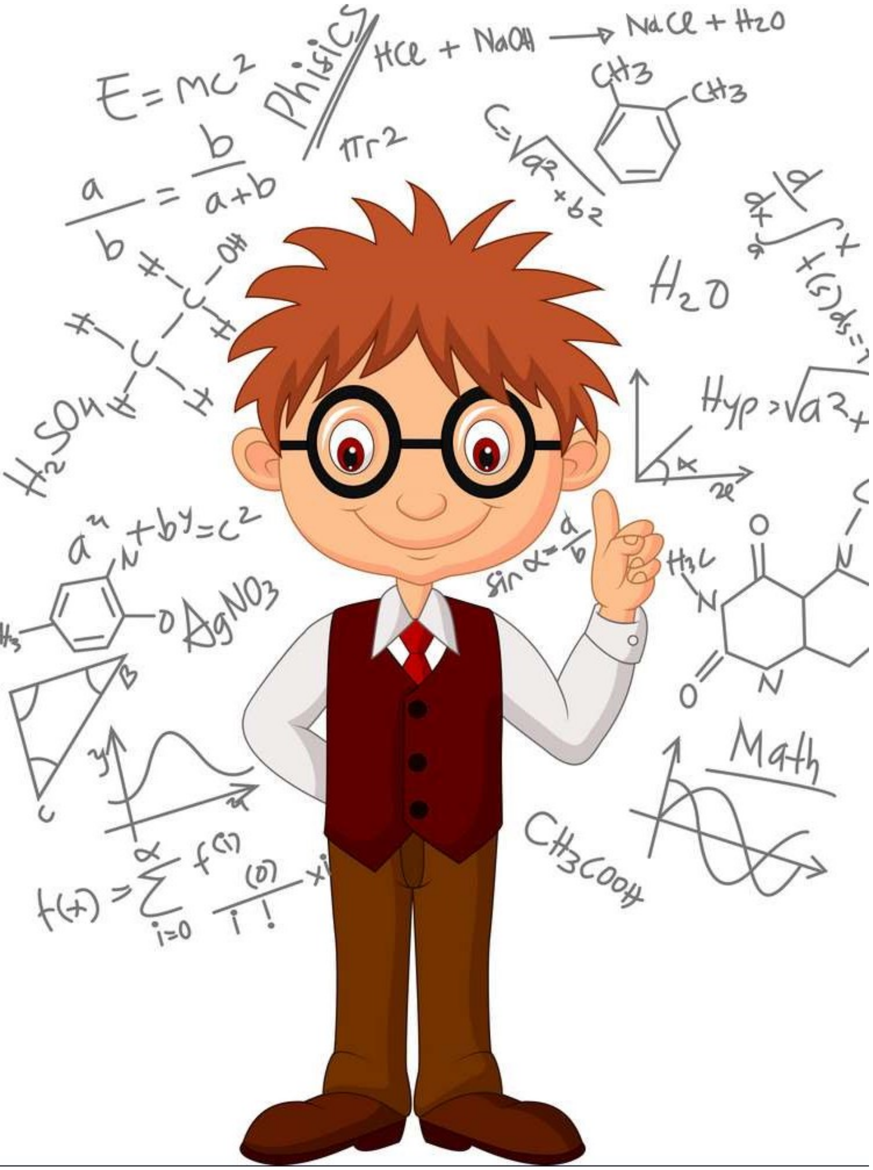
Statistics for Data Analytics

Lecturer: Marina Iantorno
E-mail: miantorno@cct.ie

CCT College Dublin
Computing • IT • Business



In today's class we will cover:

- ❑ One-Way ANOVA
- ❑ F-test
- ❑ Normality test

ANOVA

One of the most common techniques used in Inferential Statistics is Analysis of Variance, known as ANOVA. This test is about analysing the variability in a “y” variable and trying to understand where that variability is coming from.

ANOVA could be very useful when we want to compare several populations regarding some quantitative variable. It is particularly suitable for situations involving an experiment in which a certain treatment is applied (x) to subjects and the response is measured afterwards (y).

ANOVA

ANOVA test will evaluate the variation between the mean of different variables. The t-test for two populations allows us to test two means. If we worked with a hypothesis test (t) it would state something as follow:

H_0: $\mu_1 = \mu_2$

H_1: $\mu_1 \neq \mu_2$ or H_1: $\mu_1 > \mu_2$ or H_1: $\mu_1 < \mu_2$

With ANOVA we extend this idea to "k" different means from "k" different populations, but the only possibility for H_1 is $\neq$.

ANOVA

When we want to compare two different populations, a t-test would be enough, but when we want to analyse more than two populations we will be in ANOVA territory. The ANOVA procedure is built around a hypothesis called F-test, which compares how much the groups differs from each other compared to how much variability is within each group.

In other words, ANOVA is a parametric test and it is used to compare the means of three or more samples.

ANOVA

We will follow some steps in a one-way ANOVA:

1. Check the ANOVA conditions, using the data collected from each of the k populations.

2. Set up the hypothesis H0: $\mu 1 = \mu 2 = \ldots = \mu k$ versus the H1 Hypothesis which will state that at least one mean is different.

3. Conduct an F-test on the data and find the p-value.

4. Make your conclusions: If you reject H0 you conclude that at least one of the means is different from the others, otherwise you conclude that you did not have enough evidence to say that the means are different.

ANOVA

Let's study this with an example!

You are a financial advisor from an insurance company, and you manage three regions: East, Southwest and Northwest. You want to verify if the average charges of the three regions are the same or not.

A random sample was taken, and you can find it on Moodle. The file is called "insurance_data.csv".

ANOVA

1. Checking the ANOVA conditions

The conditions that have to be met in order to conduct the ANOVA test are:

- ➢The k populations are independent. In other words, their outcomes do not affect each other.
- ➢The k populations have a normal distribution.
- ➢The variances of the k normal distributions are equal.

ANOVA

1. Checking the ANOVA conditions

Independency of the variables: The first study that comes to our mind when we want to know the association between variables is the correlation test, and we would be right, but in this case, we are analysing numerical and categorical variables (charges is numerical and region is categorical). As the samples are randomly taken there is no reason to presume dependence from one variable to another one.

ANOVA

1. Checking the ANOVA conditions

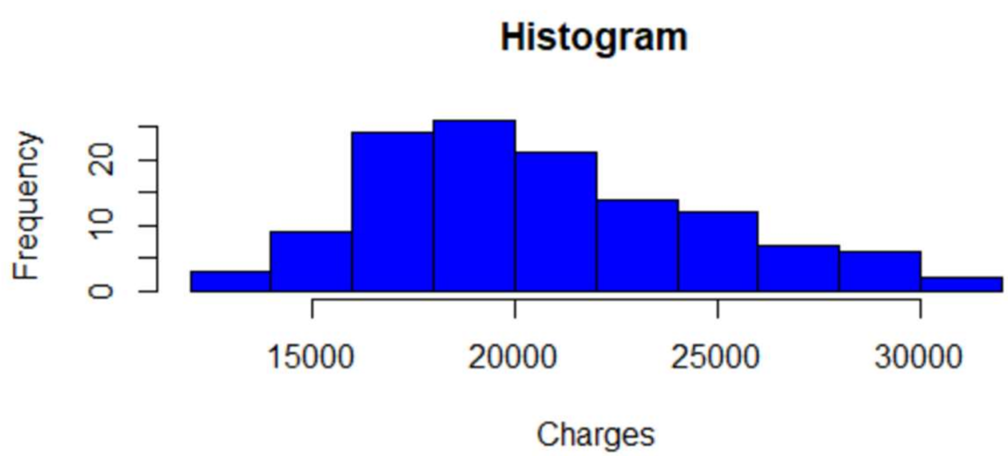
Normality of the distributions: The ANOVA test can be placed ONLY if the samples came from a normal distribution. To check this, we have 2 options:

- Plotting the data.
- Using the Shapiro Wilk test for normality (we will also study other tests during the semester).

ANOVA

1. Checking the ANOVA conditions

When we plot the data, we must do it with our numerical variable. Keep in mind that this will not be exactly symmetric, but even if it is a bit skewed, as log as you can see the bell, it should be ok.



ANOVA

1. Checking the ANOVA conditions

However, to be 100% sure, it is advisable to test the normality of the numerical variable using a normality test. The most common one is Shapiro Wilk test.

```
#Shapiro wilk test

stats.shapiro(dataset.charges[dataset.region == "east"])
```

```
ShapiroResult(statistic=0.9700243473052979, pvalue=0.10049082338809967)
```

```
#Shapiro wilk test

stats.shapiro(dataset.charges[dataset.region == "southwest"])
```

```
ShapiroResult(statistic=0.9592273235321045, pvalue=0.44794216752052307)
```

```
#Shapiro wilk test

stats.shapiro(dataset.charges[dataset.region == "northwest"])
```

```
ShapiroResult(statistic=0.9469977021217346, pvalue=0.10870692878961563)
```

Pvalue > 0.05 in all the categories, thus data is normally distributed

ANOVA

1. Checking the ANOVA conditions

When we use a Shapiro Wilk Test, we are stating a hypothesis test in which our premise is that the data came from a normal distribution. See below:

H_0 : data came from a normal distribution

H_1 : data did not come from a normal distribution

We always consider as $\alpha = 0.05$.

ANOVA

1. Checking the ANOVA conditions

Equality of the variances: We have two methods to check if the variances are equal or not.

Method 1: F-test

This test will be a Hypothesis test that will state that the variances are equal as a null hypothesis and that the variances are not equal as alternative hypothesis.

ANOVA

1. Checking the ANOVA conditions

Step 1: Hypothesis

$H0 : \sigma 1 = \sigma 2$

$H1 : \sigma 1 \neq \sigma 2$

Let’s say that Variable 1 belongs to East region and Variable 2 belongs to the Southwest region.

ANOVA

1. Checking the ANOVA conditions

Step 2: Formula

We will use F-Snedecor

$$F = \frac{\text{Larger Sample Variance}}{\text{Smaller Sample Variance}}$$

This Distribution has degrees of freedom (v).

V1: n2 – 1 → Southwest → V1 = 22

V2: n1 – 1 → East → V2 = 67

Step 3: Critical values

We will place this test to the right always, but still, we must divide our significance level in 2. Let's check this value in Probabilities distribution, always looking for P(x> x) = 0.025

Fc = 1.88


```
dataset['region'].value_counts()
```

```
east         68
northwest    33
southwest    23
```

ANOVA

1. Checking the ANOVA conditions

Step 4: Decision Rule

I reject H0 if $F > 1.88$

I accept H0 if $F < 1.88$

Step 5: Calculation of F

S East = 4238.58 → 4238.58^2

S South = 3239.53 → 3239.53^2

Step 6 : Result of the test

$F < F_c$ therefore I accept H_0.

Step 7: Conclusion

The variance of the regions are equal.

$$F = \frac{4238.58^2}{3239.53^2} = 1.71$$

ANOVA

1. Checking the ANOVA conditions

If we had to do this with many variables, we would need a lot of time. Let’s Python check the homogeneity of the variances between those that are in my analysis. We will do it using the Levene test. This will also run a Hypothesis Test in which we will start saying that the variances are equal.

H_0 : The variances between the regions are equal

H_1: The variances between the regions are not equal

We will get an outcome and we will analyse the result by looking at the p-value to accept or reject the hypothesis.

ANOVA

1. Checking the ANOVA conditions

```
levene(east, south, north, center = 'mean')
```

```
LeveneResult(statistic=0.9081132811476632, pvalue=0.40601496082599176)
```

In this case, p-value is greater than alpha, then we accept the null hypothesis and therefore we can say that the variances are equal.

ANOVA

1. Checking the ANOVA conditions

We verified that:

- Variables are independent.
- They come from a normal distribution.
- There is no difference between the variances.

Let’s move on!

ANOVA

2. Set up the hypothesis H0: $\mu 1 = \mu 2 = \ldots = \mu k$ versus the H1 Hypothesis which will state that at least one mean is different from the rest.

Here we will test if all population means can be deemed equal to each other. H_0 for ANOVA will always state that the means are equal, and the alternative Hypothesis will state that at least two of the means are different.

H_0: $\mu_{east} = \mu_{southwest} = \mu_{northwest}$

H_1: At least 2 μ are different.

ANOVA

3. Conduct an F-test on the data and find the p-value.

As ANOVA is the Analysis of Variance, we need to breakdown the variance into sum of squares.

Variance is the average squared deviation (difference) of a data point from the distribution mean.

Take the distance of each data point from the mean, square each distance, add them together, and then find the average.

Take out the “Find the average” part and we are left with just the SUM OF SQUARES (SS).

SS is variance without finding the average of the sum of the squared deviations.

ANOVA

3. Conduct an F-test on the data and find the p-value.

Let's define a couple of things to make it clearer:

SST or SSC = Sums of squares for treatment. It is the variability between the groups.

SSE = Sums of squares for error. It is the variability within the groups.

SSTO = Sums of total squares. SST + SSE

MST = Mean sums of squares for treatments. It measures the mean variability between the different treatments.

MSE = Mean sums of squares for error. It measures the mean within-treatment variability.

ANOVA

3. Conduct an F-test on the data and find the p-value.

Some formulas:

$$MSC = \frac{SSC}{DF\ (c-1)}$$

$$MSE = \frac{SSE}{DF\ (N - c)}$$

- N = total number of observations
- C = number of columns
- Df = degrees of freedom

ANOVA

3. Conduct an F-test on the data and find the p-value.

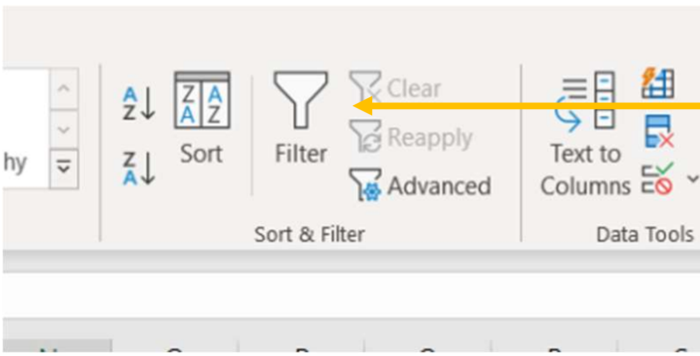
Let’s use Excel!

First, make a copy of your dataset and save it as an Excel file. Once you are done, hide the columns that are not relevant to this study.

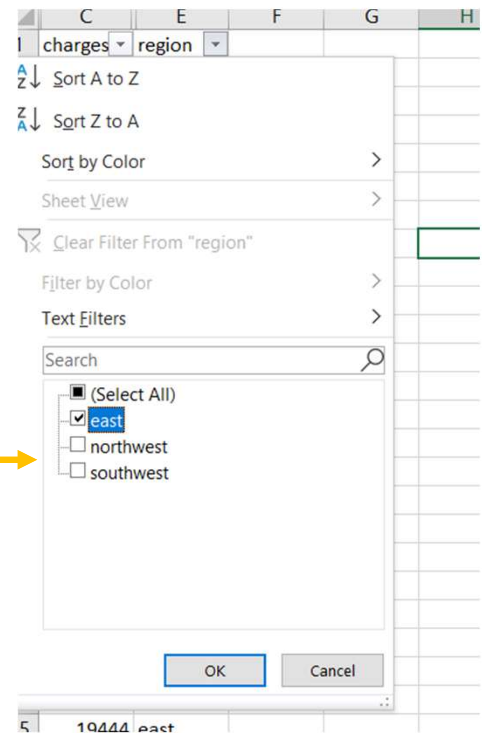
	C	E	F	G
1	charges	region		
2	16885	southwest		
3	27809	east		
4	23568	southwest		
5	23245	east		
6	14712	northwest		
7	17663	east		
8	16578	east		
9	21099	northwest		
10	30185	east		
11	22413	east		
12	15821	southwest		
13	30942	east		
14	17560	northwest		
15	19108	east		
16	17081	southwest		
17	18972	east		
18	20746	northwest		
19	19965	east		
20	21224	east		
21	15518	east		
22	21349	northwest		
23	20984	east		

ANOVA

3. Conduct an F-test on the data and find the p-value.



Apply the filter here for the two columns that are on the sheet



Filter by region and create a new table with your data

ANOVA

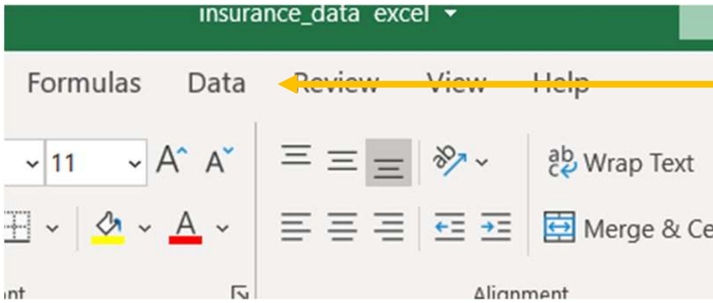
3. Conduct an F-test on the data and find the p-value.

H	I	J	K	L
east	northwest	south		
27809	14712	16885		
23245	21099	23568		
17663	17560	15821		
16578	20746	17081		
30185	21349	16298		
22413	17353	13845		
30942	24394	18608		
19108	18034	22144		
18972	28950	25382		
19965	26109	17942		
21224	28869	21082		
15518	23807	23307		
20984	17469	19041		
19516	25679	18259		
19444	17749	24520		
29523	18311	17496		
12829	15818	19933		
17085	29331	19200		
24870	21774	25309		
17180	20010	19023		
22332	23967	22479		

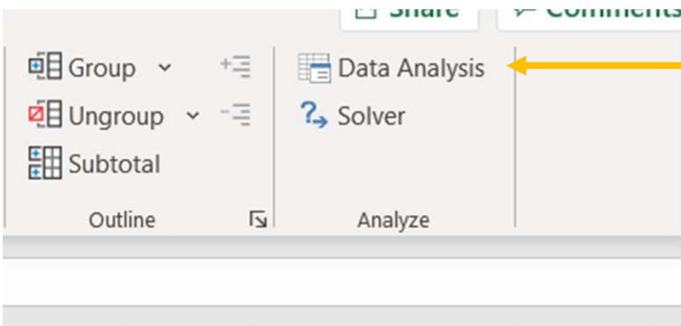
Your new table should look like this

ANOVA

3. Conduct an F-test on the data and find the p-value.



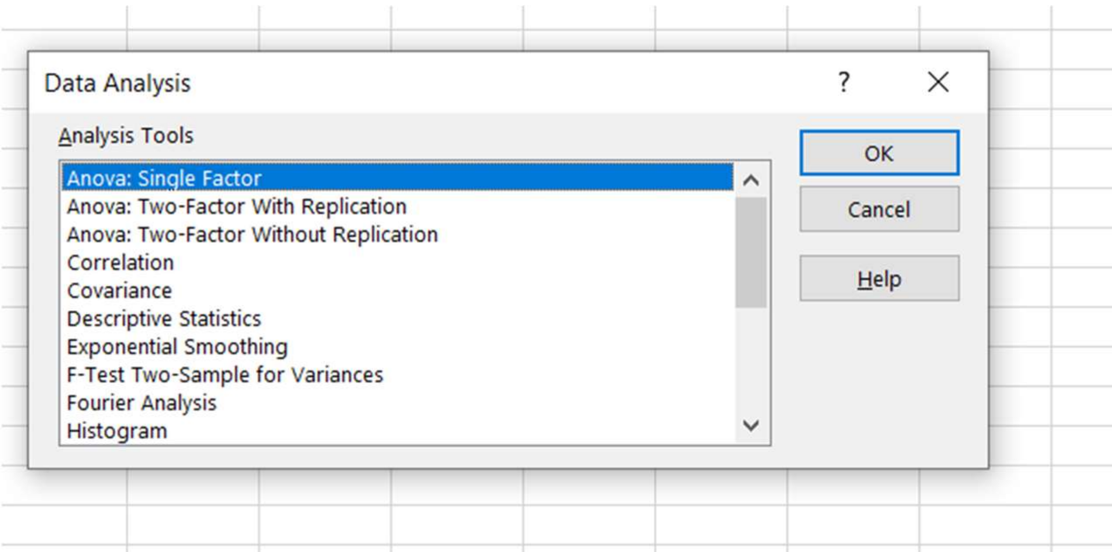
Now go to Data



Data Analysis

ANOVA

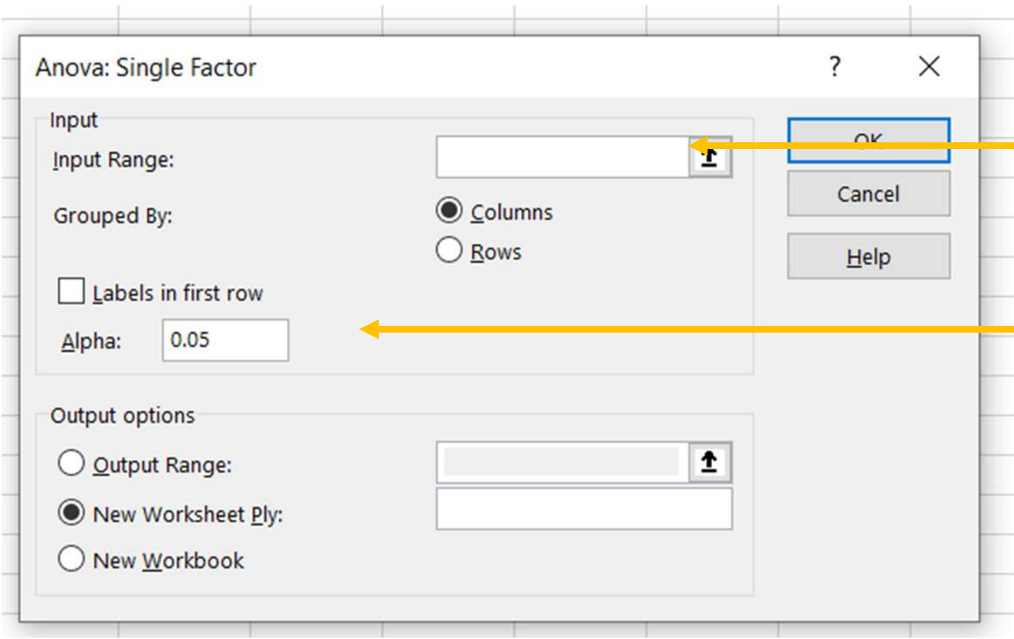
3. Conduct an F-test on the data and find the p-value.



Click the option Anova: Single Factor

ANOVA

3. Conduct an F-test on the data and find the p-value.



The input range will be the columns

Ensure your significance level

ANOVA

3. Conduct an F-test on the data and find the p-value.

Let’s see our results in the Excel sheet.

ANOVA	
Source of Variation	*SS*
Between Groups	43969820.7
Within Groups	1993463383
Total	2037433204

SST

SSE

SSTO = SSE + SST

Let’s do some Math!

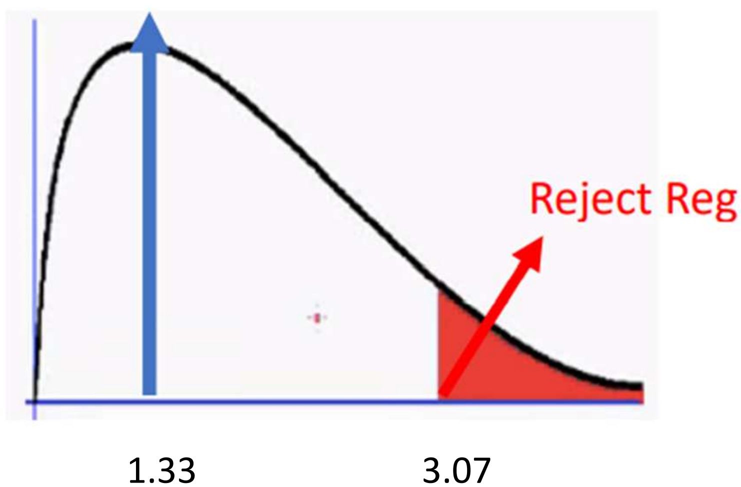
MSC = 43969820.7 /2 = 21984910.35

MSE = 1993463383/121 = 16474903.99

F = 21984910.35 /16474903.99 = 1.334448465

ANOVA

3. Conduct an F-test on the data and find the p-value.



V1: columns – 1 → 3 – 1 = 2
V2: N – columns → 124 – 3 = 121
α = 0.05

We can see this and the previous calculations on the Excel sheet

ANOVA

3. Conduct an F-test on the data and find the p-value.

V1 MSC

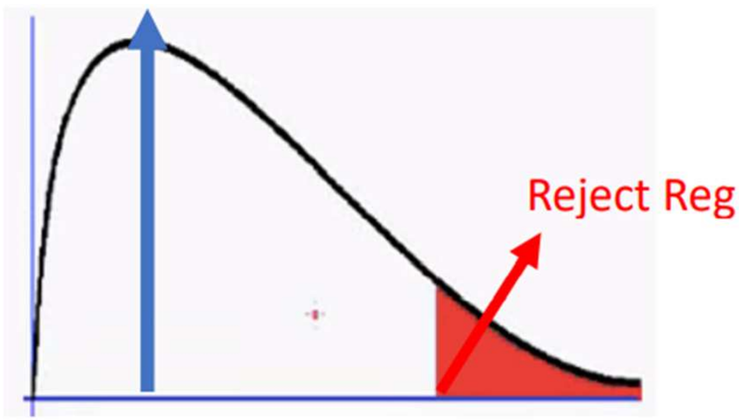
ANOVA

Source of Variation	SS	df	MS	F	P-value	F crit
Between Groups	43969820.7	2	21984910.35	1.334448	0.26715	3.07114
Within Groups	1993463383	121	16474903.99			
Total	2037433204	123				

V2 MSE

ANOVA

3. Conduct an F-test on the data and find the p-value.



1.33 3.07

F < Fc AND p-value > 0.05 → I accept H_0

V1: columns – 1 → 3 – 1 = 2
V2: N – columns → 124 – 3 = 121
$\alpha = 0.05$

ANOVA

3. Conduct an F-test on the data and find the p-value.

We can also use Python to find the F of the test and the p-value.

```
#ONE-WAY ANOVA
model = ols('charges~region', data = dataset).fit()
aov = sm.stats.anova_lm(model, type=2)
print(aov)
```

```
             df        sum_sq       mean_sq         F   PR(>F)
region      2.0  4.396982e+07  2.198491e+07  1.334448  0.26715
Residual  121.0  1.993463e+09  1.647490e+07       NaN      NaN
```

ANOVA

3. Make your conclusions

Here is when we lead to an interpretation of our analysis

I accept H_0, therefore there is no evidence to think that the means are not equal

THAT'S ALL FOR TODAY

THANK YOU