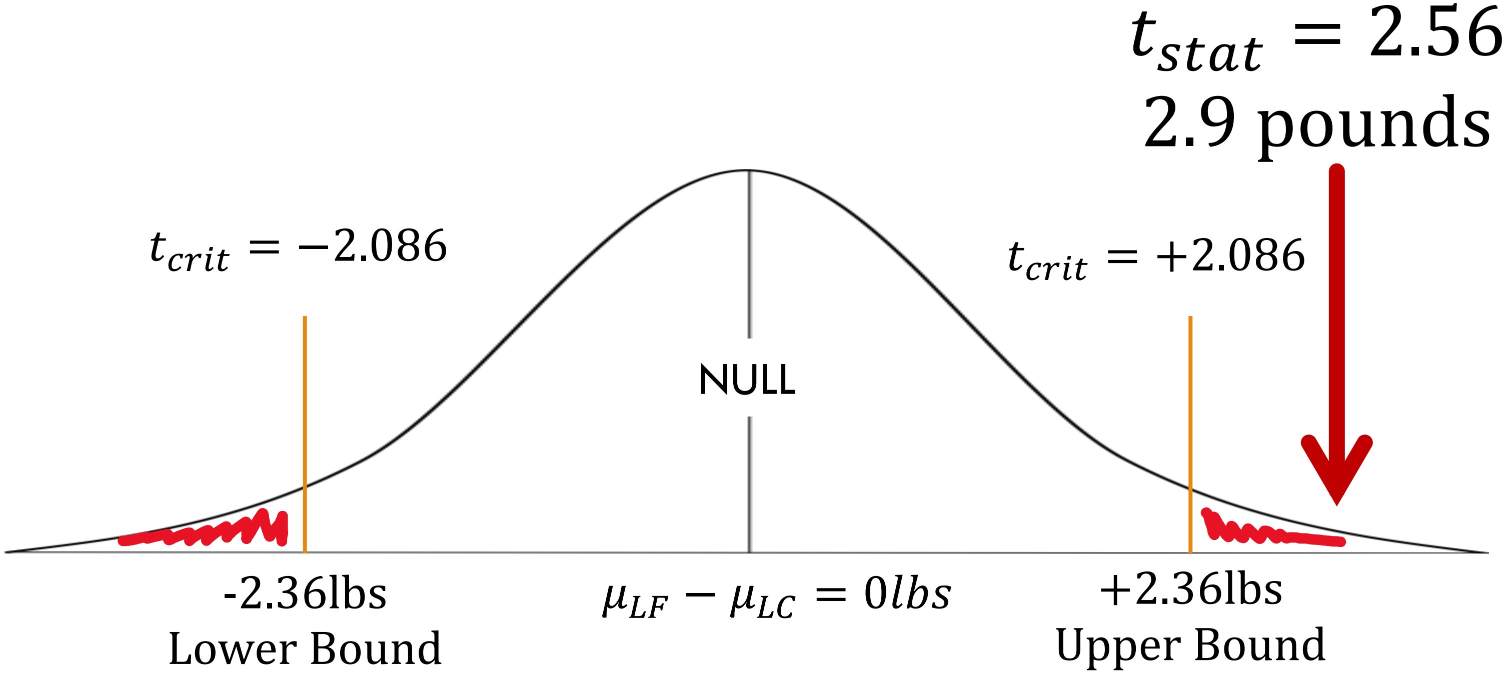
$$t_{stat} = \frac{\bar{x}_{LF} - \bar{x}_{LC}}{\sqrt{\frac{s_{LF}^2}{n_{LF}} + \frac{s_{LC}^2}{n_{LC}}}} = \frac{8.2 - 5.3}{\sqrt{\frac{2.1^2}{8} + \frac{3.2^2}{14}}} \approx \frac{2.9}{1.133} \approx 2.56$$

$$t_{stat} = 2.56$$

Note: You MUST do the squaring and dividing FIRST, then take the square root of everything LAST to get the **Pooled Standard Error**

Visualizing Critical Cut Offs

Which of the two diet methods is more effective?



Step 6: Draw Conclusions

Step 6:

Our $t_{stat} = 2.56$, and our $t_{crit} = \pm 2.086$.
Our t_{stat} is past our t_{crit}, so we reject H_0.

Reject the NULL hypothesis. The two diet methods are not equivalent. People lose more weight with the LF (low fat) diet than the LC (low carb) diet.

$$\mathrm{t}(20) = +2.56, \mathrm{p} < 0.05$$

$$\mathrm{t}(20) = +2.56, \mathrm{p} = 0.02$$

Confidence Interval for Differences

Construct a 95% confidence interval for the difference in means. Are your results consistent with the hypothesis test ($\alpha = .05$)?

Confidence Interval for Differences

Confidence interval for a single population mean (μ):

$$CI = \bar{x} \pm t * \frac{s}{\sqrt{n}};$$
$$df = n - 1$$

Confidence interval for the difference in means for two populations ($\mu_A - \mu_B$):

$$CI = (\bar{x}_A - \bar{x}_B) \pm t * SE_{Pooled}$$
$$df = n_A + n_B - 2$$

Confidence Interval for Differences

Information:

$$\bar{X}_{LF} = 8.2, s_{LF} = 2.1, n_{LF} = 8$$
$$\bar{X}_{LC} = 5.3, s_{LC} = 3.2, n_{LC} = 14$$

$$df = 8 + 14 - 2 = 20 \rightarrow t_{.025} = \pm 2.086$$

Calculation:

$$CI = (\bar{X}_A - \bar{X}_B) \pm t * \sqrt{\frac{s_A^2}{n_A} + \frac{s_B^2}{n_B}}$$

Confidence Interval for Differences

Calculation:

$$CI = (\bar{X}_A - \bar{X}_B) \pm t * \sqrt{\frac{s_A^2}{n_A} + \frac{s_B^2}{n_B}}$$

$$95\%\ CI = (8.2 - 5.3) \pm 2.086 * \sqrt{\frac{2.1^2}{8} + \frac{3.2^2}{14}}$$

$$95\%\ CI = 2.9 \pm 2.086 * 1.133$$

$$95\%\ CI = 2.9 \pm 2.36$$

$$95\%\ CI = [.54, 5.26]$$

Confidence Intervals and Nulls

- When constructing confidence intervals for a difference in means, the null hypothesis is usually $H_0\colon\ \mu_A - \mu_B = 0$.
  - So, we can compute our confidence interval for the difference in means, and check to see whether the interval contains 0.
- When the interval contains 0, it is plausible that $\mu_A - \mu_B = 0$, and this is consistent with *failing to reject* H_0.
- When the interval does *not* contain 0, it is unlikely that $\mu_A - \mu_B = 0$, and this is consistent with *rejecting* H_0.

Confidence Intervals and Nulls

Interpretation:

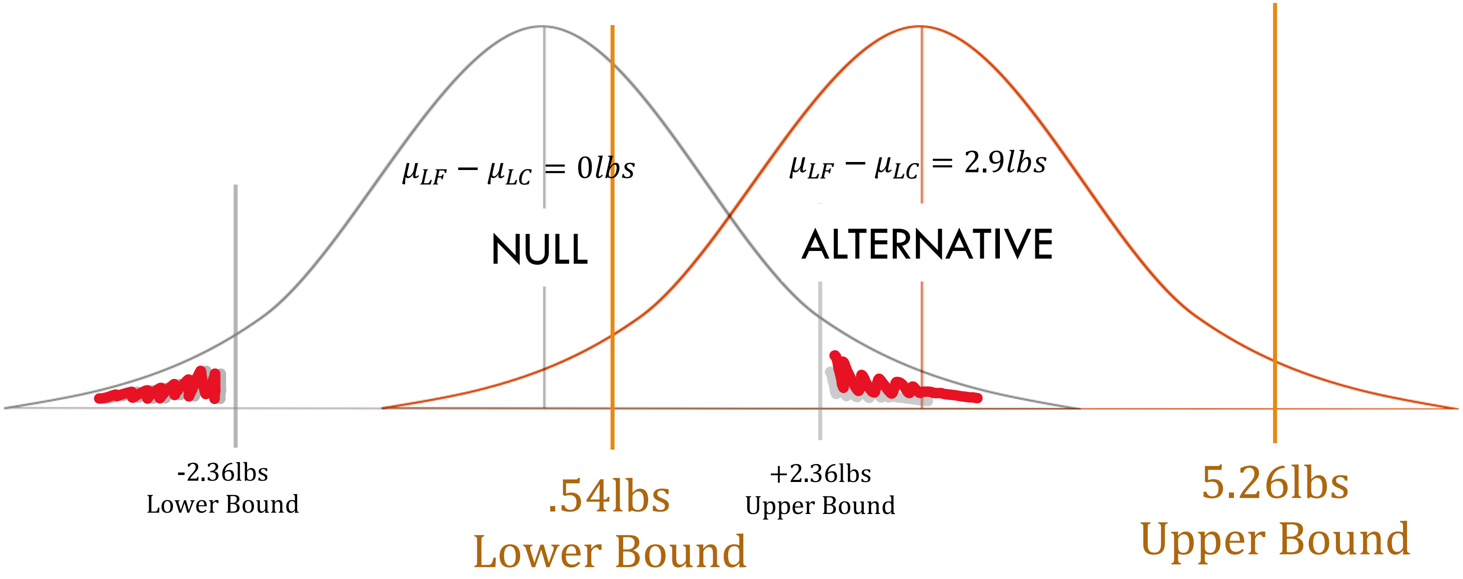
$$95\%\ CI = [.54, 5.26]$$

"If we take repeated samples (of the given sizes) and compute a 95% confidence interval for the difference in means each time, approximately 95% of the intervals would contain the true average difference in mean weight loss between the low fat groups and the low carb groups. Our 95% CI suggests people lose at least .54 lbs and up to 5.26 lbs more weight in the low fat group compared to the low carb group."

Note: This interval does not contain the null hypothesized value, 0, so it seems that 0 is not a plausible value for $\mu_{LF} - \mu_{LC}$. (i.e. The true mean difference does not seem to be 0). This is consistent with our hypothesis test results, in which we reject the null hypothesis that $H_0\colon \mu_{LF} - \mu_{LC} = 0$.

Reject the NULL Example

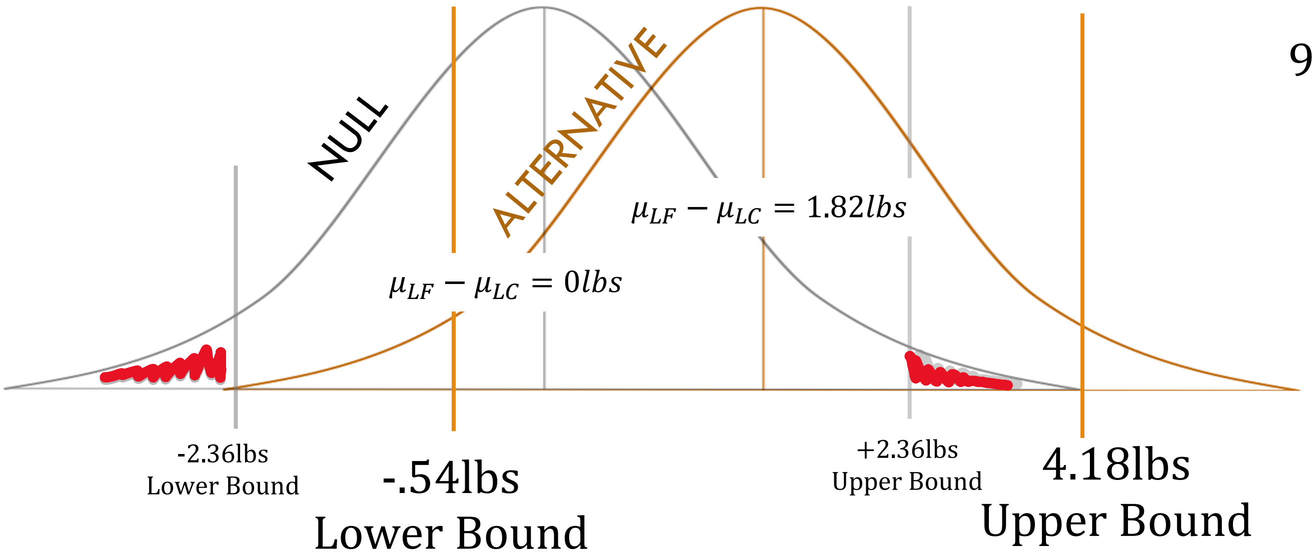
- Notice how the MEAN of the NULL is NOT in the 95% confidence interval for our difference in the means.
  - In light grey on the left is the NULL hypothesis and its range of “zero enough”
  - On the right is our 95% confidence interval for our difference in the means



Fail to Reject NULL Example

I modified the last example to show what an insignificant case would look like.

- Notice how the MEAN of the NULL is in the 95% confidence interval for our difference in the means
  - A difference of zero (the NULL) is a reasonable mean difference, so you cannot reject the null.



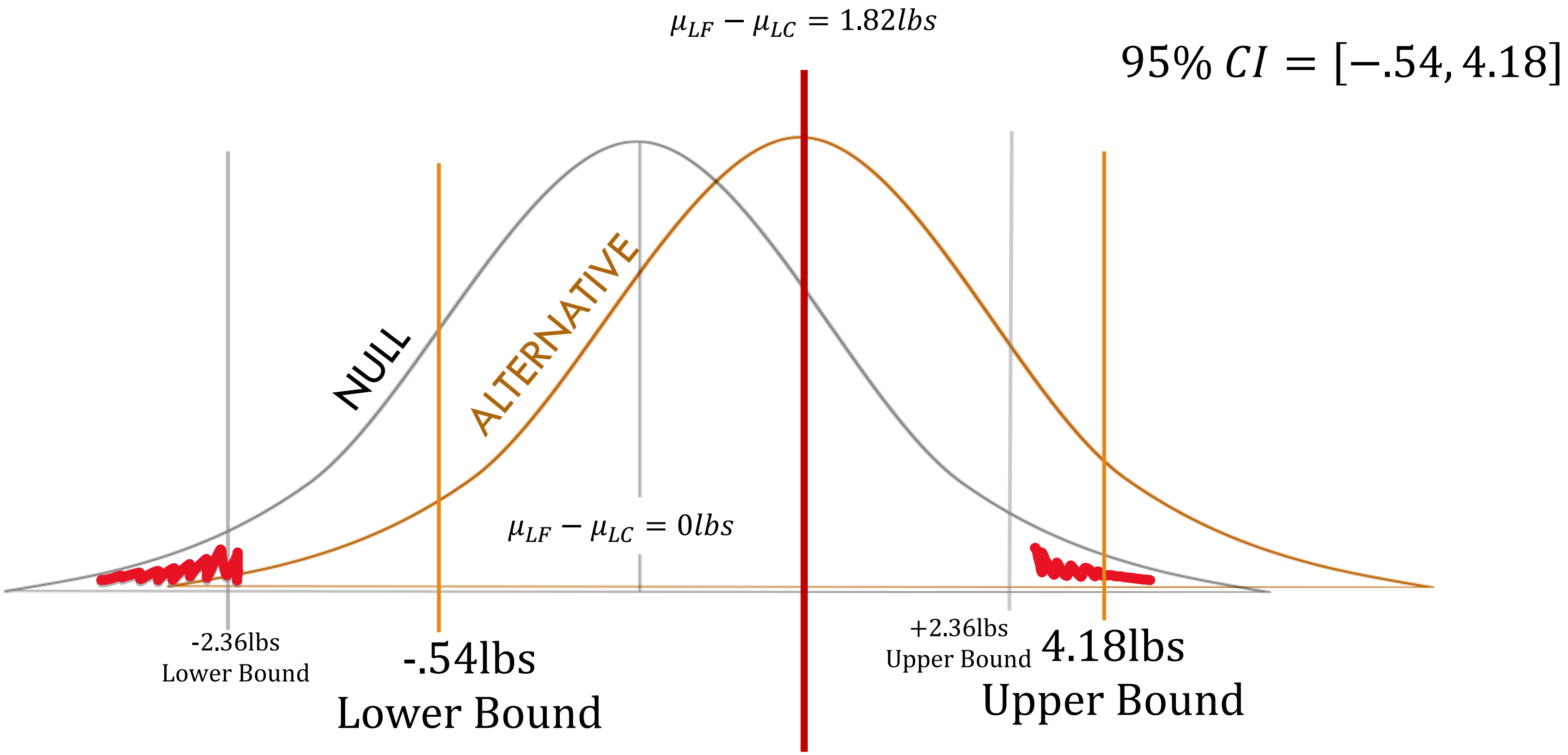
$$95\%\ CI = [-.54, 4.18]$$

Contains zero, meaning the difference could just be zero…

Fail to Reject NULL Example

I modified the last example to show what an insignificant case would look like.

- The red line (which represents the sample mean difference) is inside of the range of “zero enough”



Independent Sample t-tests and Effect Size

Effect Size - Cohen d

- We can also calculate the effect size, Cohen’s d, to get an idea of the practical significance of the results.

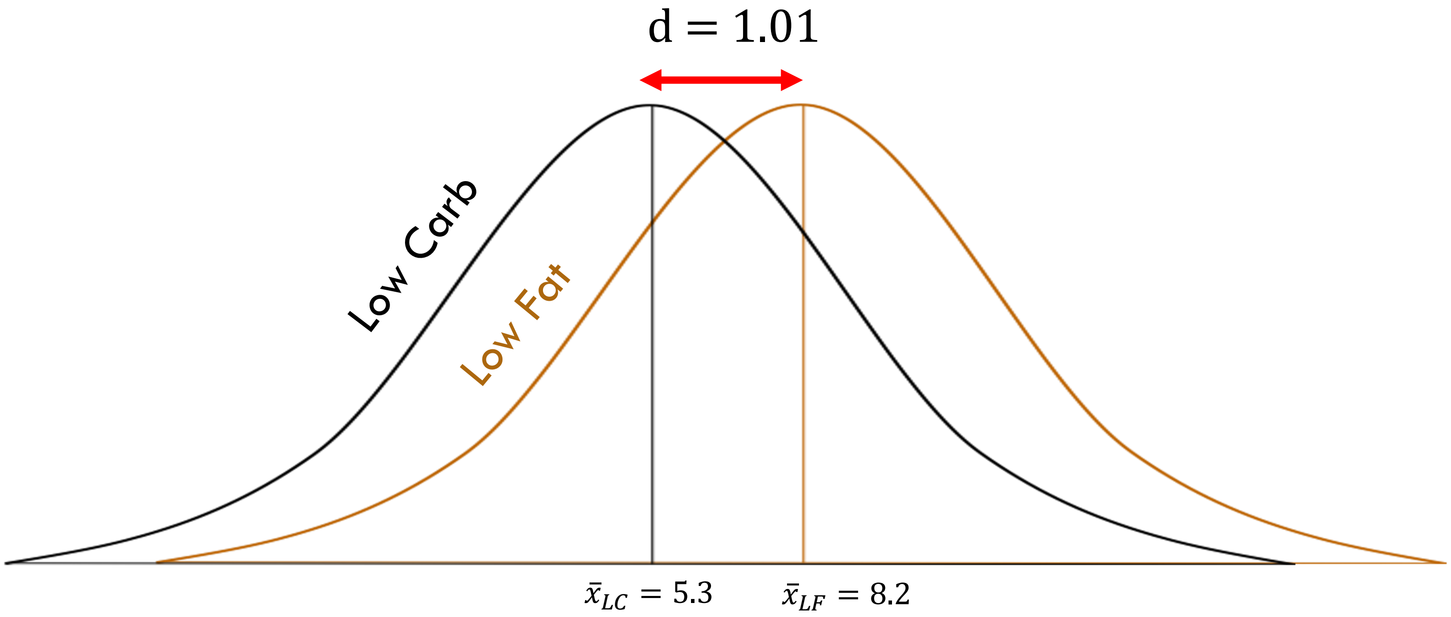
$$\mathrm{S_p} = Pooled\ Standard\ Deviation = \sqrt{\frac{(n_A - 1)s_A^2 + (n_B - 1)s_B^2}{n_A + n_B - 2}}$$

$$= \sqrt{\frac{(8 - 1)2.1_A^2\ +\ (14_B - 1)3.2_B^2}{8 + 14_B - 2}} = 2.86$$

$$Cohen's\ d = \frac{\bar{x}_{LF} - \bar{x}_{LC}}{\mathrm{S_p}} = \frac{8.2 - 5.3}{2.86} \approx \frac{2.9}{2.86} \approx 1.01$$

Effect Size - Cohen d

- For these samples, the effect size is $\mathrm{d} = 1.01$
- The Low Fat dieters lose, on average, a whole standard deviation more than the Low Carb group.



One Sample and Independent t-test

- □ Though they look different, the essence of the t-tests is the same…

$$\text{t-value} = \frac{\text{Difference between means}}{\text{Variation (SE) of the group(s)}}$$

$$t = \frac{\bar{x} - \mu}{s/\sqrt{n}}$$

One Sample t-test

$$t = \frac{(\bar{X}_A - \bar{X}_B) - (\mu_A - \mu_B)}{\sqrt{\frac{s_A^2}{n_A} + \frac{s_B^2}{n_B}}}$$

Independent Samples t-test

Up Next…

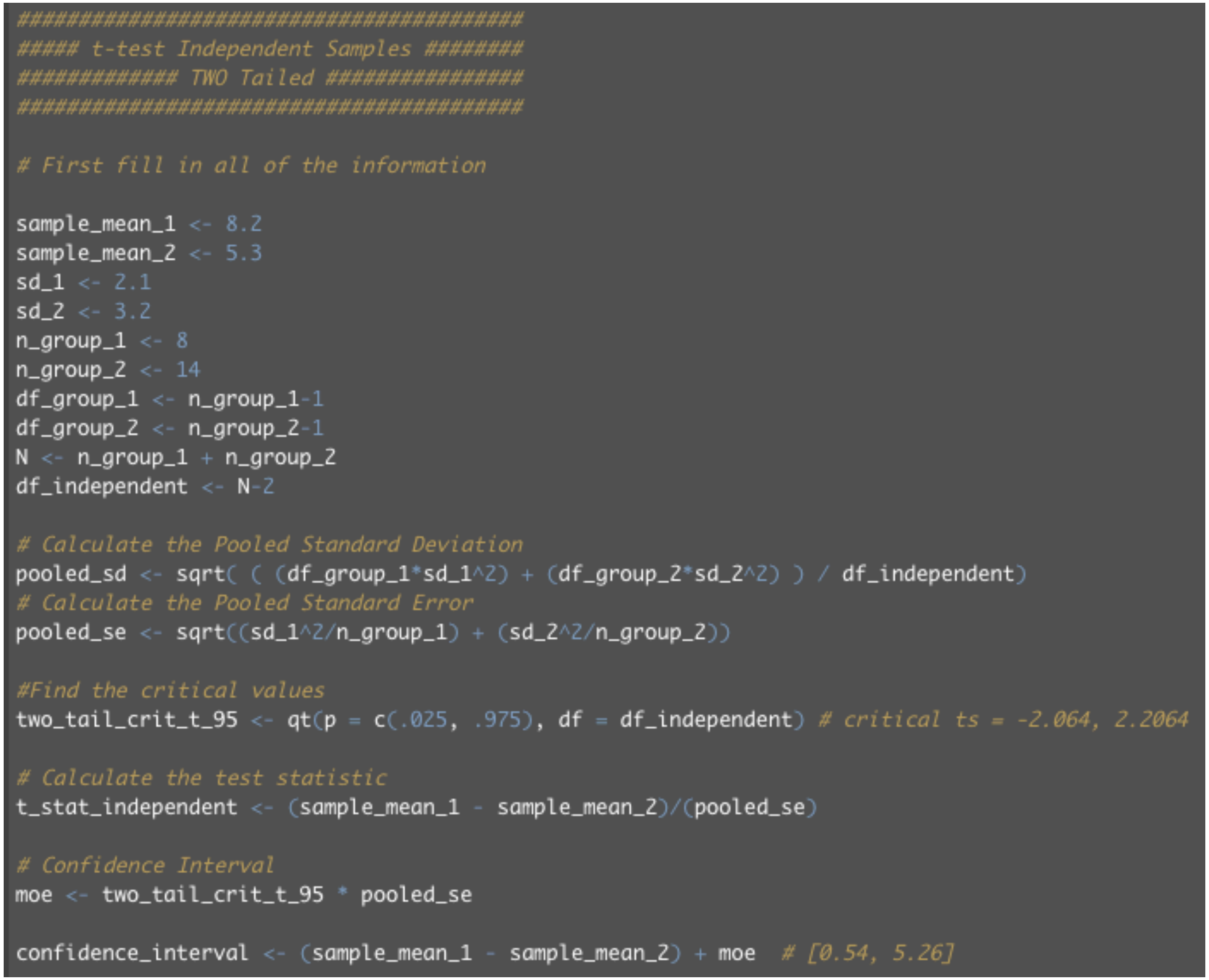
- Our last type of t-test we will cover…

Dependent Samples t-tests

Independent Samples t-test in R

Independent Samples t-test in R (Summary Statistics)

- You can calculate the t-statistic and confidence interval using summary data for an independent samples t-test



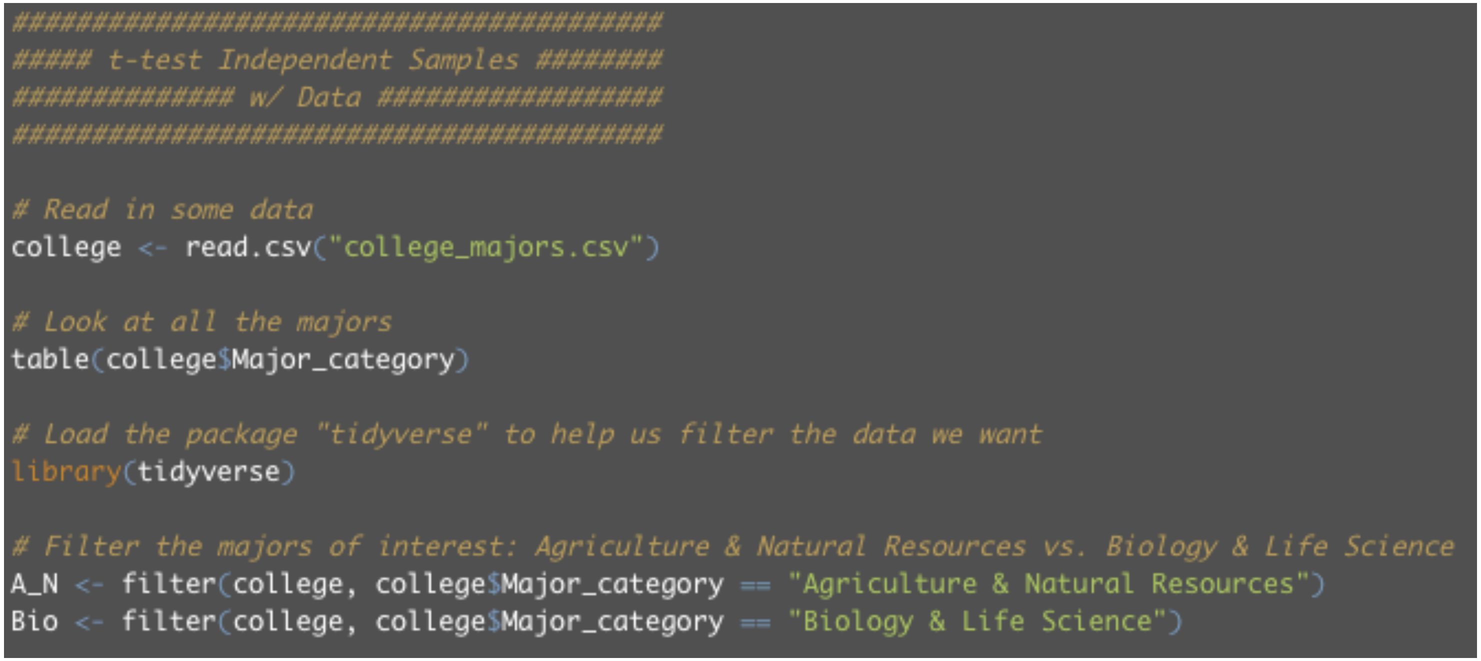
```
#########################################
##### t-test Independent Samples ########
############# TWO Tailed ################
#########################################

# First fill in all of the information

sample_mean_1 <- 8.2
sample_mean_2 <- 5.3
sd_1 <- 2.1
sd_2 <- 3.2
n_group_1 <- 8
n_group_2 <- 14
df_group_1 <- n_group_1-1
df_group_2 <- n_group_2-1
N <- n_group_1 + n_group_2
df_independent <- N-2

# Calculate the Pooled Standard Deviation
pooled_sd <- sqrt( ( (df_group_1*sd_1^2) + (df_group_2*sd_2^2) ) / df_independent)
# Calculate the Pooled Standard Error
pooled_se <- sqrt((sd_1^2/n_group_1) + (sd_2^2/n_group_2))

#Find the critical values
two_tail_crit_t_95 <- qt(p = c(.025, .975), df = df_independent) # critical ts = -2.064, 2.2064

# Calculate the test statistic
t_stat_independent <- (sample_mean_1 - sample_mean_2)/(pooled_se)

# Confidence Interval
moe <- two_tail_crit_t_95 * pooled_se

confidence_interval <- (sample_mean_1 - sample_mean_2) + moe  # [0.54, 5.26]
```

Independent Samples t-test in R (Data)

- As all of you know, there are a number of different Majors offered at universities. Different majors may result in jobs that have different levels of pay. Let’s compare a few different Major Categories and see if they are statistically significantly different from each other on their Median level of pay.
- Let’s start with Agriculture & Natural Resources vs. Biology & Life Science.

Reading In and Filtering Data

□ We will need to use a package called “tidyverse” to help us “filter( )” through some of the data, since right now we only want the Agriculture & Natural Resources and the Biology & Life Science data.



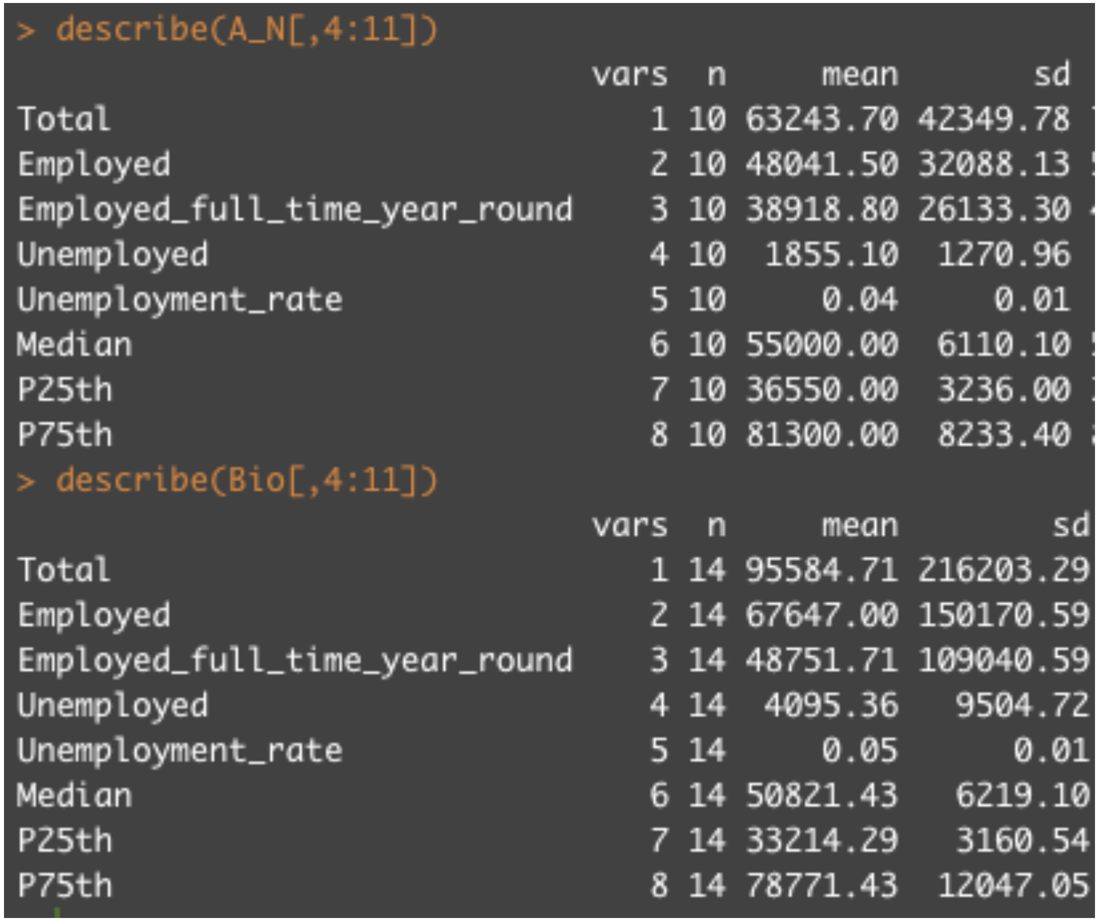
```
#########################################
##### t-test Independent Samples ########
############## w/ Data ##################
#########################################

# Read in some data
college <- read.csv("college_majors.csv")

# Look at all the majors
table(college$Major_category)

# Load the package "tidyverse" to help us filter the data we want
library(tidyverse)

# Filter the majors of interest: Agriculture & Natural Resources vs. Biology & Life Science
A_N <- filter(college, college$Major_category == "Agriculture & Natural Resources")
Bio <- filter(college, college$Major_category == "Biology & Life Science")
```

Agriculture vs. Biology t-test

- Do Agriculture & Natural Resources and Biology & Life Science differ significantly from each other in median income?
  - Test at .05 level.
- H_0: Agriculture & Natural Resources and Biology & Life Science do **NOT** differ significantly from each other in median income (There is no difference in median income).
- H_1: Agriculture & Natural Resources and Biology & Life Science **DO** differ significantly from each other in median income (There is no difference in median income).

Descriptive Statistics

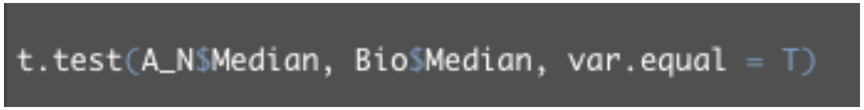
- First, let’s look at some descriptive statistics.
  - “Total” is the number of people who graduated with that major.
  - “Median” is the median income of that major.
- Does it look like Agriculture and Bio differ *significantly* in their median income?



```
> describe(A_N[,4:11])
                              vars  n     mean       sd
Total                            1 10 63243.70 42349.78
Employed                         2 10 48041.50 32088.13
Employed_full_time_year_round    3 10 38918.80 26133.30
Unemployed                       4 10  1855.10  1270.96
Unemployment_rate                5 10     0.04     0.01
Median                           6 10 55000.00  6110.10
P25th                            7 10 36550.00  3236.00
P75th                            8 10 81300.00  8233.40
> describe(Bio[,4:11])
                              vars  n     mean        sd
Total                            1 14 95584.71 216203.29
Employed                         2 14 67647.00 150170.59
Employed_full_time_year_round    3 14 48751.71 109040.59
Unemployed                       4 14  4095.36   9504.72
Unemployment_rate                5 14     0.05      0.01
Median                           6 14 50821.43   6219.10
P25th                            7 14 33214.29   3160.54
P75th                            8 14 78771.43  12047.05
```

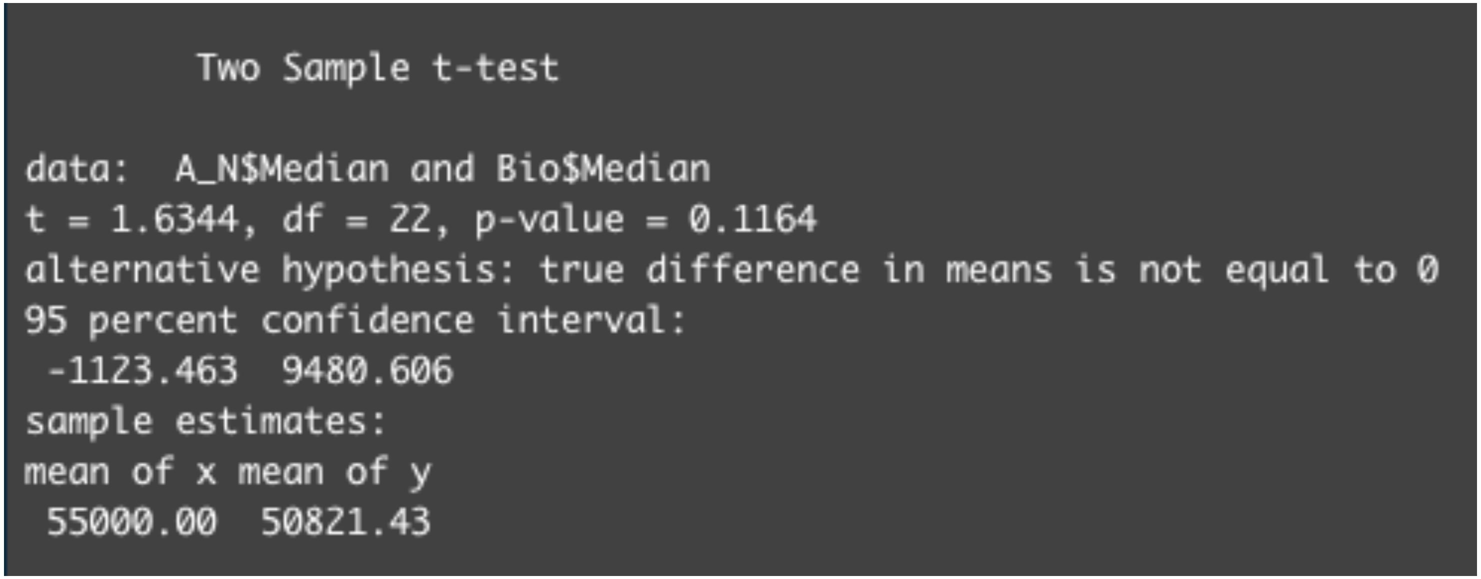
$55,000.00 – $50,821.43 = $4,178.57

Agriculture vs. Biology t-test, R Output



```
t.test(A_N$Median, Bio$Median, var.equal = T)
```

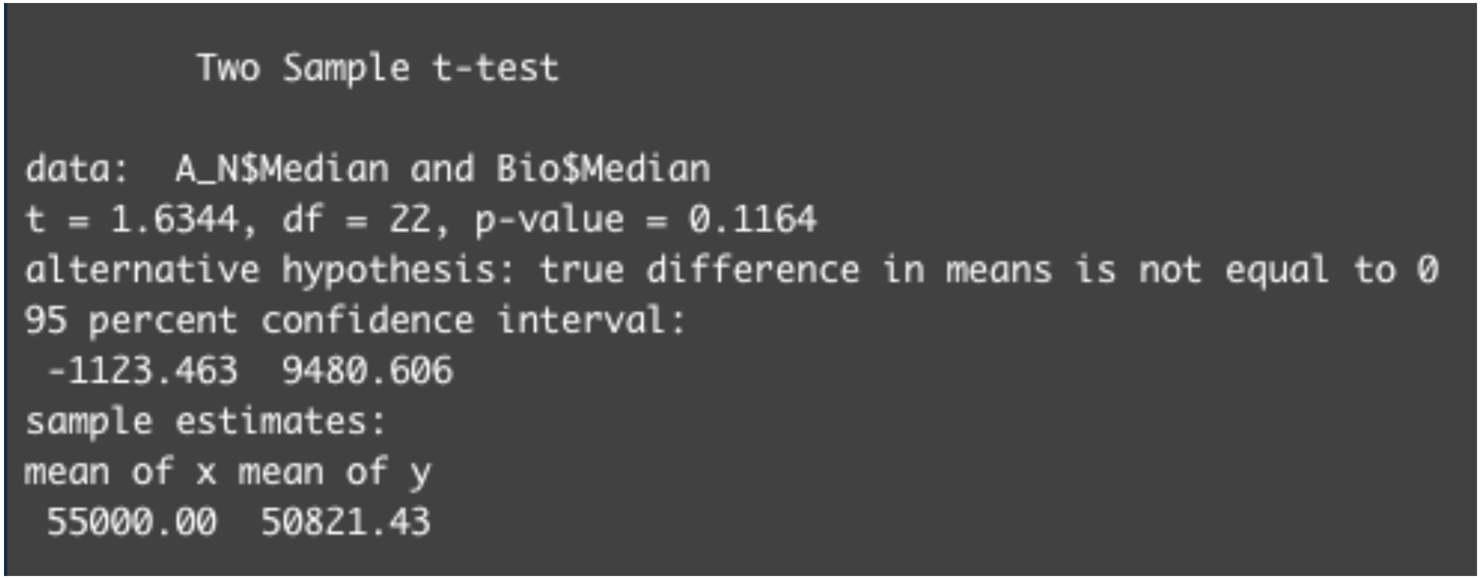
- Here we have output from a two-tailed t-test in R.
- We have 22 degrees of freedom, so we know we had (N-2) 24 observations.
- Our t-value is 1.63, which even without looking at a t-table or p-value, we know is kind of low.



```

	Two Sample t-test

data:  A_N$Median and Bio$Median
t = 1.6344, df = 22, p-value = 0.1164
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -1123.463  9480.606
sample estimates:
mean of x mean of y
 55000.00  50821.43
```

Agriculture vs. Biology t-test, R Output

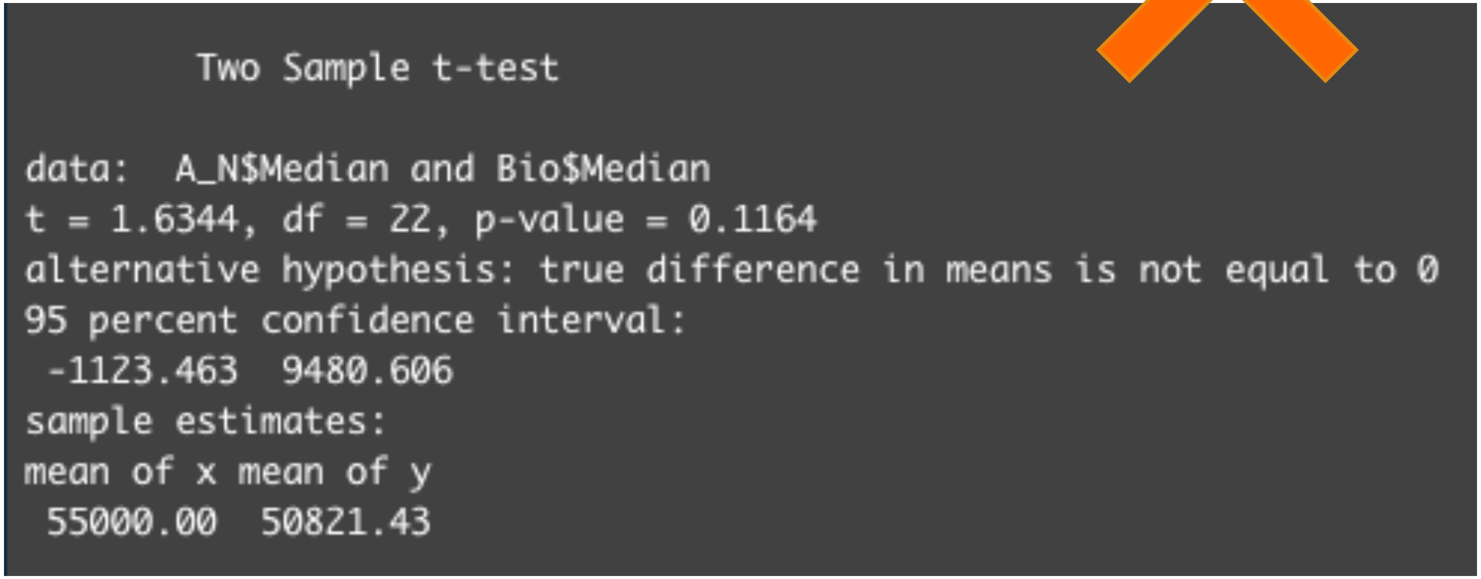
- The p-value is 0.1164, so p > .05, we fail to reject our null hypothesis… The two groups do not differ.
- The confidence interval confirms this, too
  - 95% CI [-1,123.46, 9,480.61]
  - Notice how the interval contains zero



```
        Two Sample t-test

data:  A_N$Median and Bio$Median
t = 1.6344, df = 22, p-value = 0.1164
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -1123.463  9480.606
sample estimates:
mean of x mean of y 
 55000.00  50821.43 
```

Agriculture vs. Biology t-test, R Output

- We would conclude that Agriculture & Natural Resources and Biology & Life Science are not statistically significantly different from each other in terms of Median income.
  - 95% CI [-1,123.46, 9,480.61]
  - $t(22) = 1.63, p = 0.116$



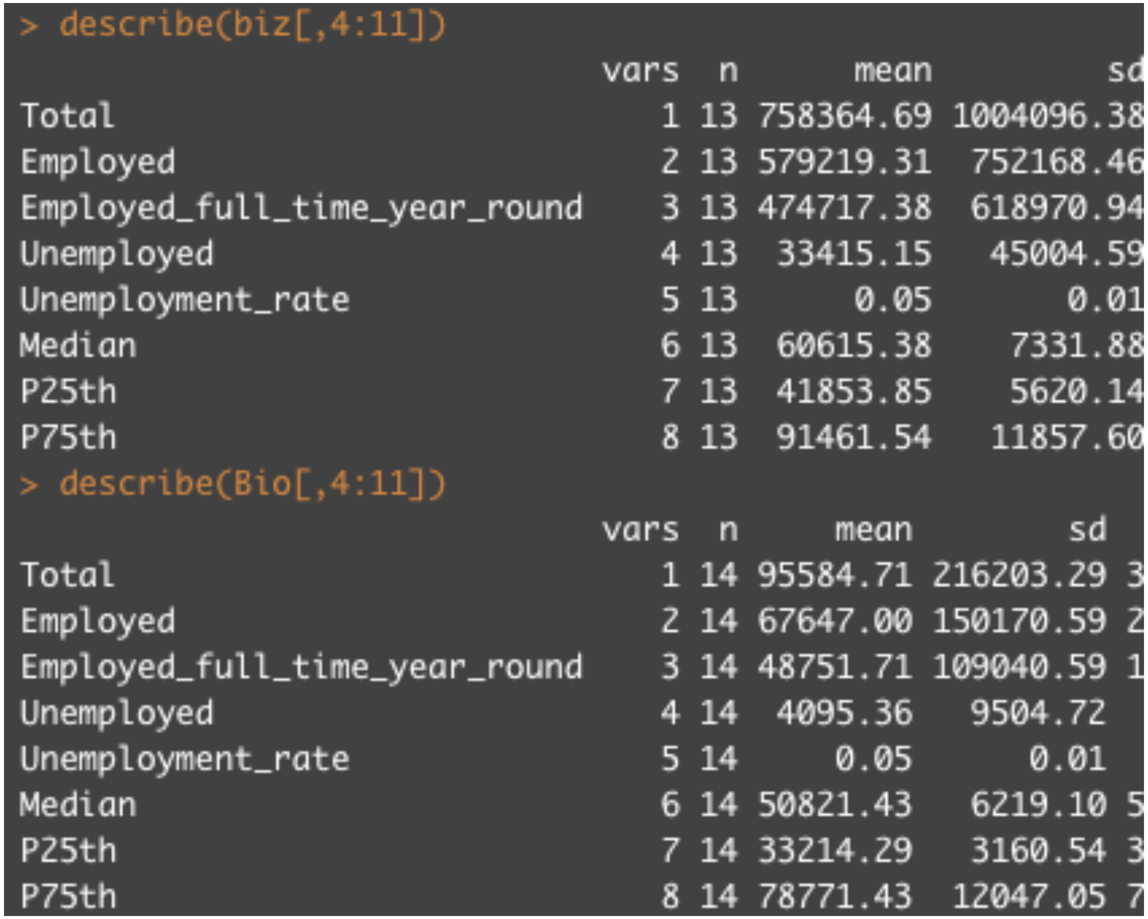
```
        Two Sample t-test

data:  A_N$Median and Bio$Median
t = 1.6344, df = 22, p-value = 0.1164
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -1123.463  9480.606
sample estimates:
mean of x mean of y 
 55000.00  50821.43 
```

Business vs. Biology t-test

- Do Business vs. Biology differ significantly from each other in median income?
  - Test at .05 level.
- H_0: Business vs. Biology do **<u>NOT</u>** differ significantly from each other in median income (There is no difference in median income).
- H_1: Business vs. Biology **<u>DO</u>** differ significantly from each other in median income (There is no difference in median income).

Descriptive Statistics

- First, let's look at some descriptive statistics using the "psych" package.
  - "Total" is the number of people who graduated with that major.
  - "Median" is the median income of that major.
- Does it look like Business and Biology differ *significantly* in their median income?

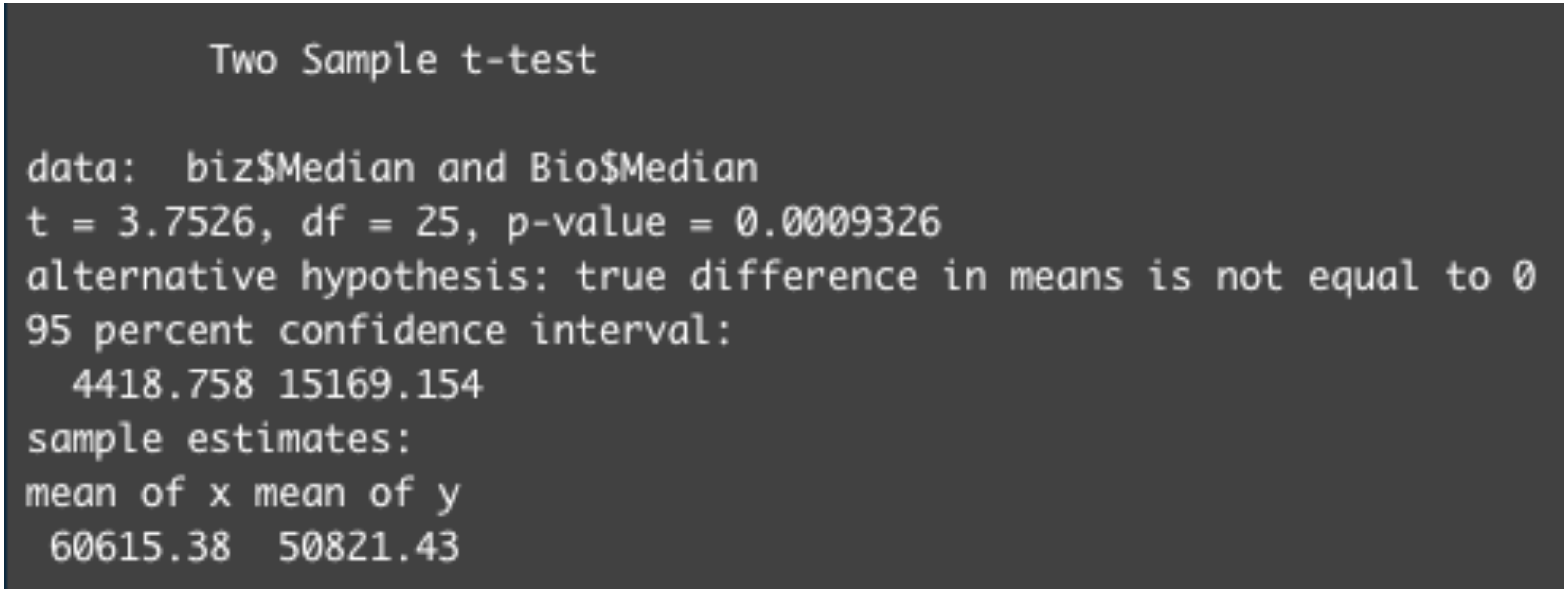


```
> describe(biz[,4:11])
                              vars  n      mean         sd
Total                            1 13 758364.69 1004096.38
Employed                         2 13 579219.31  752168.46
Employed_full_time_year_round    3 13 474717.38  618970.94
Unemployed                       4 13  33415.15   45004.59
Unemployment_rate                5 13      0.05       0.01
Median                           6 13  60615.38    7331.88
P25th                            7 13  41853.85    5620.14
P75th                            8 13  91461.54   11857.60
> describe(Bio[,4:11])
                              vars  n     mean        sd
Total                            1 14 95584.71 216203.29 3
Employed                         2 14 67647.00 150170.59 2
Employed_full_time_year_round    3 14 48751.71 109040.59 1
Unemployed                       4 14  4095.36   9504.72
Unemployment_rate                5 14     0.05      0.01
Median                           6 14 50821.43   6219.10 5
P25th                            7 14 33214.29   3160.54 3
P75th                            8 14 78771.43  12047.05 7
```

$60,615.38 - $50,821.43 = $9,793.95

Business vs. Biology t-test, R Output

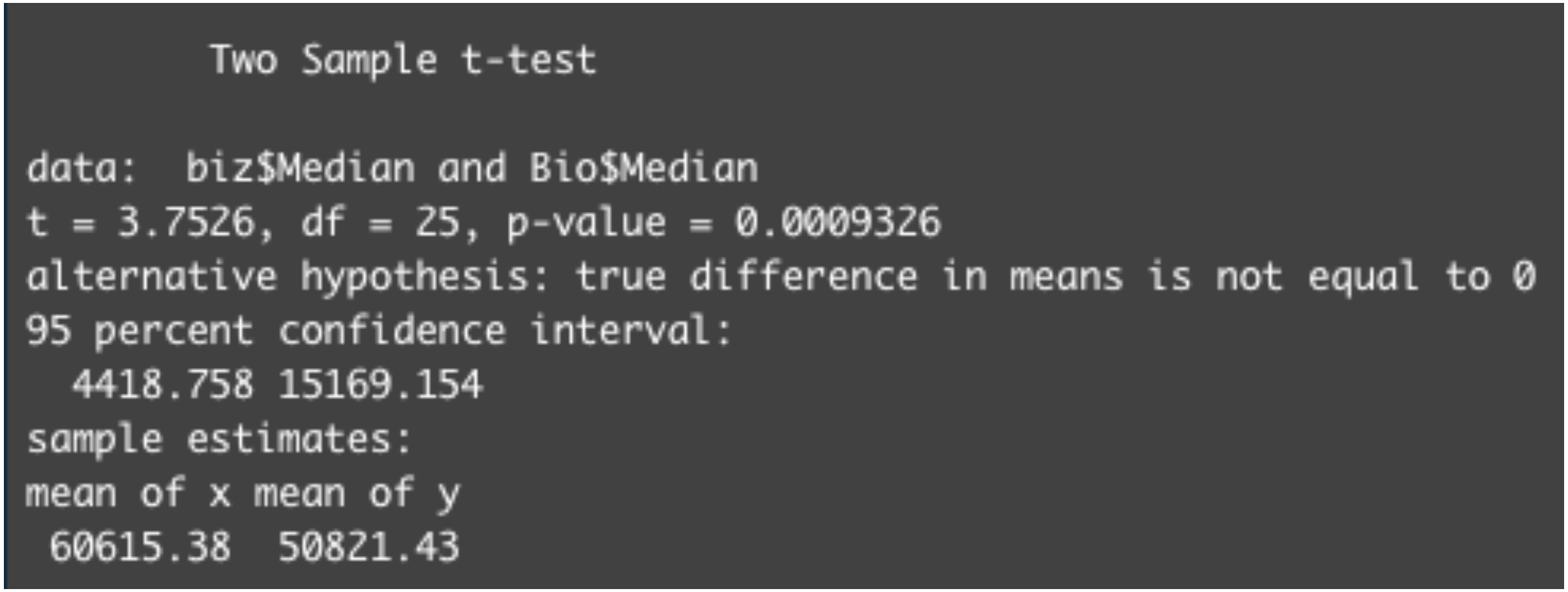
- Here we have output from a two-tailed t-test in R.
- We have 25 degrees of freedom, so we know we had (N-2) 27 observations.
- Our t-value is 3.7526, which looks high.



```
        Two Sample t-test

data:  biz$Median and Bio$Median
t = 3.7526, df = 25, p-value = 0.0009326
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
  4418.758 15169.154
sample estimates:
mean of x mean of y 
 60615.38  50821.43 
```

Business vs. Biology t-test, R Output

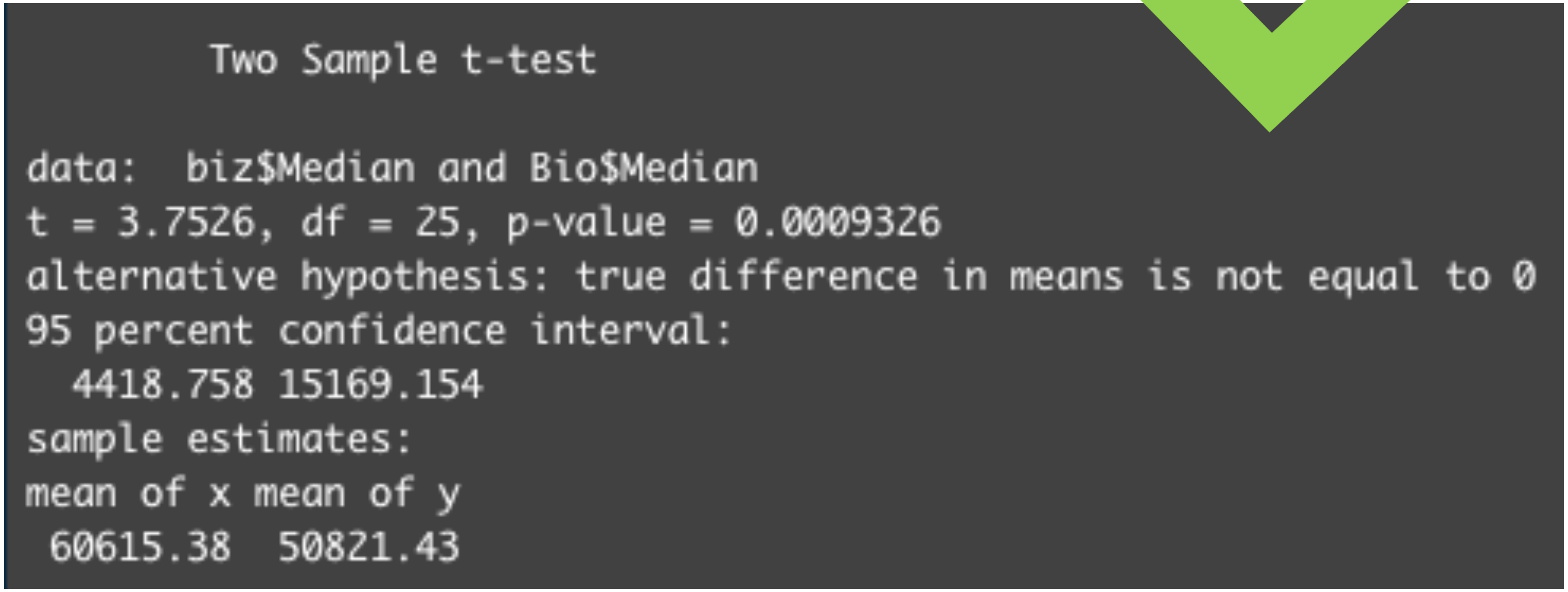
- The p-value is 0.001, so p < .05, we reject our null hypothesis that the two groups do *not* differ.
- The confidence interval confirms this, too
  - 95% CI [$4,418.76, $15,169.15]
  - Notice how the interval contain does not contain zero



```
	Two Sample t-test

data:  biz$Median and Bio$Median
t = 3.7526, df = 25, p-value = 0.0009326
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
  4418.758 15169.154
sample estimates:
mean of x mean of y 
 60615.38  50821.43 
```

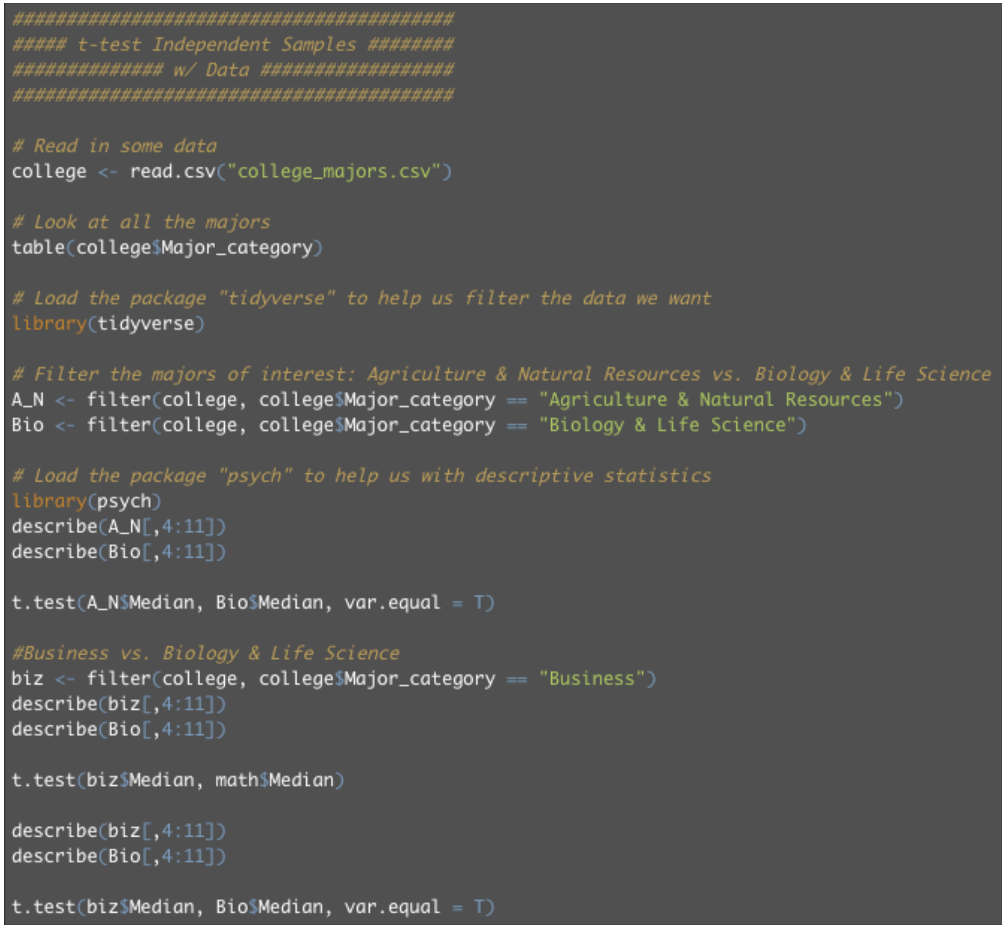
Business vs. Biology t-test, R Output

- We would conclude that Business vs. Biology are statistically significantly different from each other in Median income.
  - Business majors make more money.
  - 95% CI [$4,418.76, $15,169.15]
  - $t(25) = 3.75, p = 0.001$



```
        Two Sample t-test

data:  biz$Median and Bio$Median
t = 3.7526, df = 25, p-value = 0.0009326
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
  4418.758 15169.154
sample estimates:
mean of x mean of y 
 60615.38  50821.43 
```

R Code for t-test



```r
#########################################
##### t-test Independent Samples ########
############## w/ Data ##################
#########################################

# Read in some data
college <- read.csv("college_majors.csv")

# Look at all the majors
table(college$Major_category)

# Load the package "tidyverse" to help us filter the data we want
library(tidyverse)

# Filter the majors of interest: Agriculture & Natural Resources vs. Biology & Life Science
A_N <- filter(college, college$Major_category == "Agriculture & Natural Resources")
Bio <- filter(college, college$Major_category == "Biology & Life Science")

# Load the package "psych" to help us with descriptive statistics
library(psych)
describe(A_N[,4:11])
describe(Bio[,4:11])

t.test(A_N$Median, Bio$Median, var.equal = T)

#Business vs. Biology & Life Science
biz <- filter(college, college$Major_category == "Business")
describe(biz[,4:11])
describe(Bio[,4:11])

t.test(biz$Median, math$Median)

describe(biz[,4:11])
describe(Bio[,4:11])

t.test(biz$Median, Bio$Median, var.equal = T)
```

Data Source: The Economic Guide To Picking A College Major
https://github.com/fivethirtyeight/data/tree/master/college-majors
http://www.census.gov/programs-surveys/acs/data/pums.html
http://www.census.gov/programs-surveys/acs/technical-documentation/pums.html