Unit-II: Statistical Inference

Syllabus

Need of statistics in Data Science and Big Data Analytics, Measures of Central Tendency: Mean, Median, Mode, Mid-range. Measures of Dispersion: Range, Variance, Mean Deviation, Standard Deviation. Bayes' theorem, Basics and need of hypothesis and hypothesis testing, Pearson Correlation, Sample Hypothesis testing, Chi-Square Tests, t-test.

Statistics - Definition

Statistics is the science of collecting, analyzing and understanding data, and accounting for the relevant uncertainties.

Need of statistics in Data Science and Big Data Analytics

1. **Prediction and Classification:** Statistics help in prediction and classification of data **whether it would be right for the clients viewing by their previous usage of data.**

2. **Helps to create Probability Distribution and Estimation:** Probability Distribution and Estimation are crucial in **understanding the basics of machine learning and algorithms like logistic regressions.**

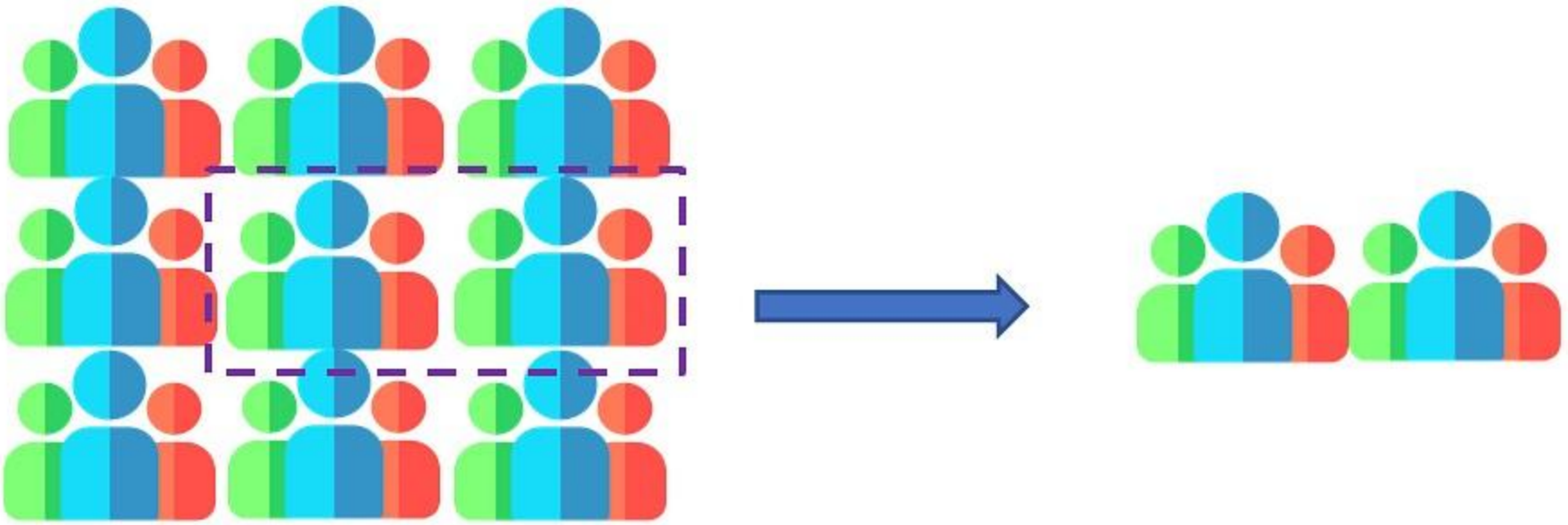
3. **Pattern Detection and Grouping:** Statistics help in picking out the optimal data and removing the irrelevant data for companies who like their work organized. It also helps spot out **anomalies** which further helps in processing the right data.

Population

In statistics, population is the entire set of items from which you draw data for a statistical study. It can be a group of individuals, a set of items, etc. It makes up the data pool for a study.

Sample

A sample represents the group of interest from the population, which you will use to represent the data.



Data Sampling

It is a statistical analysis technique used to select, manipulate and analyze a representative subset of data points to identify patterns and trends in the larger data set being examined.

Measures of Central Tendency: Mean

* **Mean**:- average

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$$

Measures of Central Tendency: Median

* Median :- the value seperating the higher half from the lower half of a data sample, a population or probability distribution.

Steps :-

1) Sort the dataset
2) Pick the middle element of the dataset
   a) pick middle element for odd number of elements.
   b) take average of two middle elements in case of even number of elements.

Measures of Central Tendency: Mode

* **Mode** :- The value that occurs most frequently in the data set. If each data occurs only once no mode.

Data sets with one mode - unimodal
two modes - bimodal
three modes - trimodal
multimodal.

Measures of Central Tendency: Mid-range

Mid-range:- average of the smallest and the largest values in the data set.

e.g. $X = [3, 5, 7, 3, 12, 5, 20]$

$$\text{Mid-range} = \frac{3+20}{2} = 11.5$$

Measures of Central Tendency: Range

Range:- The difference between the largest and the smallest values in the data set.

e.g. $X = [3, 5, 7, 3, 12, 5, 20]$

Range $= 20 - 3 = 17$

Measures of Central Tendency: Standard Deviation

Standard Deviation:-

① A measure of how spread out data is.

② The average distance from the mean of the data set to a point.

$$s = \sqrt{\frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n-1}}$$

Measures of Central Tendency: Standard Deviation

$D = [0 \quad 8 \quad 12 \quad 20]$

$\bar{x}$ = mean = 10

X	$x - \bar{x}$	$(x - \bar{x})^2$
0	-10	100
8	-2	4
12	2	4
20	10	100

Total = 208

Measures of Central Tendency: Standard Deviation

$D = [0 \quad 8 \quad 12 \quad 20]$

$\bar{x}$ = mean = 10

x	$x - \bar{x}$	$(x - \bar{x})^2$
0	-10	100
8	-2	4
12	2	4
20	10	100

Total = 208

$$s = \sqrt{\frac{\sum_{i=1}^{n} (x - \bar{x}_i)^2}{n-1}}$$

$$s = \sqrt{\frac{208}{3}}$$

$$s = 8.32$$

Measures of Central Tendency: Standard Deviation

$E = [8 \quad 9 \quad 11 \quad 12]$

$\bar{x}$ = mean = 10

x	$x - \bar{x}$	$(x - \bar{x})^2$
8	-2	4
9	-1	1
11	1	1
12	2	4

Total = 10

Measures of Central Tendency: Standard Deviation

$E = [8 \quad 9 \quad 11 \quad 12]$

$\bar{x}$ = mean = 10

x	$x - \bar{x}$	$(x - \bar{x})^2$
8	-2	4
9	-1	1
11	1	1
12	2	4

Total = 10

$$S = \sqrt{\frac{\sum_{i=1}^{n} (x - \bar{x}_i)^2}{n-1}}$$

$$S = \sqrt{\frac{10}{3}}$$

$$S = 1.82$$

Measures of Central Tendency: Standard Deviation

Conclusion :-

The first set has a much larger std deviation due to the fact that the data is much more spread out from the mean whereas the is closer to the mean in the second set.

If the set would have been [10 10 10 10] then the SD would have been 0 as none of the datapoints deviate from the mean.

Measures of Central Tendency: Variance

Variance:-

- Another measure of the spread of data in a data set.
- It is calculated as s^2 where s is the SD.

$$\text{Var}(x) = s^2 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n-1}$$

Measures of Central Tendency: Covariance

Covariance :-

- SD and variance work on 1-dimensional data.
- Covariance is measured betn 2 dimensions. If you calculate the covariance betn 1 dimension and itself, you get the variance for that dimension.

$$cov(x,y) = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{n-1}$$

$$\{cov(x,y) = cov(y,x)\}$$

Measures of Central Tendency: Covariance

e-g. = number of hours of study and marks obtained.

Hours (H)	Marks (M)
9	39
15	56
25	93
14	61
10	50
18	75
0	32
16	85
5	42
19	70
16	66
20	80

$\bar{H} = 13.92$

$\bar{M} = 62.42$

Measures of Central Tendency: Covariance

Hours (H)	Marks (M)	$H_i - \bar{H}$	$M_i - \bar{M}$	$(H_i - \bar{H})(M_i - \bar{M})$
9	39	-4.92	-23.42	115.23
15	56	1.08	-6.42	-6.93
25	93	11.08	30.58	338.83
14	61	0.08	-1.42	-0.11
10	50	-3.92	-12.42	48.69
18	75	4.08	12.58	51.33
0	32	-13.92	-30.42	423.45
16	85	2.08	22.58	46.97
5	42	-8.92	-20.42	182.15
19	70	5.08	7.58	38.51
16	66	2.08	3.58	7.45
20	80	6.08	17.58	106.89
				1352.42

$$\text{cov}(H, M) = 1352.42/11 = 122.94$$

Measures of Central Tendency: Covariance

(1) Its sign is important (+ve or -ve)

(2) If the value is positive - both the dimensions increase together

(3) If the sign is negative - as one of the dimensions increases, the other decreases.

(4) If covariance is zero - the two dimensions are independent of each other.

Measures of Central Tendency: Mean Absolute Deviation

Mean Absolute Deviation: -

- Average Absolute Deviation (AAD) of a dataset is the average of the absolute deviations from a central point.
- The central point can be a mean, median, mode or the result of any other measure of central tendancy or any reference value related to the given data set.
- AAD includes the mean absolute deviation and median absolute deviation (both abbreviated as MAD)

Measures of Central Tendency: Mean Absolute Deviation

Steps to calculate MAD of a dataset

1) Calculate the desired central tendency point (mean, median or mode) of the data set.

2) use -

$$\frac{1}{n} \sum_{i=1}^{n} |x_i - m(x)|$$

where $m(x)$ is the central point that is calculated in step 1.

Measures of Central Tendency: Mean Absolute Deviation

e.g. $X = [2, 2, 3, 4, 14]$

Measure of central Tendency $m(x)$

Mean Absolute Deviation.

Mean $= 5$

$$\frac{|2-5| + |2-5| + |3-5| + |4-5| + |14-5|}{5}$$

$$= \frac{18}{5} = 3.6$$

Measures of Central Tendency: Mean Absolute Deviation

Median = 3

$$\frac{|2-3| + |2-3| + |3-3| + |4-3| + |14-3|}{5}$$

$$= \frac{14}{5} = 2.8$$

mode = 2

$$\frac{|2-2| + |2-2| + |3-2| + |4-2| + |14-2|}{5}$$

$$= \frac{15}{5} = 3$$

Bayes' Theorem

Bayes' Theorem states that the conditional probability of an event, based on the occurrence of another event, is equal to the likelihood of the second event given the first event multiplied by the probability of the first event.

Bayes' Theorem

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

Where,

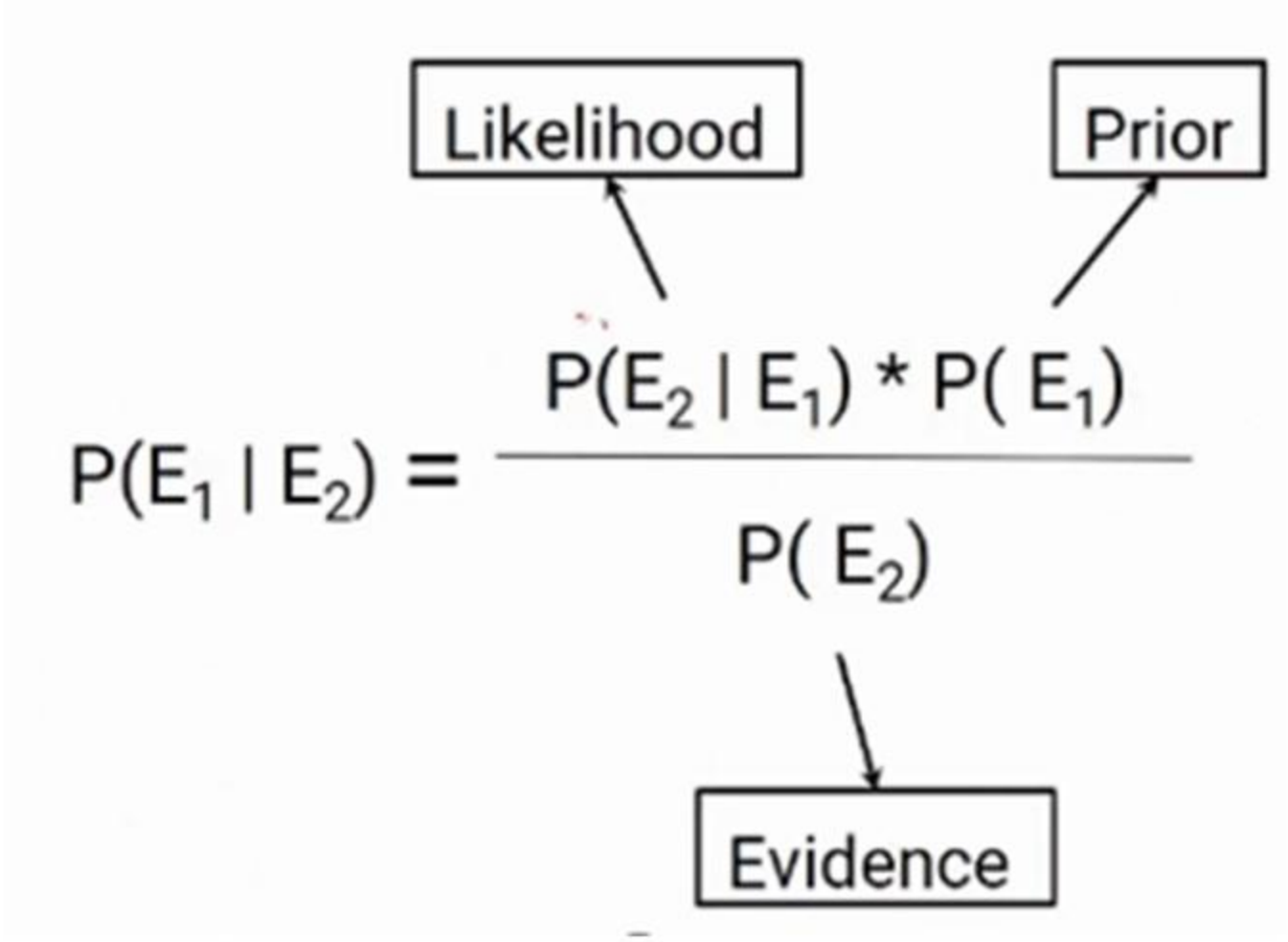
P(A|B) – the probability of event A occurring, given event B has occurred

P(B|A) – the probability of event B occurring, given event A has occurred

P(A) – the probability of event A

P(B) – the probability of event B

Bayes' Theorem



Hypothesis

- **Hypothesis testing** is a statistical technique that helps scientists and researchers test the validity of their claims about the real-world/real-life events.

- Hypothesis testing techniques are often used in statistics and data science to analyze whether the claims about the occurrence of the events is true.

Hypothesis

- **Null Hypothesis:** Null hypothesis is a statistical theory that suggests there is no statistical significance exists between the populations.
- It is denoted by $\mathbf{H_0}$.

- **Alternative Hypothesis:** An Alternative hypothesis suggests there is a significant difference between the population parameters. It could be greater or smaller. Basically, it is the contrast of the Null Hypothesis.
- It is denoted by $\mathbf{H_a}$ or $\mathbf{H_1}$.

Hypothesis - example

- **Running 5 miles a day will lead to a reduction of 10 kg of weight within a month.** Now, this is the claim which is required to be proved or otherwise.

- The **alternate hypothesis** will be formulated first as the statement that "running 5 miles a day will lead to a reduction of 10 kg of weight within a month".
- The **null hypothesis** will be the opposite of the alternate hypothesis and stated as the fact that "running 5 miles a day does not lead to a reduction of 10 kg of weight within a month".

Hypothesis - example

Null hypothesis	Running 5 miles a day **does not** result in the reduction of 10 kg of weight within a month.
Alternate hypothesis	Running 5 miles a day results in the reduction of 10 kg of weight within a month.

Hypothesis - example

- The housing price depends upon the average income of people staying in the locality.
- This is the claim which is required to be proved or otherwise.
- The **alternate hypothesis** will be formulated first as the statement that "**housing price depends upon the average income of people staying in the locality**".
- The **null hypothesis** will be formulated as the statement that **housing price does NOT depend upon the average income of people staying in the locality**.

Hypothesis - example

Null hypothesis	The housing price **does not** depend upon the average income of people staying in the locality.
Alternate hypothesis	The housing price depends upon the average income of people staying in the locality.

Correlation

Variables within a dataset can be related for lots of reasons.
For example:

- One variable could cause or depend on the values of another variable.
- One variable could be lightly associated with another variable.
- Two variables could depend on a third unknown variable.

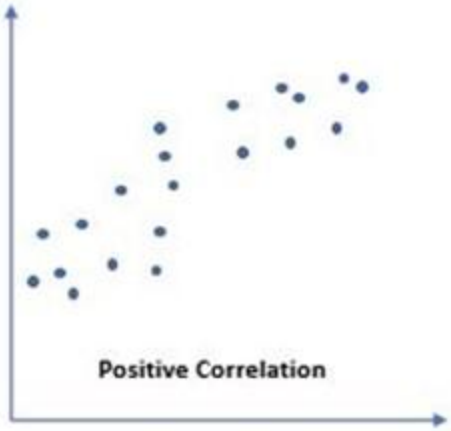
It can be useful in data analysis and modeling to better understand the relationships between variables. **The statistical relationship between two variables is referred to as their correlation.**

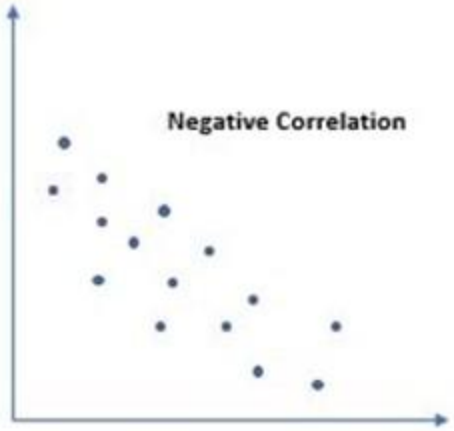
Correlation

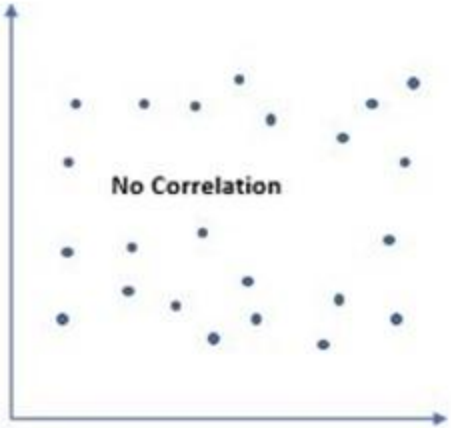
- A correlation could be positive, meaning both variables move in the same direction, or negative, meaning that when one variable's value increases, the other variables' values decrease.
- Correlation can also be neutral or zero, meaning that the variables are unrelated.

Correlation

- **Positive Correlation**: indicates an increase in one variable associated with an increase in the other
- **Negative Correlation**: indicates an increase in one variable would result in a decrease in the other
- **Neutral Correlation**: No relationship in the change of the variables.







Correlation

- Correlation measures the strength of association between two variables and the direction of the relationship.
- In terms of the strength of relationship, the value of the correlation coefficient varies between +1 and -1.
- A value of ± 1 indicates a perfect degree of association between the two variables.
- As the correlation coefficient value goes towards 0, the relationship between the two variables will be weaker.
- The direction of the relationship is indicated by the sign of the coefficient; a + sign indicates a positive relationship and a - sign indicates a negative relationship.

Correlation

Usually, in statistics, we measure four types of correlations:

- Pearson correlation
- Kendall rank correlation
- Spearman correlation
- Point-Biserial correlation

Pearson Correlation

- Pearson Correlation is one of the most used correlations during the data analysis process.
- Pearson correlation measures the linear relationship between variable X and variable Y and has a value between 1 and -1.
- In other words, the Pearson Correlation Coefficient measures the relationship between 2 variables via a line.

Pearson Correlation

SUBJECT	AGE X	GLUCOSE LEVEL Y	XY	X^2	Y^2
1	43	99	4257	1849	9801
2	21	65	1365	441	4225
3	25	79	1975	625	6241
4	42	75	3150	1764	5625
5	57	87	4959	3249	7569
6	59	81	4779	3481	6561
Σ	**247**	**486**	**20485**	**11409**	**40022**

Pearson Correlation

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

Pearson Correlation

SUBJECT	AGE X	GLUCOSE LEVEL Y	XY	X^2	Y^2
Σ	247	486	20485	11409	40022

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

Pearson Correlation

SUBJECT	AGE X	GLUCOSE LEVEL Y	XY	X^2	Y^2
Σ	247	486	20485	11409	40022

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

$$= \frac{(6*20485) - 247*486}{\sqrt{[6*11409 - (247)^2][6*40022 - (486)^2]}}$$

Pearson Correlation

SUBJECT	AGE X	GLUCOSE LEVEL Y	XY	X^2	Y^2
Σ	247	486	20485	11409	40022

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

$$= \frac{(6*20485) - 247*486}{\sqrt{[6*11409 - (247)^2][6*40022 - (486)^2]}}$$

$$= \frac{122910 - 120042}{\sqrt{[68454 - 61009][240132 - 236196]}} \qquad = \frac{2868}{\sqrt{7445 * 3936}}$$

Pearson Correlation

SUBJECT	AGE X	GLUCOSE LEVEL Y	XY	X^2	Y^2
Σ	247	486	20485	11409	40022

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

$$= \frac{(6*20485) - 247*486}{\sqrt{[6*11409 - (247)^2][6*40022 - (486)^2]}}$$

$$= \frac{122910 - 120042}{\sqrt{[68454 - 61009][240132 - 236196]}} \quad = \frac{2868}{\sqrt{7445 * 3936}}$$

$$= \frac{2868}{\sqrt{29303520}} \quad = \frac{2868}{5413.27} \quad = 0.5298$$

Pearson Correlation

- The Pearson correlation coefficient (named for Karl Pearson) can be used to summarize the strength of the linear relationship between two data samples.
- The Pearson's correlation coefficient is calculated as the covariance of the two variables divided by the product of the standard deviation of each data sample.

Pearson's correlation coefficient = covariance(X, Y) / (stdv(X) * stdv(Y))

Pearson Correlation

$$r = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 \sum (y_i - \bar{y})^2}}$$

Where,

r = Pearson Correlation Coefficient

x_i = x variable samples

y_i = y variable sample

$\bar{x}$ = mean of values in x variable

$\bar{y}$ =mean of values in y variable

Pearson Correlation

Subject	AGE x	Glucose Level y	$x_i - \overline{x}$	$y_i - \overline{y}$	$(x_i - \overline{x})(y_i - \overline{y})$	$(x_i - \overline{x})^2$	$(y_i - \overline{y})^2$
1	43	99	1.83	18	32.94	3.34	324
2	21	65	-20.17	-16	322.72	406.82	256
3	25	79	-16.17	-2	32.34	261.46	4
4	42	75	0.83	-6	-4.98	0.68	36
5	57	87	15.83	6	94.98	250.58	36
6	59	81	17.83	0	0	317.90	0
Σ	**247**	**486**			**478**	**1240.83**	**656**

$\overline{x} = 41.17$

$\overline{y} = 81$

Pearson Correlation

Subject	AGE x	Glucose Level y	$x_i - \bar{x}$	$y_i - \bar{y}$	$(x_i - \bar{x})(y_i - \bar{y})$	$(x_i - \bar{x})^2$	$(y_i - \bar{y})^2$
Σ	247	486			478	1240.83	656

$$r = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 \sum (y_i - \bar{y})^2}}$$

Pearson Correlation

Subject	AGE x	Glucose Level y	$x_i - \bar{x}$	$y_i - \bar{y}$	$(x_i - \bar{x})(y_i - \bar{y})$	$(x_i - \bar{x})^2$	$(y_i - \bar{y})^2$
Σ	247	486			478	1240.83	656

$$r = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 \sum (y_i - \bar{y})^2}}$$

$$= \frac{478}{\sqrt{1240.83 * 656}} \qquad = \frac{478}{\sqrt{813986.71}}$$

$$= \frac{478}{902.21} \qquad = 0.5298$$

Pearson Correlation

- The result of the calculation, the correlation coefficient can be interpreted to understand the relationship.
- The coefficient returns a value between -1 and 1 that represents the limits of correlation from a full negative correlation to a full positive correlation.
- A value of 0 means no correlation.
- The value must be interpreted, where often a value below -0.5 or above 0.5 indicates a notable correlation, and values below those values suggests a less notable correlation.

Pearson Correlation - Advantages

- It provides insight into how strongly the features are correlated.
- Highly effective in selecting features.
- Very easy to calculate.

Degrees of Freedom

- Degrees of freedom refers to the maximum number of logically independent values, which have the freedom to vary.
- It can be defined as the total number of observations minus the number of independent constraints imposed on the observations.

$$Df = N - 1$$

Df = degrees of freedom

N = sample size

Degrees of Freedom

- Consider a data sample consisting of five positive integers.
- The values could be any number with no known relationship between them. This data sample would, theoretically, have five degrees of freedom.
- Four of the numbers in the sample are {3, 8, 5, and 4} and the average of the entire data sample is revealed to be 6.
- This must mean that the fifth number has to be 10. It can be nothing else. It does not have the freedom to vary.
- So, the degrees of freedom for this data sample is 4.

Contingency Table

A table showing the distribution of one variable in rows and another in columns. It is used to study the relation between two variables.

Exited\ Gender	Yes	No	Total
Male	38	178	216
Female	44	140	184
Total	82	318	400

Degrees of freedom for contingency table is given as (r-1) * (c-1) where r, c are rows and columns. Here df = (2–1) * (2–1) = 1.

Categorical Variables

- Categorical variables fall into a particular category of those variables that can be divided into finite categories.
- These categories are generally **names or labels**.
- These variables are also called **qualitative variables** as they depict the quality or characteristics of that particular variable.

For example, If people responded to a survey about which what brand of car they owned, the responses would fall into **categories** like "Honda", "Toyota", and "Ford". In this case, the data is categorical.

Categorical Variables

There are broadly two types of categorical variables:

1.Nominal Variable: A nominal variable has no natural ordering to its categories.

- They have two or more categories.
- For example, Marital Status (Single, Married, Divorcee); Gender (Male, Female, Transgender), etc.

2.Ordinal Variable: A variable for which the categories can be placed in an order.

- For example, Customer Satisfaction (Excellent, Very Good, Good, Average, Bad), and so on

When the data we want to analyze contains this type of variable, we turn to the chi-square test, denoted by χ^2, to test our hypothesis.

Chi-Square Tests

- The Chi-Square test is a statistical procedure for determining the difference between observed and expected data.
- This test can also be used to determine whether it correlates to the categorical variables in our data.
- The test calculates a statistic that has a chi-squared distribution, named for the Greek capital letter **Chi (X) pronounced "ki".**

Chi-Square Tests – Example

- A research scholar is interested in the relationship between the placement of students in the department of a reputed University and their C.G.P.A.
- He obtains the placement records of the past five years from the placement cell database (at random).
- He records how many students who got placed fell into each of the following C.G.P.A. categories – 9-10, 8-9, 7-8, 6-7, and below 6.

Chi-Square Tests – Example

- If there is no relationship between the placement rate and the C.G.P.A., then the placed students should be equally spread across the different C.G.P.A. categories (i.e. there should be similar numbers of placed students in each category).
- However, if students having C.G.P.A more than 8 are more likely to get placed, then there would be a large number of placed students in the higher C.G.P.A. categories as compared to the lower C.G.P.A. categories.
- In this case, the data collected would make up the observed frequencies.

Chi-Square Tests – Example

- So the question is, **are these frequencies being observed by chance or do they follow some pattern?**
- The chi-square test helps us answer the above question by comparing the **observed frequencies to the frequencies that we might expect** to obtain purely by chance.
- Chi-square test in hypothesis testing is used to test the hypothesis about the distribution of observations/frequencies in different categories.

Types of Chi-Square Tests

- Chi-Square Goodness of Fit Test
- Chi-Square Test for Association/Independence

Chi-Square Goodness of Fit Test

- We use it to find how the **observed value** of a given event is significantly **different** from the **expected value**.
- In this case, we have categorical data for one independent variable, and we want to check whether the distribution of the data is similar or different from that of the expected distribution.
- Let's consider the above example where the research scholar was interested in the relationship between the placement of students in the statistics department of a reputed University and their C.G.P.A.

Chi-Square Goodness of Fit Test

- In this case, the independent variable is C.G.P.A with the categories 9-10, 8-9, 7-8, 6-7, and below 6.
- The statistical question here is: whether or not the observed frequencies of placed students are equally distributed for different C.G.P.A categories (so that our theoretical frequency distribution contains the same number of students in each of the C.G.P.A categories).

Chi-Square Goodness of Fit Test

We will arrange this data by using the contingency table which will consist of both the observed and expected values as below:

C.G.P.A						
	10-9	**9-8**	**8-7**	**7-6**	**Below 6**	**Total**
Observed Frequency of Placed students	30	35	20	10	5	100
Expected Frequency of Placed students	20	20	20	20	20	100

Chi-Square Goodness of Fit Test

- After constructing the contingency table, the next task is to compute the value of the chi-square statistic. The formula for chi-square is given as:

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

where,

$\chi 2$ = Chi-Square value

O_i = Observed frequency

E_i = Expected frequency

Chi-Square Goodness of Fit Test

- **Step 1:** Subtract each expected frequency from the related observed frequency. For example, for the C.G.P.A category 10-9, it will be "30-20 = 10". Apply similar operation for all the categories
- **Step 2:** Square each value obtained in step 1, i.e. (O-E)2. For example: for the C.G.P.A category 10-9, the value obtained in step 1 is 10. It becomes 100 on squaring. Apply similar operation for all the categories

Chi-Square Goodness of Fit Test

- **Step 3:** Divide all the values obtained in step 2 by the related expected frequencies i.e. $(O-E)^2/E$. For example: for the C.G.P.A category 10-9, the value obtained in step 2 is 100. On dividing it with the related expected frequency which is 20, it becomes 5. Apply similar operation for all the categories

Chi-Square Goodness of Fit Test

- **Step 4:** Add all the values obtained in step 3 to get the chi-square value. In this case, the chi-square value comes out to be 32.5
- **Step 5:** Once we have calculated the chi-square value, the next task is to compare it with the critical chi-square value. We can find this in the below chi-square table against the degrees of freedom (number of categories – 1) and the level of significance

Chi-Square Goodness of Fit Test

C.G.P.A						
	10-9	**9-8**	**8-7**	**7-6**	**Below 6**	**Total**
Observed Frequency of Placed students (O)	30	35	20	10	5	100
Expected Frequency of Placed students (E)	20	20	20	20	20	100
O- E	10	15	0	-10	-15	
$(O- E)^2$	100	225	0	100	225	
$(O- E)^2/E$	5	11.25	0	5	11.25	
$\Sigma[(O- E)^2/E]$				32.5		

Chi-Square Goodness of Fit Test

- In this case, the **degrees of freedom** are **5-1 = 4**. So, the critical value at 5% level of significance is **9.49.**
- Our obtained value of **32.5** is much larger than the critical value of 9.49.
- Therefore, we can say that the observed frequencies are significantly different from the expected frequencies. In other words, C.G.P.A is related to the number of placements that occur in the department.

Chi-Square Goodness of Fit Test

Chi-Square (χ^2) Distribution

Degrees of Freedom	Area to the Right of Critical Value							
	0.99	0.975	0.95	0.90	0.10	0.05	0.025	0.01
1	—	0.001	0.004	0.016	2.706	3.841	5.024	6.635
2	0.020	0.051	0.103	0.211	4.605	5.991	7.378	9.210
3	0.115	0.216	0.352	0.584	6.251	7.815	9.348	11.345
4	0.297	0.484	0.711	1.064	7.779	9.488	11.143	13.277
5	0.554	0.831	1.145	1.610	9.236	11.071	12.833	15.086
6	0.872	1.237	1.635	2.204	10.645	12.592	14.449	16.812
7	1.239	1.690	2.167	2.833	12.017	14.067	16.013	18.475
8	1.646	2.180	2.733	3.490	13.362	15.507	17.535	20.090
9	2.088	2.700	3.325	4.168	14.684	16.919	19.023	21.666
10	2.558	3.247	3.940	4.865	15.987	18.307	20.483	23.209
11	3.053	3.816	4.575	5.578	17.275	19.675	21.920	24.725
12	3.571	4.404	5.226	6.304	18.549	21.026	23.337	26.217
13	4.107	5.009	5.892	7.042	19.812	22.362	24.736	27.688
14	4.660	5.629	6.571	7.790	21.064	23.685	26.119	29.141
15	5.229	6.262	7.261	8.547	22.307	24.996	27.488	30.578
16	5.812	6.908	7.962	9.312	23.542	26.296	28.845	32.000
17	6.408	7.564	8.672	10.085	24.769	27.587	30.191	33.409
18	7.015	8.231	9.390	10.865	25.989	28.869	31.526	34.805
19	7.633	8.907	10.117	11.651	27.204	30.144	32.852	36.191
20	8.260	9.591	10.851	12.443	28.412	31.410	34.170	37.566
21	8.897	10.283	11.591	13.240	29.615	32.671	35.479	38.932
22	9.542	10.982	12.338	14.042	30.813	33.924	36.781	40.289
23	10.196	11.689	13.091	14.848	32.007	35.172	38.076	41.638
24	10.856	12.401	13.848	15.659	33.196	36.415	39.364	42.980
25	11.524	13.120	14.611	16.473	34.382	37.652	40.646	44.314
26	12.198	13.844	15.379	17.292	35.563	38.885	41.923	45.642
27	12.879	14.573	16.151	18.114	36.741	40.113	43.194	46.963
28	13.565	15.308	16.928	18.939	37.916	41.337	44.461	48.278
29	14.257	16.047	17.708	19.768	39.087	42.557	45.722	49.588
30	14.954	16.791	18.493	20.599	40.256	43.773	46.979	50.892

Chi-Square Goodness of Fit Test

Chi-Square (χ^2) Distribution

Degrees of Freedom	Area to the Right of Critical Value							
	0.99	0.975	0.95	0.90	0.10	0.05	0.025	0.01
1	—	0.001	0.004	0.016	2.706	3.841	5.024	6.635
2	0.020	0.051	0.103	0.211	4.605	5.991	7.378	9.210
3	0.115	0.216	0.352	0.584	6.251	7.815	9.348	11.345
4	0.297	0.484	0.711	1.064	7.779	9.488	11.143	13.277
5	0.554	0.831	1.145	1.610	9.236	11.071	12.833	15.086
6	0.872	1.237	1.635	2.204	10.645	12.592	14.449	16.812
7	1.239	1.690	2.167	2.833	12.017	14.067	16.013	18.475
8	1.646	2.180	2.733	3.490	13.362	15.507	17.535	20.090
9	2.088	2.700	3.325	4.168	14.684	16.919	19.023	21.666
10	2.558	3.247	3.940	4.865	15.987	18.307	20.483	23.209
11	3.053	3.816	4.575	5.578	17.275	19.675	21.920	24.725
12	3.571	4.404	5.226	6.304	18.549	21.026	23.337	26.217
13	4.107	5.009	5.892	7.042	19.812	22.362	24.736	27.688
14	4.660	5.629	6.571	7.790	21.064	23.685	26.119	29.141
15	5.229	6.262	7.261	8.547	22.307	24.996	27.488	30.578
16	5.812	6.908	7.962	9.312	23.542	26.296	28.845	32.000
17	6.408	7.564	8.672	10.085	24.769	27.587	30.191	33.409
18	7.015	8.231	9.390	10.865	25.989	28.869	31.526	34.805
19	7.633	8.907	10.117	11.651	27.204	30.144	32.852	36.191
20	8.260	9.591	10.851	12.443	28.412	31.410	34.170	37.566
21	8.897	10.283	11.591	13.240	29.615	32.671	35.479	38.932
22	9.542	10.982	12.338	14.042	30.813	33.924	36.781	40.289
23	10.196	11.689	13.091	14.848	32.007	35.172	38.076	41.638
24	10.856	12.401	13.848	15.659	33.196	36.415	39.364	42.980
25	11.524	13.120	14.611	16.473	34.382	37.652	40.646	44.314
26	12.198	13.844	15.379	17.292	35.563	38.885	41.923	45.642
27	12.879	14.573	16.151	18.114	36.741	40.113	43.194	46.963
28	13.565	15.308	16.928	18.939	37.916	41.337	44.461	48.278
29	14.257	16.047	17.708	19.768	39.087	42.557	45.722	49.588
30	14.954	16.791	18.493	20.599	40.256	43.773	46.979	50.892

Chi-Square Test for Association/Independence

- This test is used when we have categorical data for **two independent variables**, and we want to see if there is any relationship between the variables.
- Example, a teacher wants to know the answer to whether the outcome of a mathematics test is related to the gender of the person taking the test. Or in other words, she wants to know if males show a different pattern of pass/fail rates than females.

Chi-Square Test for Association/Independence

- So, here are two categorical variables: Gender (Male and Female) and mathematics test outcome (Pass or Fail).
- Let us now look at the contingency table:

Chi-Square Test for Association/Independence

	Boys	Girls
Pass	17	20
Fail	8	5

- By looking at the above contingency table, we can see that the girls have a comparatively higher pass rate than boys.
- However, to test whether this observed difference is significant or not, we will carry out the chi-square test.

Chi-Square Test for Association/Independence

The steps to calculate the chi-square value are as follows:

Step 1: Calculate the row and column total of the above contingency table:

	Boys	Girls	Total
Pass	17	20	37
Fail	8	5	13
Total	25	25	50

Chi-Square Test for Association/Independence

Step 2: Calculate the expected frequency for each individual cell by multiplying row sum by column sum and dividing by total number:

Expected Frequency = (Row Total x Column Total)/Grand Total

For the first cell, the expected frequency would be (37*25)/50 = 18.5.
Now, write them below the observed frequencies in brackets:

	Boys	Girls	Total
Pass(O) (E)	17 (18.5)	20 (18.5)	37
Fail(O) (E)	8 (6.5)	5 (6.5)	13
Total	25	25	50

Chi-Square Test for Association/Independence

Step 3: Calculate the value of chi-square using the formula:

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

Calculate the right-hand side part of each cell.
For example, for the first cell, $((17\text{-}18.5)^2)/18.5 = 0.1216$.

	Boys	Girls	Total
Pass(O) (E)	17 (18.5)	20 (18.5)	37
	$(17\text{-}18.5)^2/18.5 =$ **0.1216**	$(20\text{-}18.5)^2/18.5$ **= 0.1216**	
Fail(O) (E)	8 (6.5)	5 (6.5)	13
	$(8\text{-}6.5)^2/6.5 =$ **0.3461**	$(5\text{-}6.5)^2/6.5 =$ **0.3461**	
0.1216+0.1216+0.3461+0.3461 = **0.9354**			

Chi-Square Test for Association/Independence

Step 4: Then, add all the values obtained for each cell. In this case, the values are:

0.1216+0.1216+0.3461+0.3461 = 0.9354

Step 5: Calculate the **degrees of freedom**, i.e. (Number of rows-1)*(Number of columns-1) = **1*1 = 1.** So, the critical value at 5% level of significance is **3.84.**

Chi-Square Goodness of Fit Test

Chi-Square (χ^2) Distribution

Degrees of Freedom	Area to the Right of Critical Value							
	0.99	0.975	0.95	0.90	0.10	0.05	0.025	0.01
1	—	0.001	0.004	0.016	2.706	3.841	5.024	6.635
2	0.020	0.051	0.103	0.211	4.605	5.991	7.378	9.210
3	0.115	0.216	0.352	0.584	6.251	7.815	9.348	11.345
4	0.297	0.484	0.711	1.064	7.779	9.488	11.143	13.277
5	0.554	0.831	1.145	1.610	9.236	11.071	12.833	15.086
6	0.872	1.237	1.635	2.204	10.645	12.592	14.449	16.812
7	1.239	1.690	2.167	2.833	12.017	14.067	16.013	18.475
8	1.646	2.180	2.733	3.490	13.362	15.507	17.535	20.090
9	2.088	2.700	3.325	4.168	14.684	16.919	19.023	21.666
10	2.558	3.247	3.940	4.865	15.987	18.307	20.483	23.209
11	3.053	3.816	4.575	5.578	17.275	19.675	21.920	24.725
12	3.571	4.404	5.226	6.304	18.549	21.026	23.337	26.217
13	4.107	5.009	5.892	7.042	19.812	22.362	24.736	27.688
14	4.660	5.629	6.571	7.790	21.064	23.685	26.119	29.141
15	5.229	6.262	7.261	8.547	22.307	24.996	27.488	30.578
16	5.812	6.908	7.962	9.312	23.542	26.296	28.845	32.000
17	6.408	7.564	8.672	10.085	24.769	27.587	30.191	33.409
18	7.015	8.231	9.390	10.865	25.989	28.869	31.526	34.805
19	7.633	8.907	10.117	11.651	27.204	30.144	32.852	36.191
20	8.260	9.591	10.851	12.443	28.412	31.410	34.170	37.566
21	8.897	10.283	11.591	13.240	29.615	32.671	35.479	38.932
22	9.542	10.982	12.338	14.042	30.813	33.924	36.781	40.289
23	10.196	11.689	13.091	14.848	32.007	35.172	38.076	41.638
24	10.856	12.401	13.848	15.659	33.196	36.415	39.364	42.980
25	11.524	13.120	14.611	16.473	34.382	37.652	40.646	44.314
26	12.198	13.844	15.379	17.292	35.563	38.885	41.923	45.642
27	12.879	14.573	16.151	18.114	36.741	40.113	43.194	46.963
28	13.565	15.308	16.928	18.939	37.916	41.337	44.461	48.278
29	14.257	16.047	17.708	19.768	39.087	42.557	45.722	49.588
30	14.954	16.791	18.493	20.599	40.256	43.773	46.979	50.892

Chi-Square Goodness of Fit Test

Chi-Square (χ^2) Distribution

Degrees of Freedom	Area to the Right of Critical Value							
	0.99	0.975	0.95	0.90	0.10	0.05	0.025	0.01
1	—	0.001	0.004	0.016	2.706	3.841	5.024	6.635
2	0.020	0.051	0.103	0.211	4.605	5.991	7.378	9.210
3	0.115	0.216	0.352	0.584	6.251	7.815	9.348	11.345
4	0.297	0.484	0.711	1.064	7.779	9.488	11.143	13.277
5	0.554	0.831	1.145	1.610	9.236	11.071	12.833	15.086
6	0.872	1.237	1.635	2.204	10.645	12.592	14.449	16.812
7	1.239	1.690	2.167	2.833	12.017	14.067	16.013	18.475
8	1.646	2.180	2.733	3.490	13.362	15.507	17.535	20.090
9	2.088	2.700	3.325	4.168	14.684	16.919	19.023	21.666
10	2.558	3.247	3.940	4.865	15.987	18.307	20.483	23.209
11	3.053	3.816	4.575	5.578	17.275	19.675	21.920	24.725
12	3.571	4.404	5.226	6.304	18.549	21.026	23.337	26.217
13	4.107	5.009	5.892	7.042	19.812	22.362	24.736	27.688
14	4.660	5.629	6.571	7.790	21.064	23.685	26.119	29.141
15	5.229	6.262	7.261	8.547	22.307	24.996	27.488	30.578
16	5.812	6.908	7.962	9.312	23.542	26.296	28.845	32.000
17	6.408	7.564	8.672	10.085	24.769	27.587	30.191	33.409
18	7.015	8.231	9.390	10.865	25.989	28.869	31.526	34.805
19	7.633	8.907	10.117	11.651	27.204	30.144	32.852	36.191
20	8.260	9.591	10.851	12.443	28.412	31.410	34.170	37.566
21	8.897	10.283	11.591	13.240	29.615	32.671	35.479	38.932
22	9.542	10.982	12.338	14.042	30.813	33.924	36.781	40.289
23	10.196	11.689	13.091	14.848	32.007	35.172	38.076	41.638
24	10.856	12.401	13.848	15.659	33.196	36.415	39.364	42.980
25	11.524	13.120	14.611	16.473	34.382	37.652	40.646	44.314
26	12.198	13.844	15.379	17.292	35.563	38.885	41.923	45.642
27	12.879	14.573	16.151	18.114	36.741	40.113	43.194	46.963
28	13.565	15.308	16.928	18.939	37.916	41.337	44.461	48.278
29	14.257	16.047	17.708	19.768	39.087	42.557	45.722	49.588
30	14.954	16.791	18.493	20.599	40.256	43.773	46.979	50.892

Chi-Square Test for Association/Independence

- The next task is to compare it with the critical chi-square value from the table we saw above.
- The Chi-Square calculated value is **0.9354** which is less than the critical value of **3.84**.
- So in this case, we fail to reject the null hypothesis.
- This means there is **no significant association** between the two variables, i.e, boys and girls have a statistically similar pattern of pass/fail rates on their mathematics tests.