CS 251 Statistical Computing

HOP 5: R for statistical project

2/19/2020 Developed by Aya Khalil

04/02/2020 Reviewed by Apiwat Chuaphan

School of Technology and Computing @City University of Seattle (CityU)

**Before You Start**

* If you already finished this module through any CityU Technology Institute (TI) courses,  
  just skim this module and skip it.
* Version numbers may not match with the guide. But that should be fine.  
  If given the option to choose between stable release (long-term support) or most recent, please choose the stable release.
* This guide targets Windows OS users. So, MacOS users may have different commands to input in the shell/terminal.
* We cannot explain every step. **This cookbook always needs your own creative judgement.**
* **For your working directory, use your course number.** The hands-on tutorial may use a different course number as an example.

**Learning Outcomes**

* Two samples Hypothesis test
* F-test
* Anova

**Resource**

* Hui, E. G. M. (2019). [*Learn R for applied statistics: With data visualization, regressions, and statistics*](https://login.proxy.cityu.edu/sso/skillport?context=144516). Apress.
* Statistics for Data Analysis Using R online course

**Section1: Two-sample hypothesis testing**

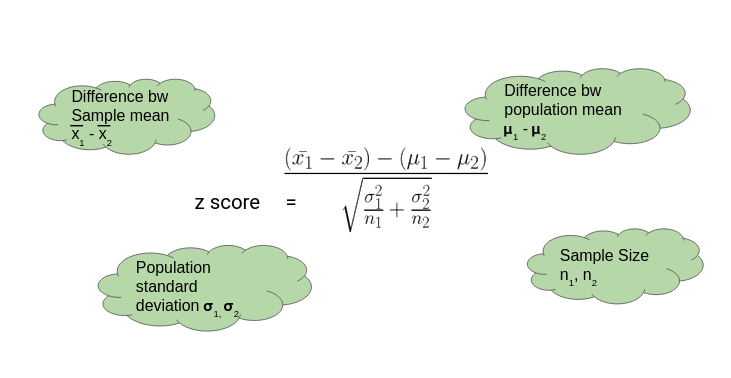
In statistical hypothesis testing, a two-sample test is a test performed on the data of two random samples, each independently obtained from a different given population. The purpose of the test is to determine whether the difference between these two populations is statistically significant. [Wikipedia](https://en.wikipedia.org/wiki/Two-sample_hypothesis_testing)

**Please watch the following video to have a general idea about Two-sample hypothesis testing**

Hypothesis Testing - Difference of Two Means - Student's -Distribution & Normal Distribution. Retrieved from: <https://www.youtube.com/watch?v=UcZwyzwWU7o&list=PL0o_zxa4K1BVsziIRdfv4Hl4UIqDZhXWV&index=57>

**Two-sample z test hypothesis testing**

We perform a Two Sample Z test when we want to compare the mean of two samples.



where equation and equation are the means of the two samples, Δ is the hypothesized difference between the population means (0 if testing for equal means), σ1 and σ2 are the standard deviations of the two populations, and n1and n2are the sizes of the two samples.

**Setup Working Environment for Module6**

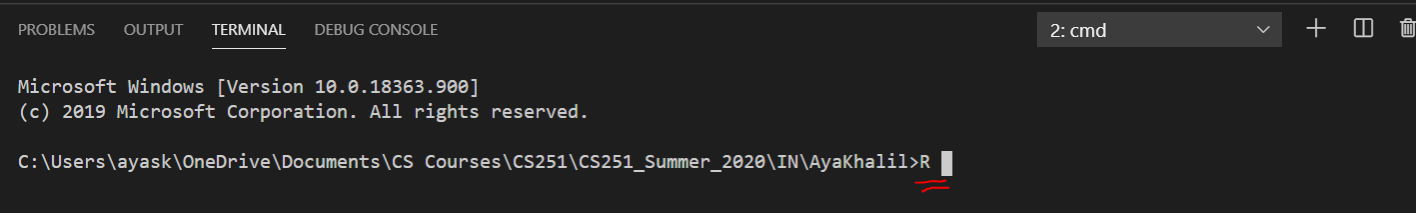
1. Open VS Code.
2. Go to your project folder

* **online student:** Open CS251 \_Fall\_2020/**ON**/FirstnameLastname /. ( File > Open )
* **onsite student:** Open CS251 \_ Fall \_2020/**IN**/FirstnameLastname. ( File > Open )

1. Then create “**Module6**” directory in the VSCode.

>>>mkdir Module6

**To open the R terminal, write in the console R**



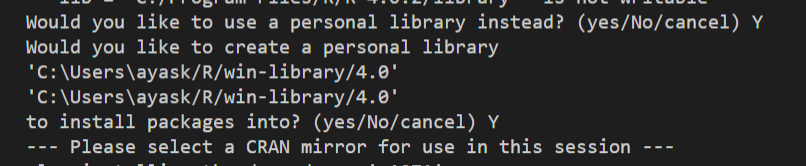
**Then enter**

* **Install BSDA package using the following command:**

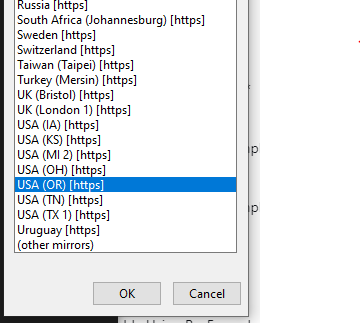
**Install.packages(“BSDA”)**



**If you were asked any of the following questions, just type y**



**Then choose any of the USA CRAN**



**Problem:**

We have two machines, machine1 & machine2. We took 100 samples from both machines. Now, we want to check whether there is any significant difference between the mean of these two machines or not. Check at 95% confidence level.

Answer:

We want to compare the mean of the two samples of the machines. In one-sample, we were comparing the sample mean with specific value, but here in two-samples, we compare between two machines whether they are producing the similar volume or not.

**So, for this problem**

**Null hypothesis**: H0: μ1 = μ2

or H0: μ1 – μ2= 0

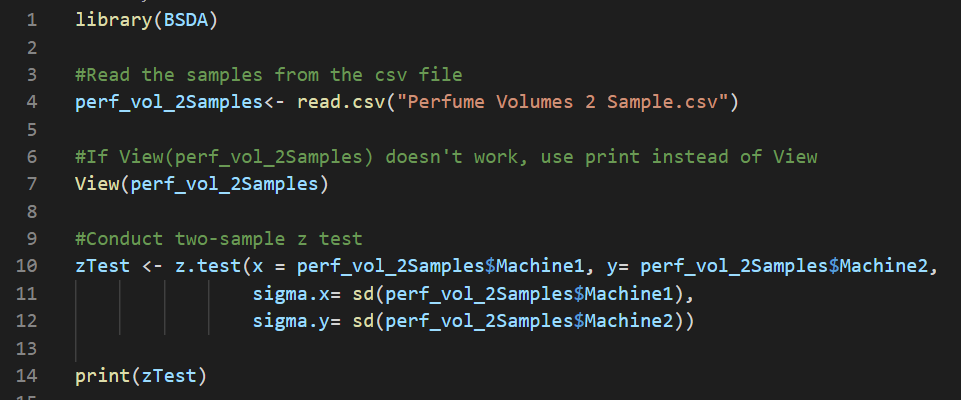
**alternative hypothesis**: Ha: μ1 ≠ μ2

**Download the attached CSV file “Perfume Volumes 2 Sample.csv” and make sure that you have it in your working directory like what we did in the previous HOP5A- Part1.**

**You should be in:**

* **online student:** CS251 \_ Fall \_2020/**ON**/FirstnameLastname
* **onsite student:** CS2511 \_ Fall \_2020/**IN**/FirstnameLastname
* **In Module6 project folder, create new file TwoSampleZtest.R**

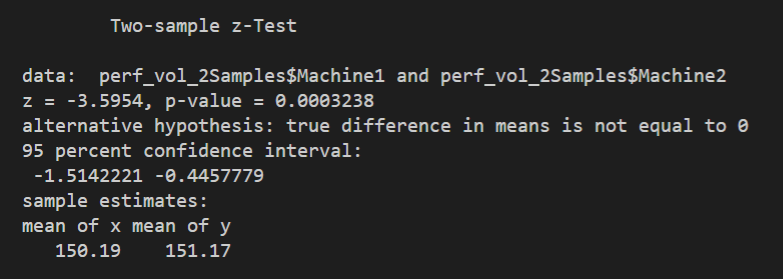
Type the following in TwoSampleZtest



**Code Explanation:**

1. BSDA package for functions and data sets for the text Basic Statistics & Data Analysis. So, BSDA allows us to use the z.test function to calculate two-sample z test
2. Read the csv file which contains the data of our samples
3. View/print the content of csv file to see what the samples look like.
4. Conduct two-sample z test.
   1. For the first argument in the function x= perf\_vol\_2Samples$Machine1 is the samples of the first machine or the first column in the table.
   2. For the second argument y= perf\_vol\_2Samples$Machine2, y is the sample data of machine2
   3. We don’t know the population standard deviation, but as long as we have a large number of data (100 samples) then we can calculate the standard deviation for the sample and use it as the population standard deviation, so the third argument in the function is calculating the standard deviation to the data of machine1
   4. For the fourth argument, calculating the standard deviation to the data of machine2
   5. We didn’t add the mean as an argument in the function because the null hypothesis is assuming the difference between the two means is equal to zero.
   6. Also, we didn’t add the confidence level because by default it’s 95% which means the significance level is equal to 0.05

**Output**



**Output Explanation**

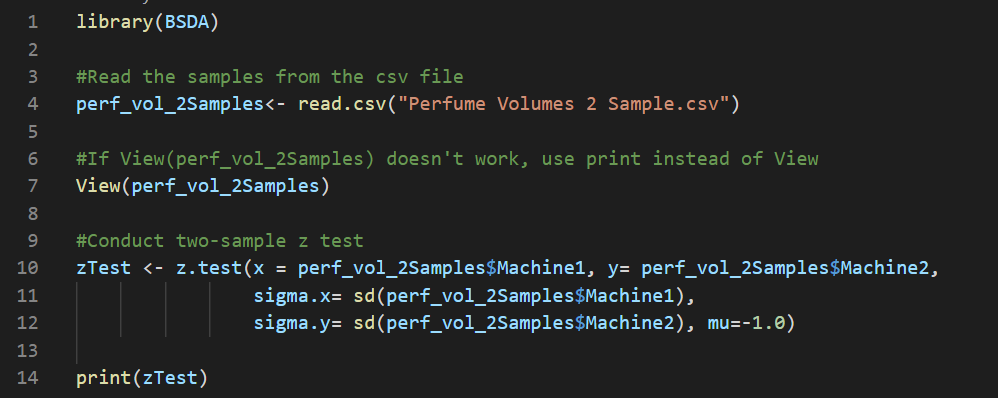
* First thing we will look at in the output is p-value and you will notice that p-value < 0.05 or the significance level, so the null hypothesis will be rejected.
* Also, we rejected the null because μ1 ≠ μ2 and you can see that the mean of x= 150.19, and the mean of y 151.17 are not equal.

**Problem**:

Let’s try to solve the previous problem one more time but now assume that there is a difference in the performance of machine1 & machine2. Machine2 is supposed to make **1cc** more than Machine1. Now, we want to check whether the null hypothesis is true or not. Check at 95% confidence level.

**Answer**:

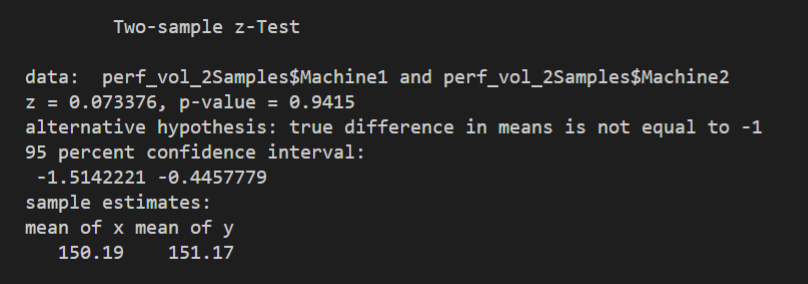
* **In Module6 project folder, create new file ZTwoSampDifMean.R**
* Type the following in **ZTwoSampDifMean**



**Code Explanation**

1. BSDA package for functions and data sets for the text Basic Statistics & Data Analysis. So, BSDA allows us to use the z.test function to calculate two-sample z test
2. Read the csv file which contains the data of our samples
3. View/print the content of csv file to see what the samples look like.
4. Conduct two-sample z test.
   1. For the first argument in the function x= perf\_vol\_2Samples$Machine1 is the samples of the first machine or the first column in the table.
   2. For the second argument y= perf\_vol\_2Samples$Machine2, y is the sample data of machine2
   3. We don’t know the population standard deviation, but as long as we have a large number of data (100 samples) then we can calculate the standard deviation for the sample and use it as the population standard deviation, so the third argument in the function is calculating the standard deviation to the data of machine1
   4. For the fourth argument, calculating the standard deviation to the data of machine2
   5. The difference between the mean of Machine1 & Machine2 is equal to -1, so the fifth argument in the function is mu=-1 which is μ1 - μ2
   6. Also, we didn’t add the confidence level because by default it’s 95% which means the significance level is equal to 0.05

**Output**



**Output Explanation**

* First thing we will look at in the output is p-value and you will notice that p-value > 0.05 or the significance level, so the null hypothesis will not be rejected or failed to reject the null hypothesis.
* Also, the difference between the mean of x= 150.19, and the mean of y 151.17 approximately is equal to 1.

**Two-samples t test hypothesis testing**

If two set of data are independent or dependent.

* If the values in one sample reveal no information about those of the other sample, then the samples are independent
  + Example: Blood pressure of male/female
  + Then it is:

Two sample t test

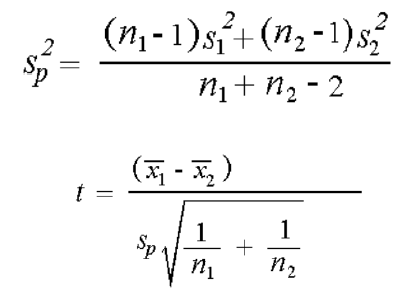
* If the values in one sample affect the values in the other sample, then the samples are dependent.
  + Example: Blood pressure before and after a specific medicine
  + Then, it is:

Paired t test

**Two-samples t test**

Is variance for two samples equal?

If yes: Pooled variance calculate Sp for finding out t

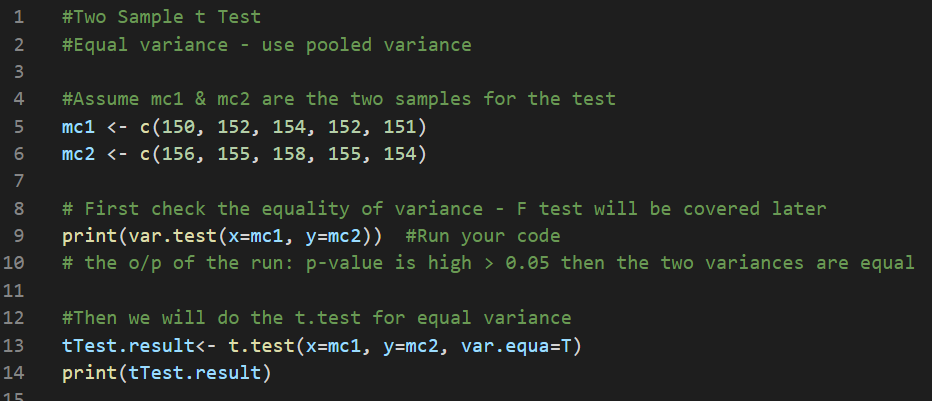


Problem:

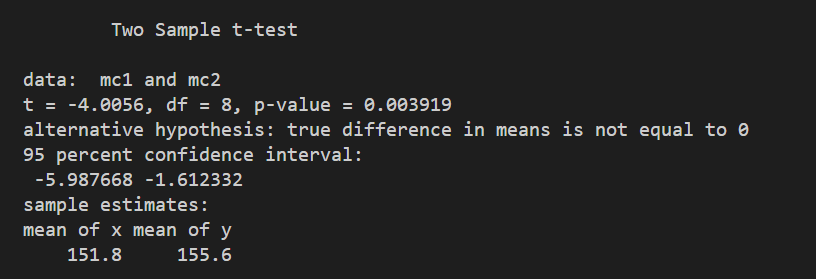
We took 5 samples from machine number1 and 5 samples from machine number2 and for these samples, we want to check whether the mean is equal or not.

Answer:

* **In Module6 project folder, create new file tTestEqualVar.R**
* Type the following in **tTestEqualVar**



Run your code

Output: 

In the output, the p-value<0.05 (alpha-value), then null hypothesis must be rejected. Also, the mean of x is not equal to y which means the null hypothesis should be rejected.

**Paired t test**

Where you have two samples in which observations in one sample can be paired with observations in the other sample

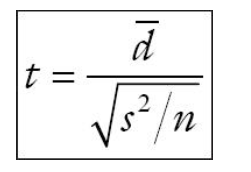
Or

If the values in one sample affect the values in the other sample, (the samples are dependent)

Example: Blood pressure before and after a specific medicine

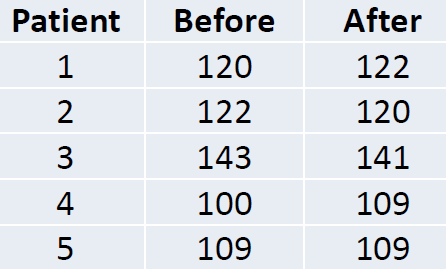
To calculate paired t test

* Find the difference between two set of readings as d1, d2 …. dn.
* Find the mean and standard deviation of these differences.



Problem:

Before and after medicine BP was measured as you can see in the below table. Is there a difference at 95% confidence level?



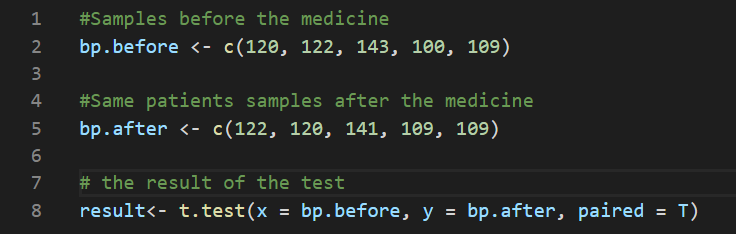
Answer:

**So, for this problem**

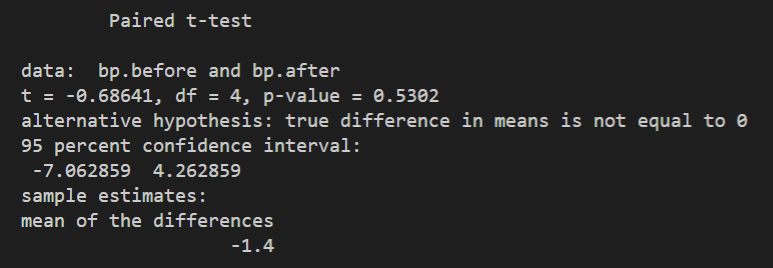
**Null hypothesis**: H0: No difference before and after taking the medicine, bp.before=bp.after

**alternative hypothesis**: Ha: there is a difference after and before the medicine, bp.before≠bp.after

* **In Module6 project folder, create new file tPairdTest.R**
* Type the following in **tPairdTest.R**



Run your code.

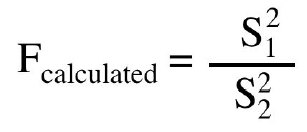
Output: 

Since the value of p is greater than 0.05 (alpha value), we fail to reject the null hypothesis. There are not enough evidences to prove that this medicine has any effect on the blood pressure.

**Section2: Two sample variance test using F Test**

F-test is used for testing equality of two variances from different populations

H0: σ21= σ22



**Please watch the following videos to understand the concept of F test**

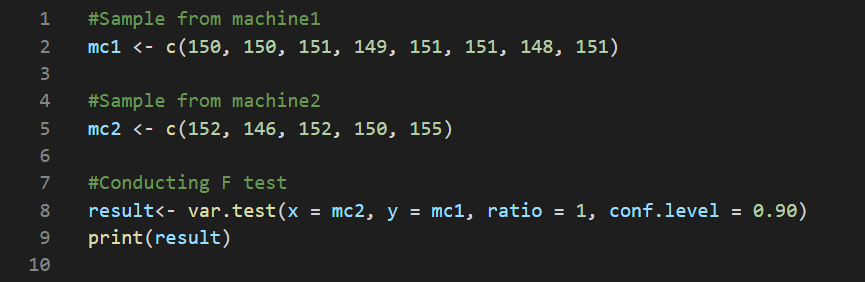
* R Tutorial: Testing Equality of Variances. Retrieved from: <https://www.youtube.com/watch?v=vnbv1jD-ntI>
* F-Test for two sample variances. Retrieved from: <https://www.youtube.com/watch?v=Pml68e3Eh3o>

Problem:

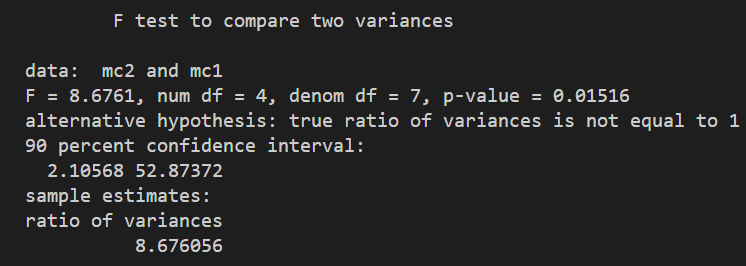
We took 8 samples from machine A and the standard deviation was 1.1. From machine B we randomly picked 5 samples and the variance was 11. Is there a difference in the variance for machine A and B? Check with 90% confidence level.

Answer:

* **In Module6 project folder, create new file Ftest.R**
* Type the following in **Ftest.R**

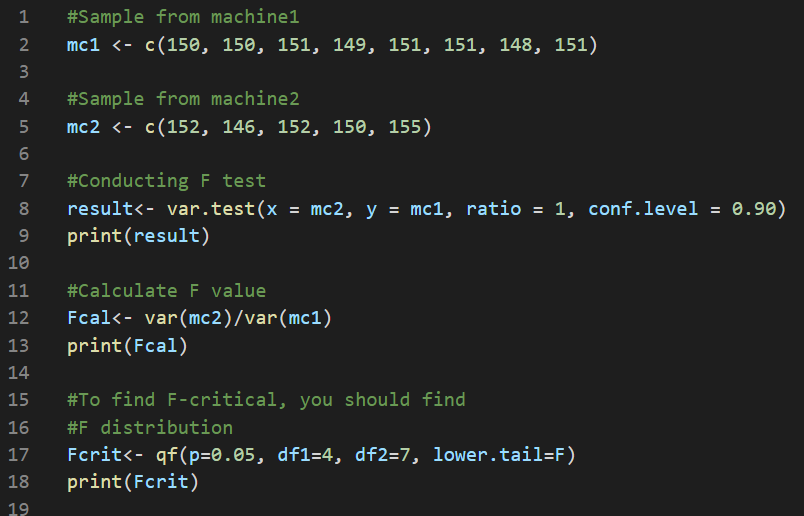


Run your code

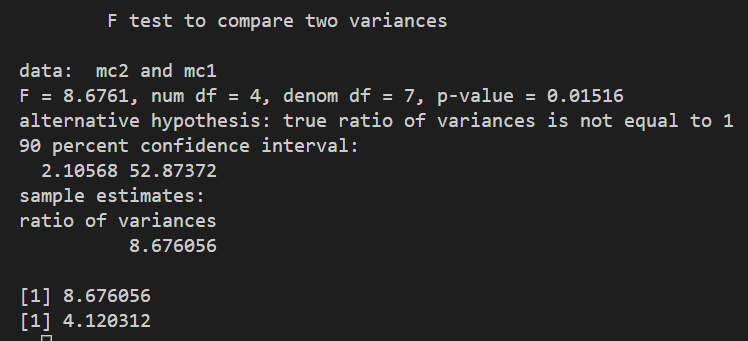
Output: 

Since the p value is less than 0.10(value of alpha) we reject the null hypothesis. This concludes that there is a significant difference between these two machines variance.

* Add the following to update **Ftest.R to calculate** F-value and F-critical



Run your code.

Output: 

Since F-critical is less than F-calculated which means F-calculated is in the rejection zone, so we should reject the null hypothesis.

**Section 3: Anova**

An ANOVA test is a way to find out if survey or experiment results are significant. In other words, they help you to figure out if you need to reject the null hypothesis or accept the alternate hypothesis.

Basically, you’re testing groups to see if there’s a difference between them. Examples of when you might want to test different groups:

Example,

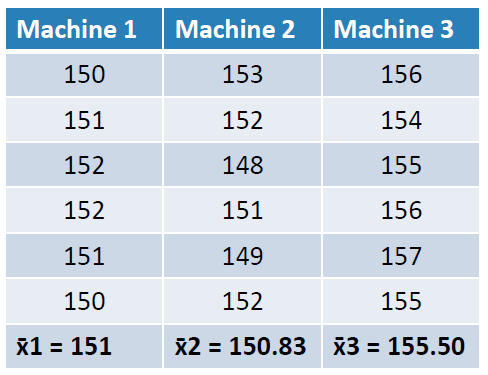
A group of psychiatric patients are trying three different therapies: counseling, medication, and biofeedback. You want to see if one therapy is better than the others.

**Please watch the following video to understand the concept of Anova**

Analysis of Variance (ANOVA) Overview in Statistics. Retrieved from: <https://www.youtube.com/watch?v=CS_BKChyPuc>

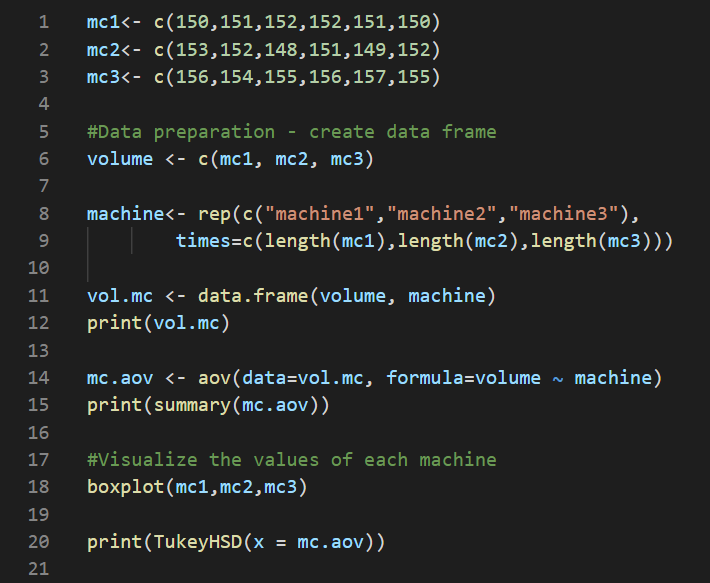
Problem:

Compare between three machines and analyze the variance to see if any of these machines is different from others regarding to the mean



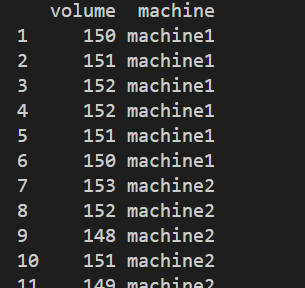
Answer

* **In Module6 project folder, create new file Anova.R**
* Type the following in **Anova.R**

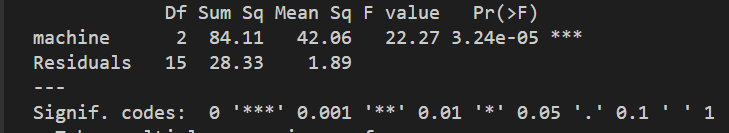


**Code Explanation:**

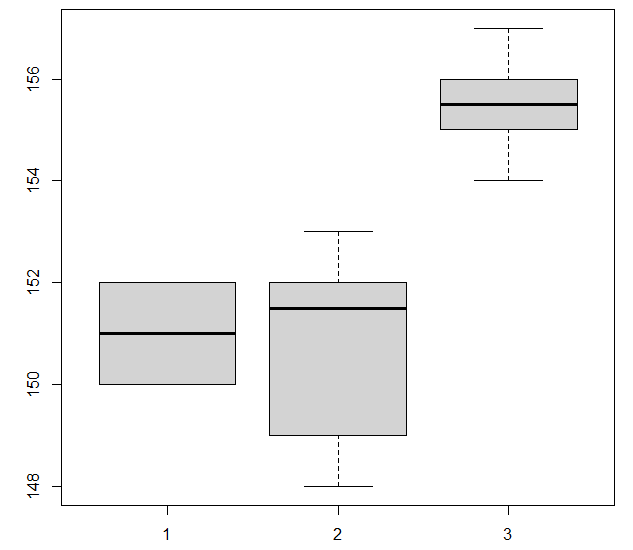
1. In the first three lines, we created three vectors that carry the values that each machine has.
2. In line#6, we created one long vector that carry the values of machine1,2 &3 that will be used later in the code.
3. In line#8, we repeated the word machine1, machine2 and machine3 with the length of the vector of mc1, mc2 &mc3.
4. In line #11, we created a data frame that contains two columns, one for the vector of volumes that we have in line#6 and one for the machine vector that we have in line#8, so the output of line#12 is the following which is our data that we will work on it.



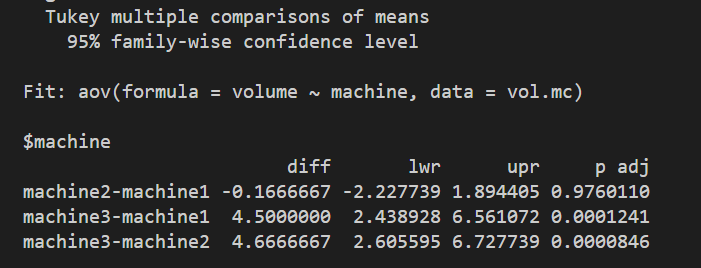
1. In line#14, we conducted the anova test, so first argument is our data which is vol.mc, and second argument contains two part, first is what you are measuring which is the volume and second is the factor that we are looking for which is the machine.
2. In line#15, we printed the summary of the output of the test which what you will get if you are doing manual calculations



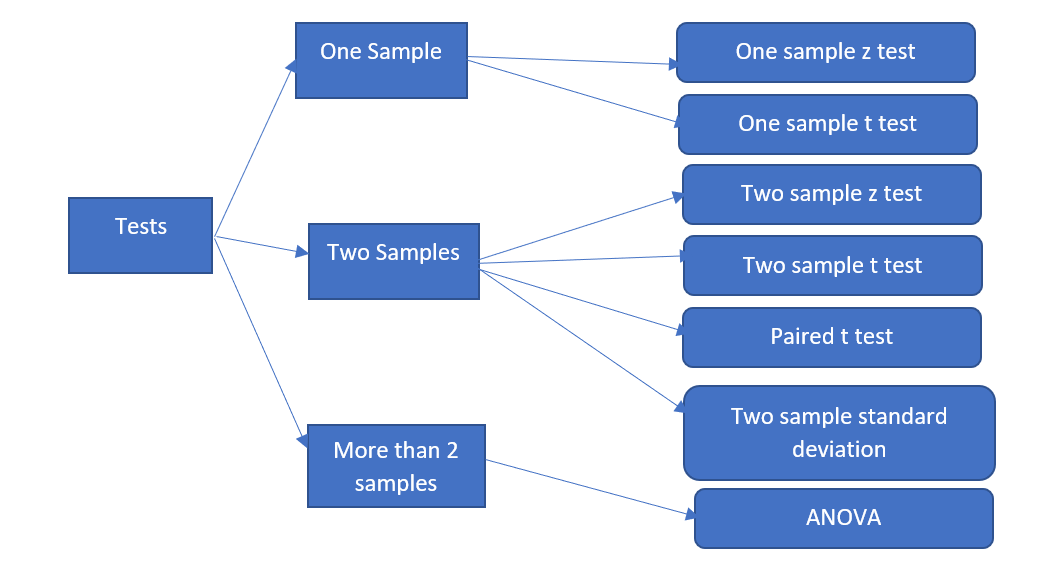
1. We visualize the values of mc1, mc2, mc3 to have an idea about how the values of each machine look like, so the output is the following and you will notice that machine3 is quite different from the other two machines



1. In line#20, we conducted the post-hoc Tukey Honset significant differences to decide which of these pairs are different, so that the test compares two machines at a time then it tells us which of these machines is different, so in the following output, you should look to the value of p. I f the value of p is low, then the null hypothesis should be rejected which mean there is a difference in machine3-machine1 & machine3-machine2



The following chart summarize what you have learned in HOP 5&6.



**Push your work to GitHub**

**Make sure you are in**

Onsite students: CS251\_ Fall \_2020/**IN**/FirstnameLastname

Online students: CS251\_ Fall \_2020/**ON**/FirstnameLastname

Run the following commands to push your work to the GitHub repository:

Open the terminal from the VSCode by hit the **control + ~** key and type the following command:

>>> git add .

>>> git commit -m “Submission for Module 6”

>>> git push origin YOUR\_BRANCH\_NAME

Note: you should change the YOUR\_BRANCH\_NAME to your own branch name. It should be firstname-lastname (e.g. maria-gracia).