```
ANOVA (ANalysis Of VAriance)
ANOVA is used to compare differences of means among more than 2 groups. It does this by
looking at variation in the data and where that variation is found (hence name).
specifically , ANOVA compares the amount of variation between groups with the amount of
variation within groups

. Null hypothesis, typically is that , all means are equal.
. The independent variables are categorical.
.Dependent variables are continue
```

In [8]:

```
import pandas as pd
hyp=pd.read_csv('Semester marks1.csv')
hyp
```

Out[8]:

	Student_ID	Semester	Total_Marks	Unnamed: 3	Unnamed: 4	Unnamed: 5	Unnamed: 6
0	1.0	1.0	365.0	NaN	NaN	NaN	NaN
1	2.0	3.0	376.0	NaN	NaN	NaN	NaN
2	3.0	1.0	411.0	NaN	NaN	NaN	NaN
3	4.0	3.0	436.0	NaN	NaN	NaN	NaN
4	5.0	2.0	366.0	NaN	NaN	NaN	NaN
...	...	...	...	...	...	...	...
196	197.0	3.0	330.0	NaN	NaN	NaN	NaN
197	198.0	1.0	311.0	NaN	NaN	NaN	NaN
198	199.0	1.0	358.0	NaN	NaN	NaN	NaN
199	200.0	2.0	468.0	NaN	NaN	NaN	NaN
200	NaN	NaN	NaN	NaN	NaN	NaN	NaN

201 rows × 7 columns

As per ANOVA we need to find continues data based on the categorical variable

. Total _marks is continious variable and . semester is categorical variable

Ho is average marks of each semester are almost equal Ha is all are not equal

In [9]:

```
import numpy as np
import statsmodels.formula.api as sm
mod=sm.ols('Total_Marks ~ Semester', data=hyp).fit()
mod
```

Out[9]:

```
<statsmodels.regression.linear_model.RegressionResultsWrapper at 0x2fb0795a9
0>
```

In [11]:

```
import statsmodels.api as k
aov_table=k.stats.anova_lm(mod,typ=1)
aov_table
```

Out[11]:

	df	sum_sq	mean_sq	F	PR(>F)
Semester	1.0	291.846746	291.846746	0.102572	0.749102
Residual	198.0	563366.873254	2845.287239	NaN	NaN

Let's see manually if the average marks of each semester are close to each other

In [12]:

```
# semester 1 Average marks
print('1st Sem avg marks =',np.ceil(hyp['Total_Marks'][hyp['Semester']==1].mean()))

# semester 2 Average marks
print('2nd Sem avg marks =',np.ceil(hyp['Total_Marks'][hyp['Semester']==2].mean()))

# semester 3 Average marks
print('3rd Sem avg marks =',np.ceil(hyp['Total_Marks'][hyp['Semester']==3].mean()))
```

```
1st Sem avg marks = 403.0
2nd Sem avg marks = 393.0
3rd Sem avg marks = 406.0
```

Realtime research on human body-

if different types of drinks (coffee,water,milk,soda etc) have the same effect (reaction time in the morning) on human body or not?

Null hypothesis- All the drinks average reaction time on human body is same Alternate hypothesis- All the drinks average reaction time on human body is NOT same