IST772 Chapter Notes Template: After Completing Please Submit as a PDF.
Originality Assertion: By submitting this file you affirm that this writing is your own.

Name: Wanyue Xiao
Date: September 8, 2020
Chapter Number: # 4
Title of Chapter: Introducing the Logic of Inference Using Confidence Intervals

## Inference

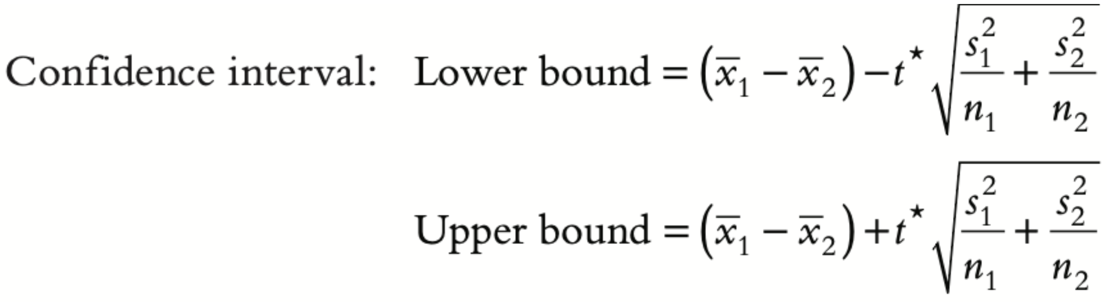
**Inference** refers to a reasoning process that begins with some information and leads to some conclusion.
**Induction** refers to the process that reasons from specific cases to the more general.

- keep in mind when reasoning from samples of data: **You cannot *prove* anything from samples or by using statistical inference.**
- Goal of statistical inference is to use a sample of data to **make estimates and/or inferences about a population**, with some degree of certainty or confidence
- Each of these sample means is uncertain: each mean is what statisticians refer to as a **point estimate**

**T-test:** generalize to a population of mean differences using sample data from two independent groups of observations.
**Confidence interval:** is a range of values that's likely to include a population value with a certain degree of **confidence**. It is often expressed a % whereby a population means lies between an upper and lower **interval**.
**Reminder:** when we say "95% confidence interval," we are referring to the **proportion of constructed confidence intervals that would likely contain the true population value**. If we ran our transmission and fuel economy study 100 times, *in about 95 of those replications the samples of transmission data would lead to the calculation of a confidence interval that overlapped the true mean difference* in mpg. As well, about *five of those 100 replications* would *give us a confidence interval that was either too high or too low*—both ends of the confidence interval would either be above or below the population mean.

## Formulas for the Confidence Interval

Formula:



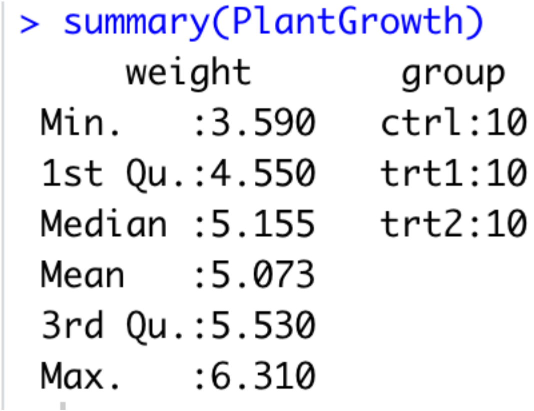
$$\text{Confidence interval:}\quad \text{Lower bound} = \left(\bar{x}_1 - \bar{x}_2\right) - t^{\star}\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

$$\text{Upper bound} = \left(\bar{x}_1 - \bar{x}_2\right) + t^{\star}\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

The first part of the formula is called "x-bars" which is the difference between two sample mean. The second part, which is called "the margin of error", is used to get the width of the confidence interval. The t* here is called t-distribution which will changed based on the sample size and the selected confidence level. The part under the square root is standard error which is the standard deviation of the sampling distribution of means. To get this value ($\frac{S_1^2}{N_1}$), we can square the standard deviation to get the variance and then divide the variance by the sample size.

## Exercise Review

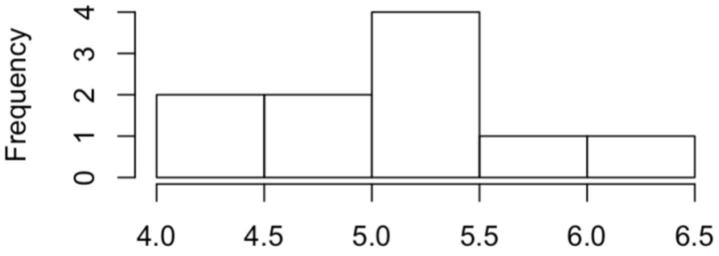
7.



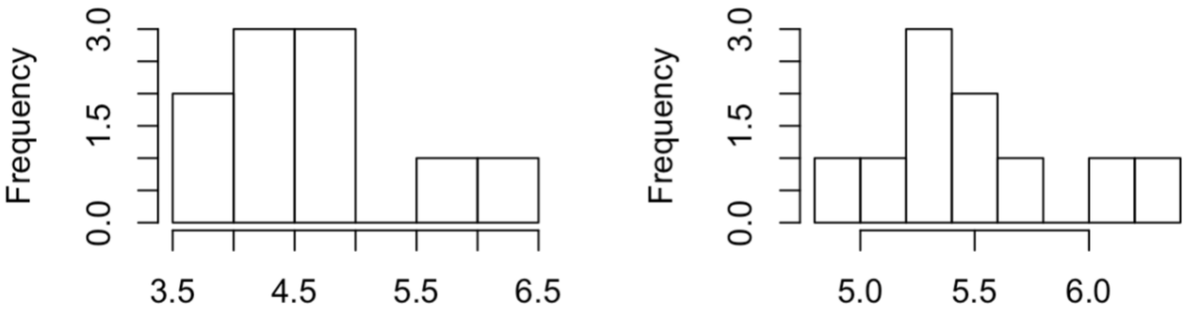
```
> summary(PlantGrowth)
     weight        group   
 Min.   :3.590   ctrl:10  
 1st Qu.:4.550   trt1:10  
 Median :5.155   trt2:10  
 Mean   :5.073            
 3rd Qu.:5.530            
 Max.   :6.310            
```

According to the result, we can summaries that the weight variable ranges from 3.590 to 6.310. Given that the mean (which is 5.073) is slightly lower that the median (which is 5.155), the distribution is slightly left-skewed. By looking at the group variable, we can summaries that it contains three types. Each type has exactly 10 observations.

**Hist of group "ctrl".**

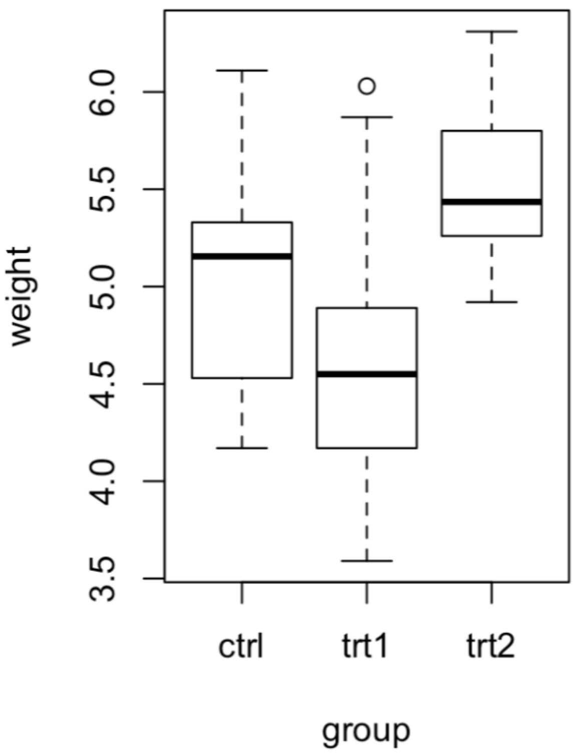


**Hist of group "trt1" (left) and "trt2" (right).**



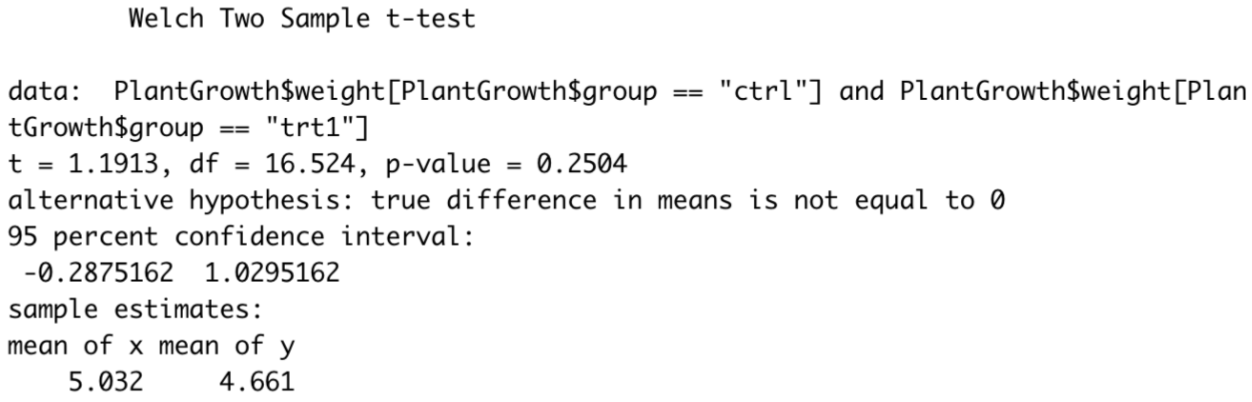
The difference of those two plots is the range of the x-axis. The range of trt1 is 3.5 to 6.5 while that of trt2 is about 5.0 to 6.3. Most of the values lie in the area of 4.5 in left plot while most of the values lie in the 5.5 area in the right plot.

8.



If we focus on the difference of range, ctrl has a largest range while trt2 has the smallest range. If we focus on the mean value, trt2 has the highest mean value while trt1 has the smallest mean value. Trt2 has the highest value while trt1 has the lowest value in their observations. Besides, trt1 has outliers.

9.

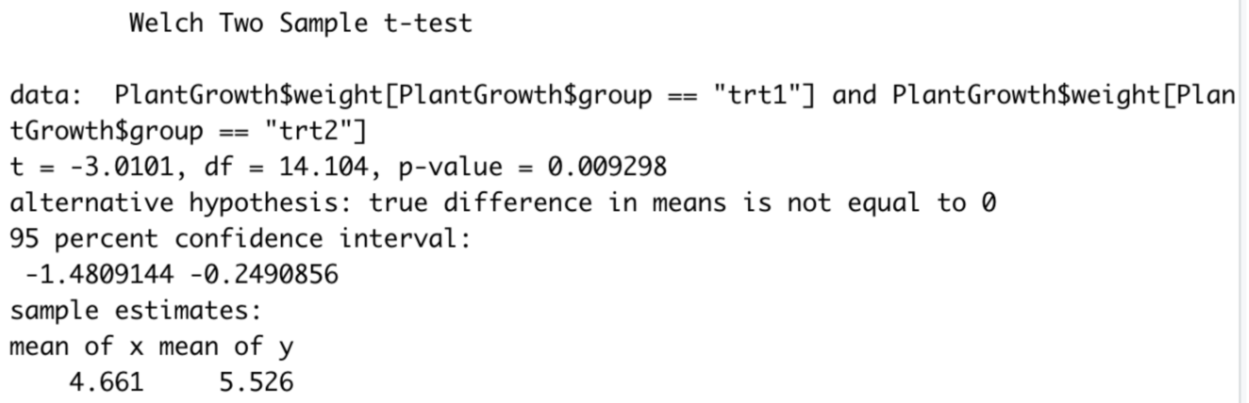


```
	Welch Two Sample t-test

data:  PlantGrowth$weight[PlantGrowth$group == "ctrl"] and PlantGrowth$weight[Plan
tGrowth$group == "trt1"]
t = 1.1913, df = 16.524, p-value = 0.2504
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -0.2875162  1.0295162
sample estimates:
mean of x mean of y 
    5.032     4.661 
```

There are 95 of those replication samples that the population difference in weight between group ctrl and group trt1 could be either a positive number or a negative number somewhere in the region of 0.37 plus or minus about 0.66. Here the 0.37 is the central of the region which is also the best point estimate.

10.



```
        Welch Two Sample t-test

data:  PlantGrowth$weight[PlantGrowth$group == "trt1"] and PlantGrowth$weight[Plan
tGrowth$group == "trt2"]
t = -3.0101, df = 14.104, p-value = 0.009298
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -1.4809144 -0.2490856
sample estimates:
mean of x mean of y
    4.661     5.526
```

There are 95 of those replication samples that the population difference in weight between group ctrl and group trt1 could be a negative number somewhere in the region of -0.865 plus or minus about 0.615. The confidence interval does not prove that there is a difference in group ctrl and group trt1 in the weight variable, but it does suggest that there is a possibility. Here the -0.865 is the central of the region which is also the best point estimate.

## R Code Fragment and Explanation

**R code:**

t.test(x, y = NULL, alternative = c("two.sided", "less", "greater"), mu = 0, paired = FALSE, var.equal = FALSE, conf.level = 0.95)

- **x: is a numeric vector of data values**
- y: is an *optional* numeric vector of data values. If excluded, the function performs a one-sample t-test on the data contained in x, if it is included it performs a two-sample t-tests using both x and y.
- mu argument provides a number indicating the true value of the mean (or difference in means if you are performing a two-sample test) under the null hypothesis. By default, the test performs a two-sided t-test.
- paired argument: indicate whether or not you want a paired t-test. The default is set to FALSE but can be set to TRUE if you desire to perform a paired t-test.
- Var.euqal: indicates whether or not to assume equal variances when performing a two-sample t-test. The default assumes unequal variance and applies the Welsh approximation to the degrees of freedom; however, you can set this to true to pool the variance.
- **conf.level:** determines the confidence level of the reported confidence interval for μ in the one-sample case and μ1−μ2 in the two-sample case.

## Question for Class

Is that mean, using the sample mean x-bar, we can get the true population mean but with some errors (which is the confidence interval)? Besides, what is the difference between z-test and t-test?