ANOVA (Analysis of Variance)

**Help me choose my Dream Car with different Drive Wheels ?**

Most of us always like to compare groups and check whether there is any variation in the average values between each group and within groups.

Some of the typical use cases are:

1. We may be interested in checking whether there is a difference in the average teaching evaluation scores of younger instructors, middle-aged instructors, and old instructors.

A teacher teaching her students

Description automatically generated with medium confidence

1. Mary wants to maintain a healthy lifestyle and would like to look fit and trim. She is looking for a suitable weight loss program and is unable to decide on the options presented to her as all are looking good to her. The different weight loss programs are low calorie, low fat, and low carbohydrate diet. Before opting for these programs she wants to check whether there is a difference in mean weight loss due to these programs.

A picture containing food, dish, vegetable

Description automatically generated

1. Bob wants to buy his dream car and has gone to a car showroom where he is presented with the prices of cars. He also wants to know whether there is variation in the average price of cars with four-wheel drive, rear wheel drive, and front wheel drive.

A car parked in a parking lot

Description automatically generated

1. A professor wants to know if three different studying techniques lead to different exam scores.



1. A botanist wants to know whether or not plant growth is influenced by sunlight exposure and watering frequency.



There are many such examples but how do we find this difference? There should be some statistical test to check this variance.

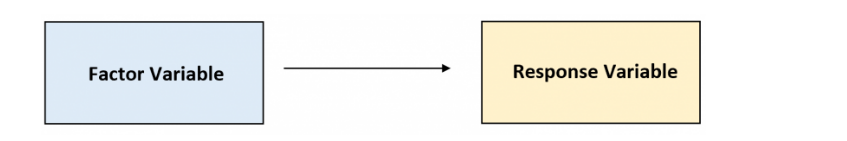
ANOVA (Analysis of Variance) is our knight in shining armor who will rescue us from this situation.

Analysis of variance (ANOVA) is a statistical technique that is used to check if the means of two or more groups are significantly different from each other.

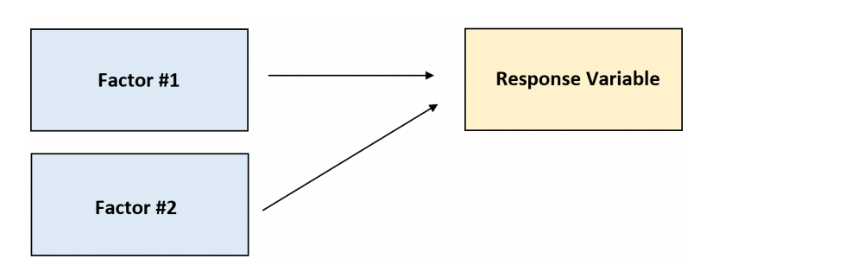
ANOVA checks the impact of one or more factors by comparing the means of different samples.

The two most common types of ANOVAs are one-way ANOVA and two-way ANOVA.

**One-way ANOVA:** Used to determine how one factor affects a response variable.



**Two-way ANOVA:** Used to determine how two factors affect a response variable, and to determine whether there is an interaction between the two factors on the response variable or not.



In case of Bob’s example, we have only one factor drive wheels influencing the response variable price.

However, with respect to the Botanist example we have 2 factors sunlight exposure and watering frequency influencing the plant growth.

For any test we need an assumption which can be tested either by experiment or observation. That is where we need an educated guess called hypothesis.

Basically, whenever we need to test ANOVA, we need 2 types of Hypothesis.

1.**Null Hypothesis:** Here we say that the sample means of all the groups are the same.

2. **Alternate Hypothesis**: Is valid when at least one of the sample means is different from the rest of the sample means

So now let’s understand ANOVA in simple terms.

ANOVA is stated as variation between group means/mean variation within the groups.

Sounds complicated right lets understand with a small case study and then explore it in python

I was aware of the one-line code from a built-in function in python which gives me the results soon for ANOVA.

But then I became curious and felt like exploring it step by step in the way ANOVA works, the statistics way using python code and wanted to validate this with the one-line code.

Let us rewind back to Bob’s case

Bob wants to buy his dream car and has gone to a car showroom where he is presented with the prices of cars. He also wants to know whether there is variation in the average price of cars with four-wheel drive, rear wheel drive and front wheel drive.

I considered an automobile dataset for my reference.

This data set consists of three types of entities:

1. the specification of an auto in terms of various characteristics,
2. its assigned insurance risk rating,
3. its normalized losses in use as compared to other cars.
4. The second rating corresponds to the degree to which the auto is riskier than its price indicates.

The description of this data is available in this link: [https://archive.ics.uci.edu/ml/datasets/Automobile](https://archive.ics.uci.edu/ml/datasets/Automobile?utm_medium=Exinfluencer&utm_source=Exinfluencer&utm_content=000026UJ&utm_term=10006555&utm_id=NA-SkillsNetwork-Channel-SkillsNetworkCoursesIBMDeveloperSkillsNetworkDA0101ENSkillsNetwork20235326-2021-01-01" \t "https://cf-courses-data.s3.us.cloud-object-storage.appdomain.cloud/IBMDeveloperSkillsNetwork-DA0101EN-SkillsNetwork/jupyterlite/lab/_blank).

Further the dataset is downloaded from this location

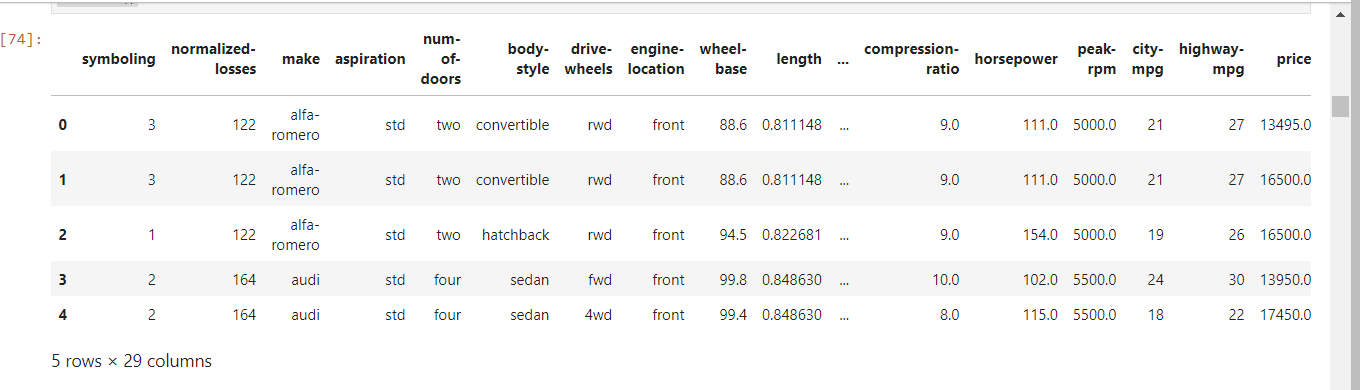
[https://archive.ics.uci.edu/ml/machine-learning-databases/autos/imports-85.data](https://archive.ics.uci.edu/ml/machine-learning-databases/autos/imports-85.data?utm_medium=Exinfluencer&utm_source=Exinfluencer&utm_content=000026UJ&utm_term=10006555&utm_id=NA-SkillsNetwork-Channel-SkillsNetworkCoursesIBMDeveloperSkillsNetworkDA0101ENSkillsNetwork20235326-2021-01-01" \t "https://cf-courses-data.s3.us.cloud-object-storage.appdomain.cloud/IBMDeveloperSkillsNetwork-DA0101EN-SkillsNetwork/jupyterlite/lab/_blank)

**Step by Step Analysis of statistics of ANOVA the Pythonic way**

So firstly, I will read the csv and convert that to a data frame.

df = pd.read\_csv("auto.csv")

df.head()

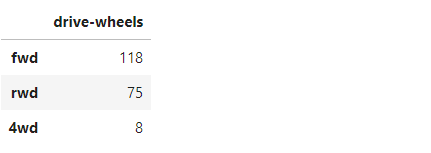


This is the snapshot of my data frame.

Now, my column of interest for ANOVA is drive-wheels. So, firstly let me obtain the value count of each wheel type.

Value counts is a good way of understanding how many units of each characteristic/variable we have. We can apply the "value\_counts" method on the column "drive-wheels".

df['drive-wheels'].value\_counts().to\_frame()



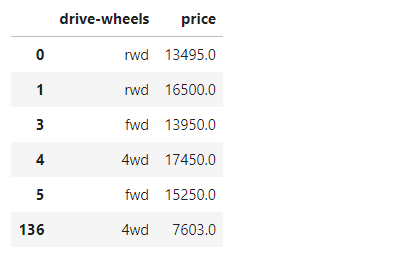
Here we have 3 groups front wheel drive, rear wheel drive and four-wheel drive.

So, I will create a small subset of a data frame from the existing data frame.

As my columns of interest are only drive-wheels and price, I create the data frame as follows

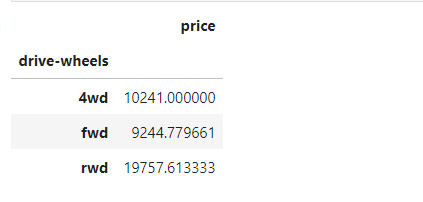
grouped\_test=df[['drive-wheels', 'price']].groupby(['drive-wheels'])

grouped\_test.head()



Next step is to calculate average price of each group.

grouped\_test.mean()



Let’s then calculate the mean price of all the groups.

grouped\_test['price'].mean().sum()/3



The next step is I need to total the squared difference between each group mean and all groups mean which is calculated as:

VC1\*(-)2 +VC2\*(-)2+VC3\*(-)2

Where VC1 ->Value counts of 4wd which is 8

VC2->Value counts of fwd which is 118

VC3->Value counts of rwd which is 75

, and are means of 4wd,fwd and rwd groups.

is total group mean.

This is evaluated as follows:

I am using the power function pow() here for calculating the square.

8\*pow((10241-13081.13), 2) +118\*pow((9244.77-13081.13),2) +75\*pow((19757.61-13081.13),2)

This comes up to 5144368246.468.

Now, finally we need to do get Mean Square variation between groups/Mean square variation within groups

Mean Square Variation between Groups=SSB/(k-1)

Where, SSB=Sum of squared differences between group mean and overall mean which is 5144368246.468 in this case.

k= Number of groups which is 3

SSB/(k-1) is thus evaluated as:

5144368246.468/2 =2572184123.234

For calculating Mean square variation within groups, I need to consider the sum of the squared difference between each price value of that group with the mean price of the group.

To make it simpler I will create 3 data frames with only each group data.

rwd\_df=df[df['drive-wheels']=='rwd']

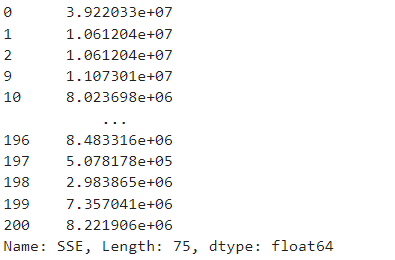
fwd\_df=df[df['drive-wheels']=='fwd']

fourwd\_df=df[dft['drive-wheels']=='4wd']

Next, I will do,

rwd\_df['SSE']=pow((rwd\_df['price']-19757.613333),2) where, 19757.613333 is the mean of the group rwd.

rwd\_df['SSE']



Later I will find the sum as follows:

rwd\_df['SSE'].sum()

I will repeat the same with the other 2 groups and finally add the sum of all the 3 groups.

rwd\_df['SSE'].sum()+fourwd\_df['SSE'].sum()+fwd\_df['SSE'].sum()

This is my output.



Mean Square Variation within Groups=SSE/(N-K)

Where, SSE=sum of the squared difference between each price value of that group with the mean price of the group and then again adding them up for all the groups.

N=Total Number of observations

K=Number of groups.

The length of the data frame will give the total number of observations which can be deduced as follows:

len(df)



MeanSquare Variation within Groups=7490000894.057853/(201-3) which is 37828287.34372653

Finally, we need to calculate the F-Test score.

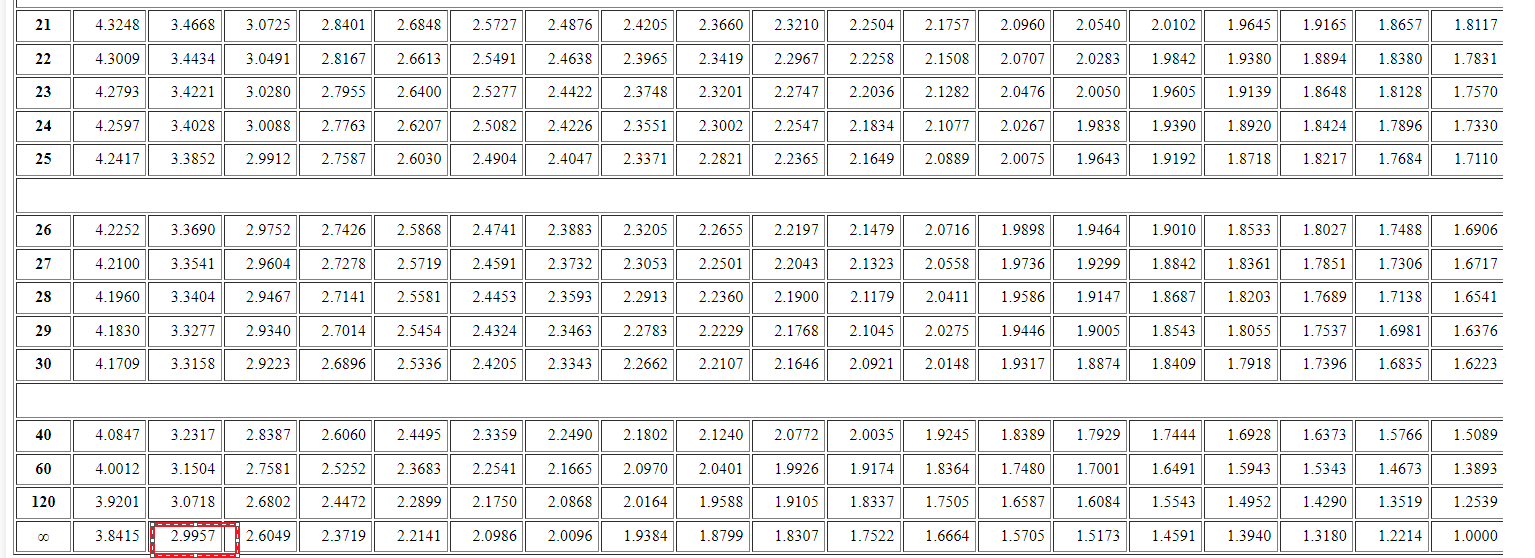
**F-t**est score: ANOVA assumes the means of all groups are the same, calculates how much the actual means deviate from the assumption, and reports it as the F-test score. A larger score means there is a larger difference between the means.

F-Test score=Mean Square variation between groups/Mean square variation within groups

2572184123.234/37828287.34372653

This comes upto 67.99631450035956

To see whether the difference is more I will validate against <http://www.socr.ucla.edu/Applets.dir/F_Table.html> for α = 0.05



Which is 2.997 and 67.99631450035956 > 2.997 .

Hence there is a huge difference in the sample means.

In our case I have considered only the influence of only one factor Drive wheels on response variable price. So, it is only one-way ANOVA.

**Step by Step Analysis of statistics of ANOVA using the Built in Function.**

We can use the built-in python function 'f\_oneway' in the module 'stats' to obtain the F-test score and P-value

# ANOVA

f\_val, p\_val = stats.f\_oneway(grouped\_test2.get\_group('fwd')['price'], grouped\_test2.get\_group('rwd')['price'], grouped\_test2.get\_group('4wd')['price'])

print( "ANOVA results: F=", f\_val)

ANOVA results: F= 67.95406500780399

So, we find that the built-in function and the analysis which I did through step by step way gives the same result.

For more details you can download a copy of my notebook which has the entire code from my GitHub link.