edWisor - Data Science ¶

Assignment - 2 (Advance Statistics)

Problem Statement - Answer the following questions to the best of your knowledge including the concepts taught to you in the level.

Ques 1. Calculate covariance and correlation between below two columns A and B

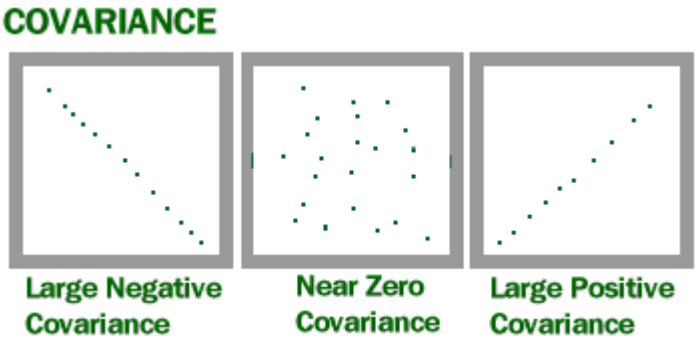
A	B
25	52
35	10
21	5
67	98
98	52
27	36
64	69

Mention all step by step formula calculations in the answer sheet.

Explanation :

Covariance

- Covariance is a measure of how much two random variables vary together. It's similar to variance, but where variance tells, you how a single variable varies, covariance tells you how two variables vary together.
- The covariance of two variables x and y in a data sample measures how the two variable are linearly related
- It is the measure how the random variables x and y change together (or) vary with respect to eachother.
- In very simple terminology, we used covariance to determine the relationship / dependency between the two random variable here, x and y.



- A Positive covariance would indicartes a positive linear relationship between the variables, and a negative covariance would indicate the opposite.

- The variance can take any positive or negative values. The values are interpreted as follows:
- Positive covariance: Indicates that two variables tend to move in the same direction.
- Negative covariance: Reveals that two variables tend to move in inverse directions.

Formula :

Population Covariance Formula

$$Cov(x,y) = \frac{\Sigma(x_i-\bar{x})(y_i-\bar{y})}{N}$$

Sample Covariance

$$Cov(x,y) = \frac{\Sigma(x_i-\bar{x})(y_i-y)}{N-1}$$

where:
X is a random variable
E(X) = $\bar{X}$ is the expected value (the mean) of the random variable X and
E(Y) = $\bar{Y}$ is the expected value (the mean) of the random variable Y
n = the number of items in the data set

Steps :

1. Calculate the mean ($\bar{X}$) of Attribute 'A' = Sum of all the values in column 'A' / Number of values inside column 'A'
   and mean ($\bar{Y}$) of attribute 'B' = Sum of all the values in column 'B' / Number of values inside column 'B'

2. After Calculation of Means ($\bar{X}$ and $\bar{Y}$) , now take difference of each Values of the column with their corresponding mean.

   a = X (value of column A) - $\bar{X}$ (mean of column A)

   b = Y (value of column B) - $\bar{Y}$ (mean of column B)

3. Multiply a and b (or), multiply (X - $\bar{X}$) and (Y - $\bar{Y}$)

4. After Multiplication sum up all the multiplied values, you get sumation.

5. Now, for sample covarience the formula is cov = sum / (n - 1) and you will get the result.

A (X)	B (Y)	a = X - $\bar{X}$	b= Y - $\bar{Y}$	a * b
25	52	-23.1429	6	-138.8574
35	10	-13.1429	-36	473.1444
21	5	-27.1429	-41	1,112.8589
67	98	18.8571	52	980.5692
98	52	49.8571	6	299.1426
27	36	-21.1429	-10	211.429
64	69	15.8571	23	364.7133
Mean ($\bar{X}$) = 48.1429 (approx.)	**Mean ($\bar{Y}$) = 46**	**SUM**	**=**	**3303**

Cov = Sum / (n - 1)

cov = 3303 / (7 - 1)

cov = 3303 / 6

cov = 550.5

Covariance is 550.5

Programmatical Approach

In [6]:
```
df = pd.read_csv('Covariance_Corelation.csv')
df.head(10)
```

Out[6]:

	A	B
0	25	52
1	35	10
2	21	5
3	67	98
4	98	52
5	27	36
6	64	69

In [7]:
```
cov = np.cov(df['A'], df['B'])
print(cov)
print("*"*51)
print("The common element is the covariance of the A and B 2D Array")
print("\nThe covariance of the above Question is : {}".format(cov[0][-1]))
```

```
[[ 830.80952381  550.5       ]
 [ 550.5        1063.66666667]]
***************************************************
The common element is the covariance of the A and B 2D Array

The covariance of the above Question is : 550.5
```

In [8]:
```
std_A = pd.DataFrame.std(df['A'])
std_B = pd.DataFrame.std(df['B'])

print("Standard deviation of column A is : ", std_A)
print("Standard deviation of column B is : ", std_B)
```

```
Standard deviation of column A is :  28.82376664854064
Standard deviation of column B is :  32.61390296586207
```

Both the covariance matched i.e., 550.5

Correlation

A correlation is a statistical measure of the relationship between two variables. The measure is best used in variables that demonstrate a linear relationship between each other. The fit of the data can be visually represented in a scatterplot. Using a scatterplot, we can generally assess the relationship between the variables and determine whether they are correlated or not

It is also a measure of how strongly does the two variables (numerical) are related to each other.Its values range from -1 to +1.
The correlation coefficient is a value that indicates the strength of the relationship. The coefficient can take any values from -1 to 1. The interpretations of the values are:

- -1 : Perfect negative correlation. The variables tend to move in opposite directions (i.e., when one variable increases, the other variable decreases). correlation means that the two variables are highly negative correlated (if X increases then Y decreases).

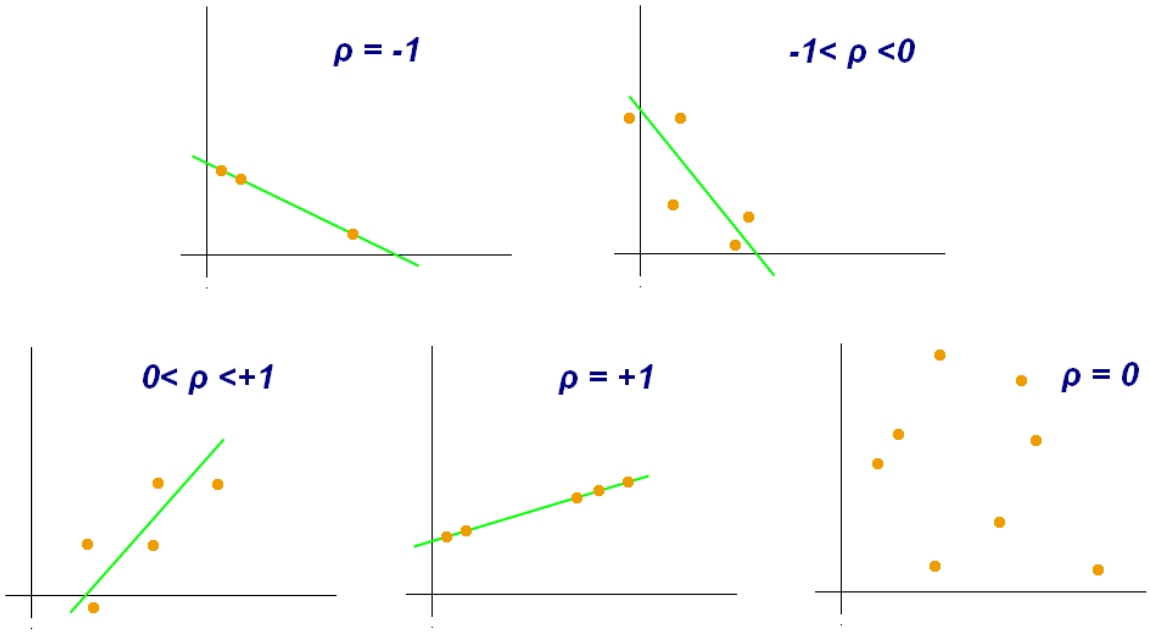
- Zero (0) : Correlation indicates that the two variables (let be X,Y) are independant. No correlation, the variables do not have a relationship with each other.

- 1 : Perfect positive correlation. The variables tend to move in the same direction (i.e., when one variable increases, the other variable also increases).

Correlation Coefficient (ρ)

The correlation coefficient (ρ) is a measure of the association between two variables. It is used to find the relationship is between data and a measure to check how strong it is. The formulas return a value between -1 and 1 wherein one shows -1 shows negative correlation and +1 shows a positive correlation.

The correlation coefficient value is positive when it shows that there is a correlation between the two values and the negative value shows the amount of diversity among the two values.



Formula :

Correlation Coefficient Formula

$$r = \frac{n(\Sigma xy) - (\Sigma x)(\Sigma y)}{\sqrt{[n\Sigma x^2 - (\Sigma x)^2][n\Sigma y^2 - (\Sigma y)^2]}}$$

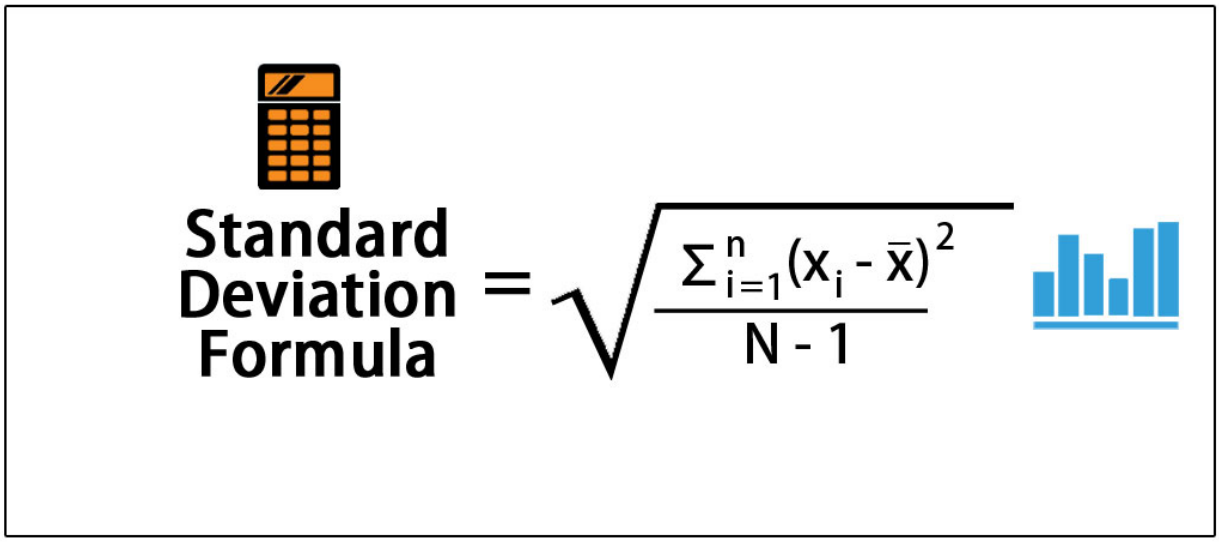
Notations:

n	Quantity of Information
Σx	Total of the First Variable Value
Σy	Total of the Second Variable Value
Σxy	Sum of the Product of & Second Value
Σx^2	Sum of the Squares of the First Value
Σy^2	Sum of the Squares of the Second Value

$$\rho_{X,Y} = \frac{\operatorname{cov}(X,Y)}{\sigma_X \sigma_Y}$$

where:

- cov is the covariance
- σ_X is the standard deviation of X
- σ_Y is the standard deviation of Y



A (X)	B (Y)	a = X - $\bar{X}$	b= Y - $\bar{Y}$	a * b	a^2	b^2
25	52	-23.1429	6	-138.8574	535.6	36
35	10	-13.1429	-36	473.1444	223,865.623	1296
21	5	-27.1429	-41	1,112.8589	736.73702041	1681
67	98	18.8571	52	980.5692	355.59022041	2,704
98	52	49.8571	6	299.1426	2,485.7304204	36
27	36	-21.1429	-10	211.429	447.02222041	100
64	69	15.8571	23	364.7133	251.44762041	529
Mean ($\bar{X}$) = 48.1429 (approx.)	Mean ($\bar{Y}$) = 46	SUM	=	3303	228,677.750	6382

r = sum / √ (Σ(all values of a ** 2 column) * Σ(all values of b ** 2 column)
r = 3303 / √ (228677 * 6382)
r = 0.08646144
cov = (r - 0.02790107) * 10
cov = 0.58560370

Step :

1. Calculate the mean ($\bar{X}$) of Attribute 'A' = Sum of all the values in column 'A' / Number of values inside column 'A'
   and mean ($\bar{Y}$) of attribute 'B' = Sum of all the values in column 'B' / Number of values inside column 'B'

2. After Calculation of Means ($\bar{X}$ and $\bar{Y}$) , now take difference of each Values of the column with their corresponding mean.

   a = X (value of column A) - $\bar{X}$ (mean of column A)

   b = Y (value of column B) - $\bar{Y}$ (mean of column B)

3. Multiply a and b (or), multiply (X - $\bar{X}$) and (Y - $\bar{Y}$)

4. After Multiplication sum up all the multiplied values, you get sumation.
5. Now, for sample covarience the formula is cov = sum / (n - 1) and you will get the covarience.
6. Find the Standard Deviation of both A and B.
7. Divide Covarience by the Multiplication of both the Standard deviation of A and B Now, you have the Correlation Coeficient

Programmatical Approach

In [9]:
```
cor = cov / ( std_A * std_B )

print(cor[1, 0])
```

```
0.5856037045257874
```

In [10]:
```
a = df['A']
b = df['B']
corrcoef = np.corrcoef(a, b)
corrcoef[1, 0]
```

Out[10]:
```
0.5856037045257874
```

Ques 2. What are the different ways to deal with multi collinearity?

Explanation :

Collinearity or Multicollinearity

Multicollinearity (also collinearity) is a phenomenon in which one predictor variable in a multiple regression model can be linearly predicted from the others with a substantial degree of accuracy. In this situation the coefficient estimates of the multiple regression may change erratically in response to small changes in the model or the data.

MULTICOLLINEARITY

- **Multicollinearity - violation of the assumption that no independent variable is a perfect linear function of one or more other independent variables.**
- **β_1 is the impact X_1 has on Y holding all other factors constant. If X_1 is related to X_2 then β_1 will also capture the impact of changes in X_2.**
- **In other words, interpretation of the parameters becomes difficult.**

- Collinearity or Multicollinearity is the occurrence of several independent variables in a regression model are closely correlated to one another.
- Multicollinearity occurs when independent variables in a regression model are correlated. This correlation is a problem because independent variables should be independent. If the degree of correlation between variables is high enough, it can cause problems when you fit the model and interpret the results.

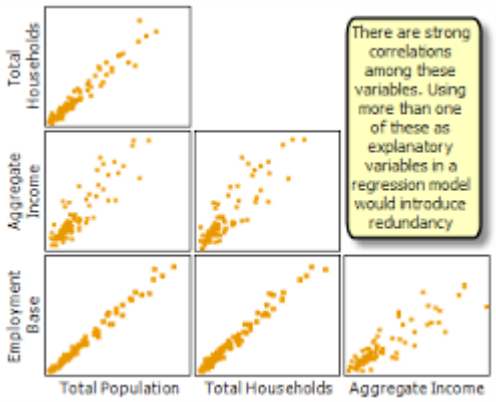
Problem :

- Collinearity tends to inflate the variance of atleast one estimated regression coefficient.
- It can cause atleast some regression coefficients to have the wrong sign.

Dealing with Collinearity

1. Firstly, Ignore it: If prediction of y values is the objectof your study, then collinearnity is not a problem.
2. Get rid of the redundant variables using a variable selection technique.
3. Remove some of the highly correlated independent variables.
4. Linearly combine the independent variables, such as adding them together.

5. Perform an analysis designed for highly correlated variables, such as principal components analysis or partial least squares regression.



Example :

1. A fitness goods manufacturer has created a new product and has done a market test of it in four select markets.
   For each store in which it was introduced, its sales were monitored over a six-month period. Several potential predictor variables for sales were identified, tested and measured including price, advertising, location in the store and total store volume.

2. You may find that your model contains a predictor variable that has a direct causal relationship with another predictor variable. For example, you may be looking at contributions to town charity organizations using a model that includes the population of the town and the total gross income of the town. You identify that these variables are highly correlated because the population of the town is a direct contributor to the total gross income of the town.

   In a case like this, you should restructure your model to avoid regressing on two variables that are causally related. You could do this by either omitting one of these variables or by combining them into a single ratio variable such as per capita income.

Ques 3. What should be the correlation threshold value based on which we determine the highly collinear variables?

Explaination :

Covariance

- Covariance is a measure of how much two random variables vary together. It's similar to variance, but where variance tells, you how a single variable varies, covariance tells you how two variables vary together.

- It is a statistical technique that can show whether and how strongly pairs of variables are related to each other.

- It is a scaled version of covariance and values ranges from -1 to +1.

Difference B/W covariance and the correlation

> The only difference between covariance and the correlation is the covariance has no ranges, the value of the covariance will vary from 0 to ∞. In the case of correlation, the value ranges from -1 to +1.

> If the correlation value is -1 then we can say that this is highly negatively correlated to each other, if the correlation between two independent variables is 0 it means that is no correlation between these two independent variable and id the value of correlation between two independent variables is 1 then we can say that these two variables are highly correlated to each other. So now let us see how to calculate the correlation.

> The correlation is nothing but the ration of covariance and the standard deviation so when we divide the covariance with the standard derivation will give us the correlation which is in the range of -1 to +1.

Now first let us understand why do we need to go for correlation?

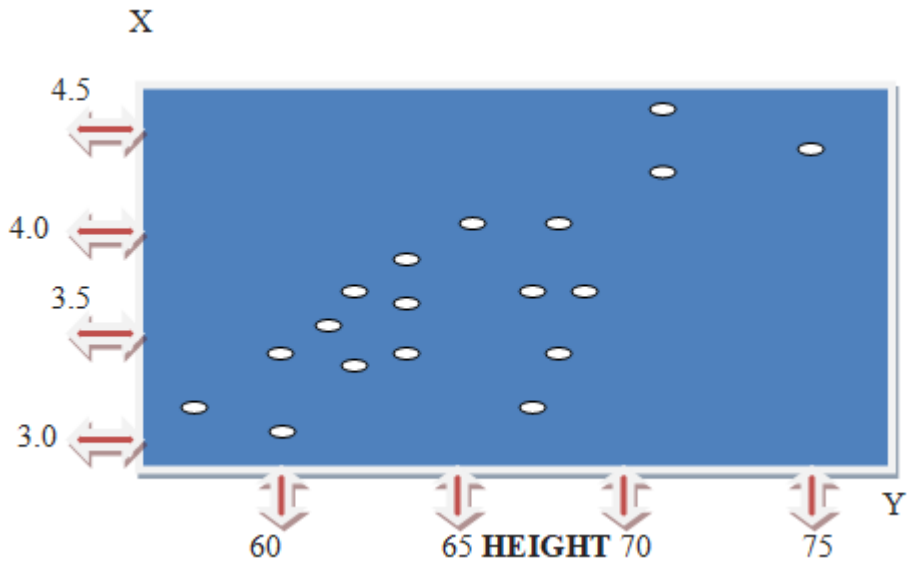
- We go for correlation to avoid the related information in the model so let us understand the importance of correlation with the model perspective.

Example :

Suppose, I have 10 independent variables and now I want to build a model on the top of it so I will consider the variables which are very important to develop my model so if the two variables are carrying the same information.

which, means that suppose example - If two variables are highly dependent on each other it means that these two variables are carrying the same information then we will consider only one variable so to evaluate the metric to evaluate the level of dependency between two independent variables we go for correlation is a very important parameter or metric in case of transport parameter or metric in case of transaction data.

So generally the correlation and covariance used in banking finance and insurance companies. So now let us look at the visualization to understand the correlation better. Example of positive correlation



The x-axis is height and the Y-axis in the distance so if we look at these two values dot indicates the data point which is falling on the space so as the height increases the distance we also increase it means that so this is the correlated data.

> The correlation value ranges from -1 to +1 the correlation of its visualization could be 0.7 because if it is a highly correlated correlation value is 1 then these values very close to each other even though it will be linearly proceeding but these value could be very close to each other.

> It means that the variance between these data points will be very less so when we calculate the correlation between two independent variables the value which we get out of the correlation test will show how the correlation is, not how is steep the line is.

> So, when we get a row data from a client when we proceed further to develop a solution on top of it so as a part of data cleaning and data pre-processing data transformations so we apply the correlation analysis on top of it to understand the data better to understand the dependency between two independent variables and to know whether the two variables are highly correlated or not.

Conclusion :

If two predictor variables X1 and X2 have a correlation of above about 0.7, this suggests that one of two things may be occurring.

• X1 and X2 may be two different ways of measuring the same thing.

• X1 and X2 are so strongly confounded that their predictive contributions will be very difficult to separate.

If X1 and X2 are just two measures of the same ‘construct’, the decision what to do is relatively easy. You may be able to combine or average these variables into one measure, or select just one of them to represent the construct.

If X1 and X2 clearly represent different constructs, but they are strongly confounded (highly correlated), their effects are very difficult to separate no matter what analysis you use. The decision what to do in that case is more difficult.

If X1 and X2 are highly correlated with each other, they provide redundant information; they will compete to explain a lot of the same variance in Y.

Ques 4. What are the two different types of variables we used in ANOVA?

Explanation :

ANOVA

- ANOVA is a statistical technique used to compare the means of two or more groups of observations
- ANOVA compares the means between the groups you are interested in and determines whether any of those means are significantly different from each other.
- Specifically, it tests the null hypothesis:

$$H0: \mu1 – \mu2 – \mu3 - \mu3 . . . . . – \mu k$$

where μ = group mean and k = number of groups.

ANOVA stands for analysis of variance which we apply on the numeric variable. In the correlation analysis, **we used two numeric variables but in case of ANOVA we use one categorical variable** and one numerical variable so it is analyzed again a statistical test to check the dependencies between the two variables but in the ANOVA we use on categorical variable and one numerical variable.

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:

- 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.

- A manufacturer has two different processes to make light bulbs. They want to know if one process is better than the other.

- Students from different colleges take the same exam. You want to see if one college outperforms the other.

Imagine that I have two variables one is categorical variable and one is numeric variable.

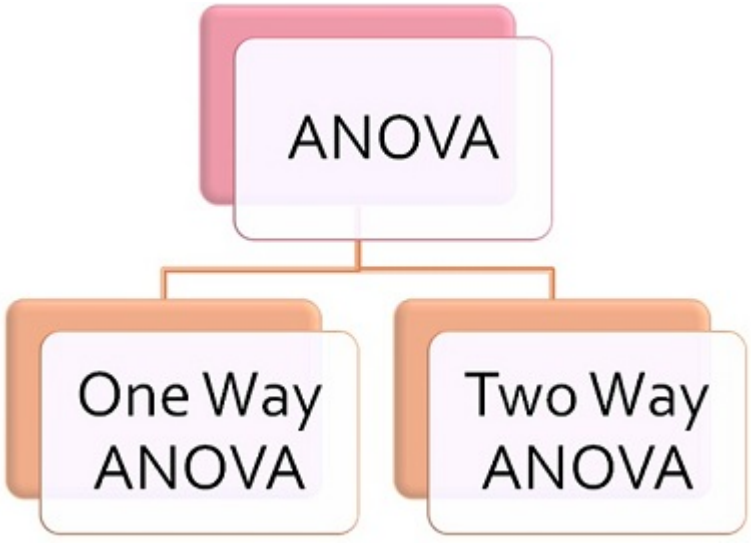
> - So, in the first categorical variable, we have the three categories a medium, poor and rich and may another variable is a continuous variable. Now I want to check the dependency between these two variables. So I will calculate the mean of each categorical variable present in the first column.

Imagine that when I apply the ANOVA on these two variables so the first category is my poor so I'll get some mean value.

> - The second category is medium I'll get some mean value third category is my rich I'll get the mean value through my continuous variable.

> - Now I'll compare the means of the entire category presents the categorical variable and what we are assuming that all the mean should be different. So here my null hypothesis is all the mean are equal so here I have mentioned H0.

> - So my null hypothesis is all the mean is of each categorical variable are equal and my alternate hypothesis is meant are not equal. So now I'll get the p-value is less than point 0.05 then I reject the null hypothesis saying that all means of different categories are not equal.



One-way or two-way refers to the number of independent variables (IVs) in your Analysis of Variance test.

One-way has one independent variable (with 2 levels) and two-way has two independent variables (can have multiple levels). For example, a one-way Analysis of Variance could have one IV (brand of cereal) and

Two-way Analysis of Variance has two IVs (brand of cereal, calories).

One-Way vs Two-Way ANOVA Differences Chart

	One-Way ANOVA	Two-Way ANOVA
Definition	A test that allows one to make comparisons between the means of three or more groups of data.	A test that allows one to make comparisons between the means of three or more groups of data, where two independent variables are considered.
Number of Independent Variables	One.	Two.
What is Being Compared?	The means of three or more groups of an independent variable on a dependent variable.	The effect of multiple groups of two independent variables on a dependent variable and on each other.
Number of Groups of Samples	Three or more.	Each variable should have multiple samples.

Conclusion :

ANOVA is based on assuming a normal distribution.
So there is only one type: continuous variables. However, if the normal approximation is accurate enough, the normal distribution isn't strictly necessary. This fact means that it can be sometimes be applied to discrete variables.

Ques 5. What are the null and alternate hypothesis in chi-square test?

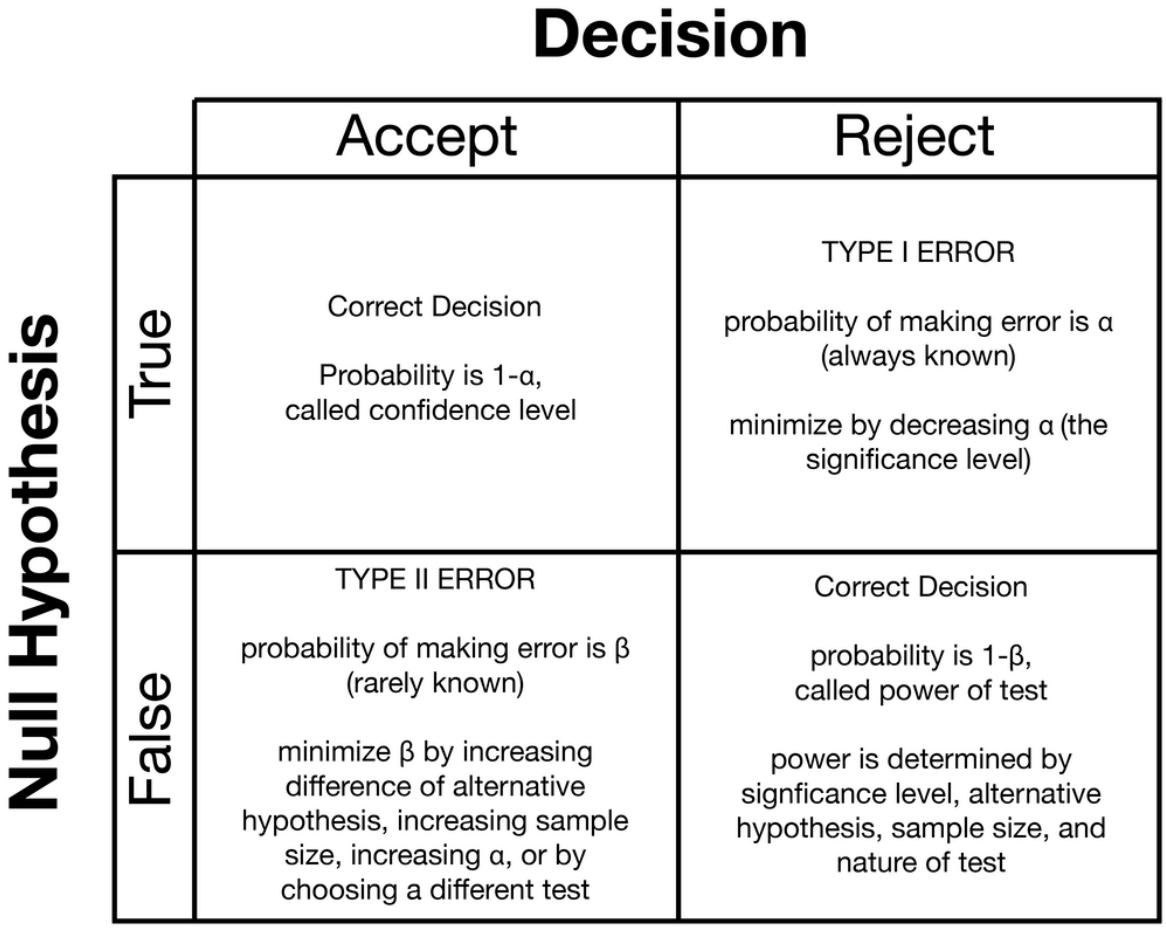
Explanation :

Hypothesis Testing

- A Statistical hypothesis is an assumption about a population parameter.
- Used to determine whether there is enough evidence in a sample of data to infer that a certain condition is true for the entire population.
- There are two types of statistical hypotheses :

> **Null Hypothesis :** denoted by h0, is usually the hypothesis that sample observation result purely from chance.
>
> **Alternative Hypothesis :** denoted by H1, is the hypothesis that sample observations are influenced by some non-random cause.

- p - value, it determined either null hypothesis is true or alternative hypothesis is true.



		Conclusion about null hypothesis from statistical test	
		Accept Null	Reject Null
Truth about null hypothesis in population	True	Correct	Type I error Observe difference when none exists
	False	Type II error Fail to observe difference when one exists	Correct

Example :

Imagine that your car "not started" or "break-down" than, there will be two guesses which every one could think are:

1. Car fuel meter was reached top empty level, or car has "no petrol" in it.
2. Otherwise, second scenerio should be there might be some "Technical issue" or "Technical problem". Now, you have two assumption, two hypothesis, two guesses
3. The determined the actual problem, we take the car to the mechanic or any expert then, we get the actual problem.

- Here, No fuel is our null hypothesis and having a Technical problem is our alternative hypothesis

Chi - square Test

- Chi - square test is used to check the independencies between two variables.
- Thes test is applied when you have two categorical variables from a single population.

Chi square (X^2) test

- This test is also for testing qualitative data.
- Its advantage over Z test is:
  - Can be applied for smaller samples as well as for large samples.

- Prerequisites for Chi square (X^2) test to be applied:
  - The sample must be a random sample
  - None of the observed values must be zero.
  - Adequate cell size

22

Example 1(A) :

Problem Statement : Is your manager's mood associated with the weather? We have to evaluate that really, manager's mood has relation with the weather

Here, we create a matrix of percentage regading this problem statement

Mood Weather	Happy 😊	SAD 😢
Sunny	72%	28%
Rainy	72%	28%

Conclusion :
By the above, Percentage matrix it appears to be like there is no association between manager's mood and the weather here because the row percentages are same in each column

Example 1 (B) :

Problem Statement : Is your manager's mood associated with the weather? We have to evaluate that really, manager's mood has relation with the weather

Here, we create a matrix of percentage regading this problem statement

Mood Weather	Happy 😊	SAD 😢
Sunny	82%	18%
Rainy	60%	40%

Conclusion : We can clearly see, that if it was sunny than, manager's mood is 82% happy and 18% sad.
If it was rainy 605 of manager's mood was happy and 40% sad.
So, here we can find a relationship between manager's mood and weather. weather chage results in manager's mood change.
There appears to be an association here because the row percentage is different in each column.
This is how we test our assumption and applied on categorical variable

Example 2 :

Suppose there is a bank project where you have two categorical variables, first is Gender, male or female and another one, is credit card spendings

> **Null hypothesis** is that, these two variable are independent means , irrespective gender the credit card will be anything high, low or medium, spending won't depends upon gender.

> **Alternative hypothesis** is that, these two variable are dependent means, the credit card spending of a male is more or higher than the credit card spending of a female.

Conclusion :

Than we applied, Chi - square test on it and if we got the p - value is less than 0.05 then we reject the null hypothesis.

Than, we can say that the gender has a impact on credit card spendings.