Chi-Square test

→ A chi-square (χ^2) statistic is a test that measures how a model compares to actual observed data. The data used in calculating a chi-square statistic must be random, raw, mutually exclusive, drawn from independent variables, and drawn from a large enough sample.

→ It is a measure of the difference between the observed and expected frequencies of the outcomes of a set of events or variables.

→ Chi-Square is useful for analyzing such differences in categorical variables, especially those nominal in nature.

→ χ^2 depends on the size of the difference between actual and observed values, the degrees of freedom, and the sample size.

→ It can be used to test whether two variables are related or independent from one-another.

→ It can also be used to test the goodness-of-fit between an observed distribution and a theoretical distribution of frequencies.

Formula of Chi-Square is :-

$$\chi^2_c = \sum \frac{(O_i - E_i)^2}{E_i}$$

where :

c = Degree of freedom
O = observed value(s)
E = Expected value(s)

Types of Chi- Square Tests :-

There are two commonly used Chi-Chi-Squared tests :-

- Chi - square Goodness of fit test.
- Chi - Square test of Independence.

	Chi-Square Goodness of Fit Test	Chi-Square Test of Independence
Number of variables	One	Two
Purpose of test	Decide if one variable is likely to come from a given distribution or not.	Decide if two variables might be related or not.
Example	Decide if bags of candy have the same number of pieces of each flavor or not.	Decide if movie goers' decision to buy snacks is related to the type of movie they plan to watch.
Hypothesis in example	H_0: Proportion of flavors of candy are the same. H_a: Proportions of flavors are not the same.	H_0: Proportion of people who buy snacks is independent of the movie type H_a: Proportion of people who buy snacks is different for different types of movies.
Theoretical Distribution	Chi-Square	Chi-Square

	Chi-Square Goodness of fit	Chi-Square Test of Independence
Degrees of freedom	Number of categories minus 1. • In our example, number of flavors of candy minus 1	Number of categories for first variable minus 1, multiplied by number of categories for second variable minus 1. • In our example, number of movie categories minus 1, multiplied by 1 (because each Purchase is a Yes/No variable and 2-1 = 1)

Assumptions of Chi-Square Test

1. Both variables are categorical
2. All observations are independent
3. Cells in the contingency table are mutually exclusive.
4. Expected value of cells should be 5 or in greater in at least 80%. of cells

* Variance :-

→ Variance is a measure of variability.

→ It is calculated by taking the average of squared deviations from the mean.

$$\text{Variance}(\sigma^2) = \frac{\sum_{i=1}^{N} (x_i - \bar{x})^2}{N}$$

→ It tells the degree of spread in our dataset. The more spread the data, the larger the variation is in relation to the mean.

* Standard Deviation

→ It is the average amount of variability in the dataset. It tells us, on average, how far each value lies from the mean.

→ A high standard deviation means that values are generally far from the mean, while a low standard deviation indicates that values are clustered close to the mean.

→ It is the square root of the variance.

Formula

$$\text{Standard Deviation}(\sigma) = \sqrt{\frac{\sum_{i=1}^{N} (x_i - \bar{x})^2}{N}}$$

Covariance

→ It is a measure of the relationship between two random variables and to what extent, they change together.

→ It defines the changes between the two variables, such that change in one variable is equal to change in another variable.

Types:-

1. Positive Covariance:-
   - If the covariance for any two variables is positive, that means, both the variables move in the same directions.
   - Here the variables show similar behaviour.

2. Negative Co-variance

If the covariance for any two variables is negative, that means, both the variables move in opposite direction.

Population Covariance Formula

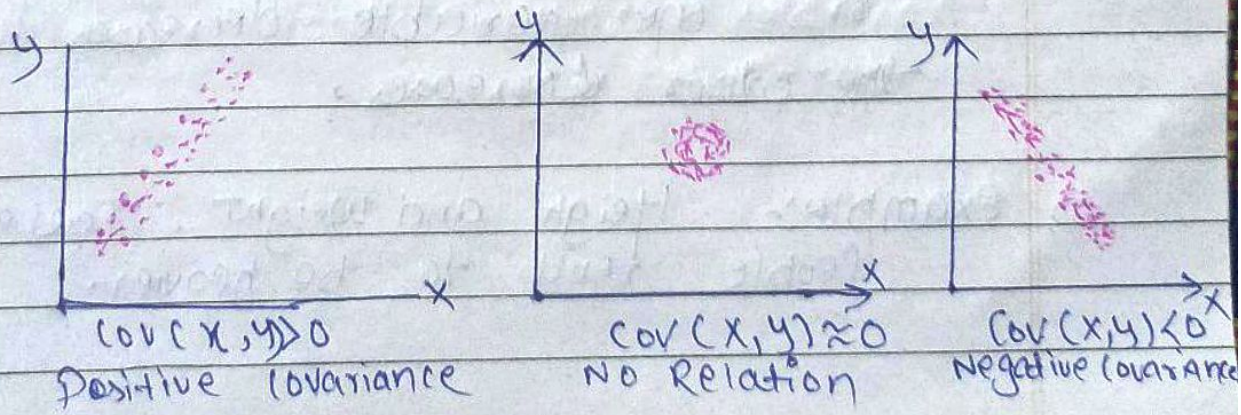
$$Cov(x,y) = \frac{\Sigma(x_i - \bar{x})(y_i - \bar{y})}{N}$$

Sample Covariance

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

Where

x_i = data value of x
y_i = data value of y
$\bar{x}$ = mean of x.
$\bar{y}$ = mean of y.
N = Number of data values.



→ Covariance values are not standardized. Therefore, the covariance can range from negative infinity to positive infinity.

Correlation

→ It refers to the statistical relationship between two entities.

→ In other words, it's how two variables move in relation to one another.

→ There are three possible results of a correlation study :-

Positive Correlation:

→ It is a relationship between two variables in which both variables move in the same direction.

→ Therefore, when one variable increases as the other variable increases, or one variable decreases while the other decreases.

→ Example:- Heigh and weight. Taller People tend to be heavier.

Negative Correlation:-

- It is a relationship between two variables in which an increase in one variable is associated with a decrease in the other.
- Example:- Height above sea level and Temperature.
  As you climb the mountain (increase in height) it gets colder (decrease in temperature).

Zero Correlation

- It exists when there is no relationship between two variables.
- For example there is no relationship between the amount of tea drunk and level of intelligence.

Types of Correlation Coefficients

While correlation studies how two entities relate to one another, a correlation coefficient measures the strength of the relationship between the two variables.

In Statistics, there are 3 types of correlation coefficients.
They are as follows:-

- **Pearson Correlation:-**

→ The Pearson correlation is the most commonly used measurement for a linear relationship between two variables.

→ The stronger the correlation between these two datasets, the closer it will be to +1 or -1.

- **Spearman Correlation:-**

→ This type of correlation is used to determine the monotonic relationship or association between two data sets.

→ Unlike the Pearson correlation coefficient, it's based on the ranked values for each data set and uses skewed or ordinal variables rather than normally distributed ones.

Kendall Correlation :-

This type of correlation measures the strength of dependence between two data sets.

Equation to Calculate Correlation :-

$$\frac{\Sigma (x(i) - \bar{x})(y(i) - \bar{y})}{\sqrt{(\Sigma (x(i) - \bar{x})^2 \; \Sigma (y(i) - \bar{y})^2}}$$

$x(i)$ = value of x
$y(i)$ = value of y
$\bar{x}$ = Mean of x value
$\bar{y}$ = Mean of y value