Descriptive Stats [Refers to the measure used to determine the centre of the distribution of data].

1. Measure of Central Tendency – {Mean, Median, Mode}
2. Measure of Dispersions – {Variance, S.D}
   ↳ How they spread from the mean.

Statistics:- It is the science of collecting, Organising & Analysing data.

↓

Descriptive Stats – It consists of Organising & Summarising data

Inferential Stats – Technique where in we used the data that we have measured to form conclusions.

Sampling Techniques

1. **Simple Random Sampling:-**

Variance

Population	Sample
$\sigma^2 = \sum_{i=1}^{N} \frac{(x_i - \mu)^2}{N}$	$s^2 = \sum_{i=1}^{n} \frac{(x_i - \bar{x})^2}{n-1}$

1. Simple Random Sampling :- Randomly they selecting of data points from sample.

2. Stratified Sampling - where the population (N) is splited in non-overlapping groups (Strata). [Individual data].

3. Systematic Sampling - Population (N) - n^{th} individual

4. Convenience Sampling - Availability, willingness or ease of access. Based then they will select samples.

Variable :-

→ Quantative — Measure by numerical → Discrete → phone
→ Continuous
→ Intervals or Decimals.

→ Qualitative — Catorical Variable

Only Discrete (Discrete Continuous)

Measure of Variable Types :-

1. Nominal - Catorical data
2. Ordinal - Catorical data (Nomial) data but it follows Order
3. Interval - 1-5, 5-15, 16-30
4. Ratio - gap between Interval Equal
   1-5, 6-10, 11-15

→ Percentiles & Quartiles

Percentage:- 1, 2, 3, 4, 5

% of the numbers that are odd

$$\% = \frac{\text{\# of numbers that are odd}}{\text{Total Numbers}}$$

$$= \frac{3}{5} = 0.6 = 60\%$$

Percentiles:- (Data, CAT, GMAT, SAT)

↳ A percentile is a value below which a certain percentage of observation lie.

Data set:- 2, 2, 3, 4, 5, 5, 5, 6, 7, 8, 8, 8, 8, 8, 9, 9, 10, 11, 11, 12

(5)

What is the percentile ranking of 10?

$$\text{percentile Rank of 10} = \frac{\text{No of Value below 10}}{20} \times 100$$

$$= \frac{16}{20} \times 100$$

$$= 80\%$$

What value exists at percentile ranking of 25%?

$$\text{Value} = \frac{\text{Percentile}}{100} \times (n+1)$$

$$= \frac{25}{100} \times (20+1) = \frac{25}{100} \times 21 = 5.25$$

↓

(5) — 25%

Five Number Summary

1. Minimum
2. First Quartile (Q_1) (25%)
3. Median (50%)
4. Third Quartile (Q_3) (75%)
5. Maximum

IQR

* Spread of the middle 50% of a dataset.
* Is a key statistic in detecting variability and outliers.

Removing the Outliers:-

{1, 2, 2, 2, 3, 3, 4, 5, 5, 5, 6, 6, 6, 6, 7, 8, 8, 9, 27}

$Q_1 = 3$ $\quad Q_3 = 7$

$IQR = Q_3 - Q_1 = 7 - 3 = 4$

[Lower fence → Higher fence]

↓ ↓

$[Q_1 - 1.5(IQR)]$ $[Q_3 + 1.5(IQR)]$

Lower Fence $= Q_1 - 1.5(IQR)$

$= 3 - 1.5(4) = 3 - 6 = -3$

Higher Fence $= Q_3 + 1.5(IQR)$

$= 7 + 1.5(4) = 7 + 6 = 13$

$[-3, +13]$

Remaining data

1, 2, 2, 2, 3, 3, 4, 5, 5, 5, 6, 6, 6, 6, 7, 8, 8, 9, ~~27~~

↓ Removed

Minimum = 1

$Q_1 = 3$

Median = 5

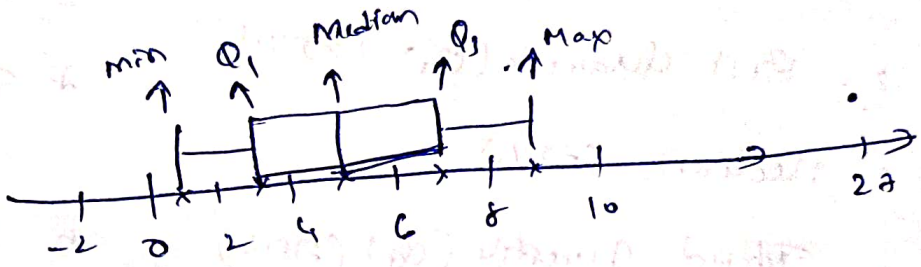
$Q_3 = 7$

Max = 9

→ 5 Number Summary

↓↓

Boxplot.



$$\sum_{i=1}^{n} \frac{(x_i - \bar{x})^2}{n-1}$$ → Bessel's correction

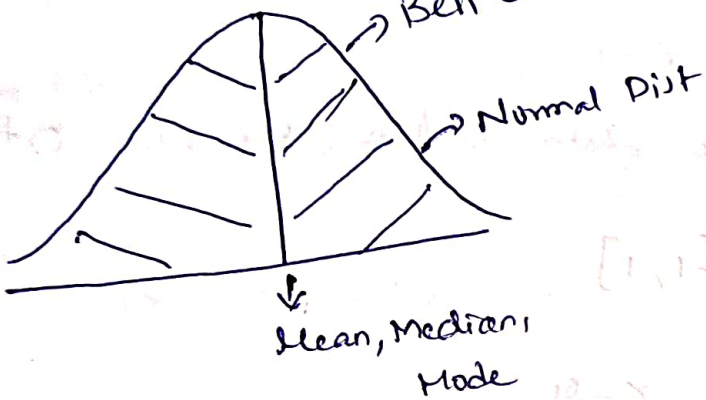
↓ → Degrees of freedom

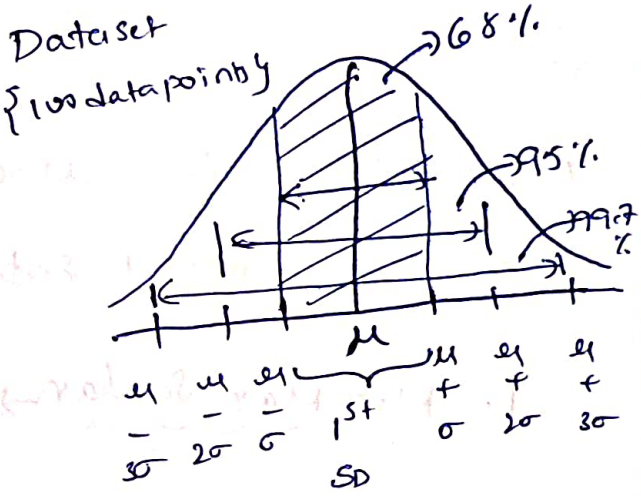
[To correct bias and get an unbiased Estimate of the population Variance]

Distributions:-

- ↳ Normal Distribution
- ↳ Standard Normal Distribution
- ↳ Z Score
- ↳ Log Normal Distributions
- ↳ Bernoulli's Distribution
- ↳ Binomial Distribution

① · Gaussian / Normal Dist:-





Empirical formula

68% – 95% – 99.7% Rule

↓

This is present in 1st S.D i.e., Bell Shape

4.5 → Clear to know

4.75 → Not clear

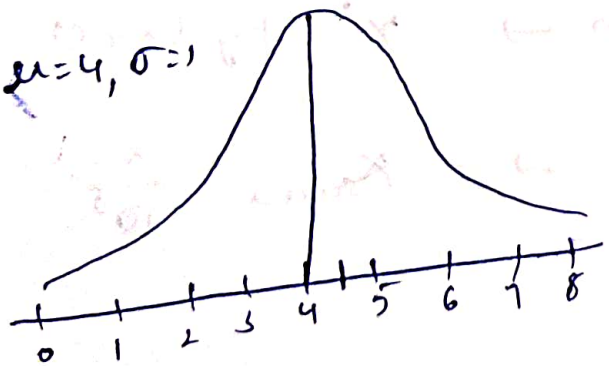
↓

$$Z\ score = \frac{x - \mu}{\sigma}$$

$$= \frac{4.75 - 4}{1}$$

$$= 0.75\ SD$$

Ex:- $\mu = 4, \sigma = 1$



$\mu = 4, \sigma = 1$

$\{1, 2, 3, 4, 5, 6, 7\}$ $\left[\text{Z score} = \frac{x_i - \mu}{\sigma} \right]$

↳ Normal Distribution / Gaussian Distri ↓

$\{-3, -2, -1, 0, 1, 2, 3\}$ $z(1) = \frac{1-4}{1} = -3$

↳ Standard Normal Distribution

$\mu = 0, \sigma = 1$ $z(2) = \frac{2-4}{1} = -2$

⋮

Standardization:-

The process of making data comparable by putting it on a common scale. $z = \frac{x_i - \mu}{\sigma}$ $z(7) = \frac{7-4}{1} = 3$

Normalization:- $\mu = 0, \sigma = 1$

↳ Convert entire data in ~~[illegible]~~ b/w 0 to 1

1. Min Max Scalar → $[-1, 1]$
2. Z Score → $\frac{x - \mu}{\sigma}$
3. Mean Normalization → $\frac{x - \mu}{x_{max} - x_{min}}$
4. Robust Scaling → $\frac{X - \text{Median}(x)}{\text{IQR}(x)}$
5. Log Transformation → $x' = \log(x + c)$
6. Decimal Scaling → $x_{scaled} = \frac{x}{10^j}$

Probability :-

* Is a measure of the likelihood of an Events.

→ Additional Rule (probability, "or")

* Mutual Exclusive Event

→ Two Event are mutual Exclusive if they cannot occur at the same time.

Ex:- Rolling a dice $\{1,2,3,4,5,6\}$ → $P(3 \text{ or } 6) = P(3) + P(6)$

$\{1,2\}$ $= (\frac{1}{6}) + (\frac{1}{6})$

$= 2/6 = 1/3$

Tossing a coin $\{H \text{ "or" } T\}$ → $P(H \text{ or } T) = P(H) + P(T)$

$= 1/2 + 1/2$

$= 1$

* Non mutual Exclusive Event (probability, "AND")

→ Multiple Events can occur at the same time

Ex:- Deccan of cards $\{\clubsuit, \heartsuit\}$

$\therefore P(A \text{ or } B) = P(A) + P(B) - P(A \text{ and } B)$

$P(\heartsuit) = \frac{4}{52}$; $P(\heartsuit) = \frac{13}{52}$

But $P(\heartsuit \text{ and } \heartsuit) = \frac{1}{52}$

$\Rightarrow P(\heartsuit \text{ or } \heartsuit) = P(\heartsuit) + P(\heartsuit) - P(\heartsuit \text{ and } \heartsuit)$

$= \frac{4}{52} + \frac{13}{52} - \frac{1}{52} = \frac{16}{52}$.

* Naive Bayes (Conditional probability):

→ It is a probabilistic classification algorithm based on Bayes Theorem.

→ It is widely used in Machine learning for tasks like Spam detection, Sentiment Analysis, Text Classification.

Bayes Theorem:

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

Where

$P(A|B)$ = Probability of Event A happening given B has occurred (posterior probability)

$P(B|A)$ = probability of B happening given A has occurred (likelihood)

$P(A)$ = probability of Event A occuring (prior probability)

$P(B)$ = probability of Event B occuring (Evidence)

Permutations:-

A permutation is an arrangement of objects in a specific orde

$$P(n,r) = {}^{n}P_{r} = \frac{n!}{(n-r)!}$$

Ex:- Passwords, seating arrangements, race rankings.

Combinations:-

* A combination is an selection of objects where the order does not matter

$$C(n,r) = {}^{n}C_{r} = \frac{n!}{r!(n-r)!}$$

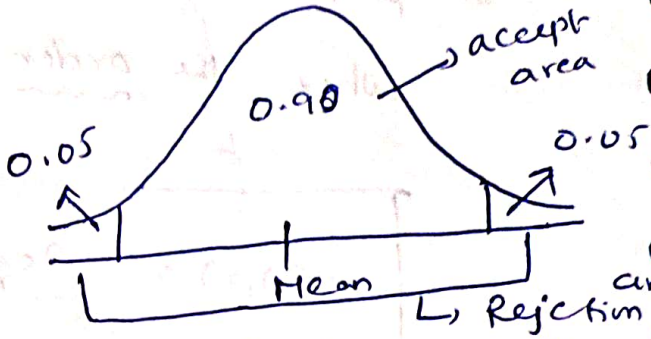
Ex:- Lottery, Selecting committees, teams.

* Both are used for selecting of objects.
  * If order matters - Use permutations
  * If order doesn't matter - Use combinations.

Hypothesis Testing:

* It is a statistical method used to make decisions about a population based on sample data.

* It helps us determine wheather an assumption (hypothesis) about a dataset is true or not.



Steps:-

Step-1 :- Define the Hypothesis.

1. Null hypothesis (H_0) :-

The assumption that there is no significant effect or no difference

2. Alternative Hypothesis (H_1):-

The assumption that there is a significant effect or difference

Step-2 :- Select the Significance level (α)

The Significance level (α) is the probability of rejecting (H_0) when it's actually true

* Common choices: 0.05(5%) or 0.01(1%)

* If $\alpha = 0.05$, we are 95% confident in our results

Step-3: Choose Statistical Test

Type of data	Test to Use
1. Comparing means	t-test (Small dataset (Samples) z-test (Large Samples)
2. Comparing two Groups	Chi-Square test (χ^2)
3. Testing Relationships	Correlation (or) regressions.

Step-4: Compute the Test Statistic & P-Value

Test Statistics :- A value that helps determine if the Sample data is different from the assumption in H_0.

P-Value :- The probability of observing the data if H_0 were true.

Step-5:- Make a Decision [Reject or Accept (fail to Reject) H_0]

* If P-Value $\leq \alpha$ → Reject H_0 [There is Enough Evidance to support H_1]
* If P-Value $> \alpha$ → fail to reject H_0 [there is not Enough Evidance to support H_1].

Type I & II Error :-

Rejecting a true Null Hypothesis

Decision	H_0 is true	H_0 is false
Reject H_0	Type I Error (False positive) X	✓ Correct Decision (True positive)
Fail to Reject H_0	✓ Correct Decision (True Negative)	Type II Error (False Negative) X

Accepting a false Null Hypothesis.

	Actually positive	Actually Negative
predicted positive	True positive (TP)	False positive (FP) Type-I Error
predicted Negative	false Negative (FN) Type-II Error	True Negative (TN)

One-Tailed Test (Directional Test)

* One tailed test is used when we are testing for a specific direction (Either an increase (or) Decrease)
* We check if the sample mean is significantly greater than (or) less than the hypothesized mean, but not Both.

Ex:-

* A railway company claims the average delay of trains is less than 10 min after introducing a new system.

Hypothesis:-

Null Hypothesis (H_0):- The average delay is 10 min or more

Alternative Hypothesis (H_1):- The average delay is less than 10 min.

When to use:-

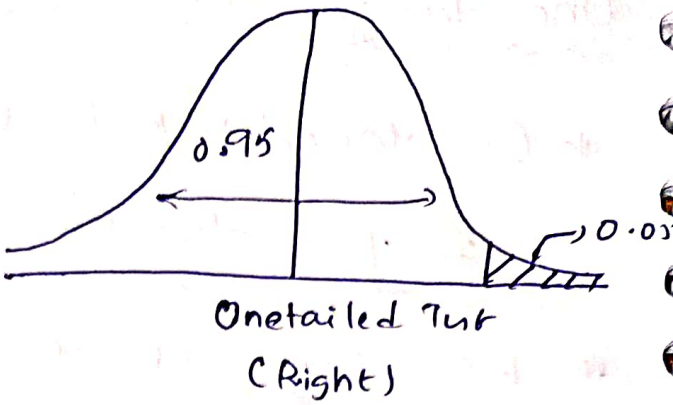
* If we expect only an increase or only a decrease (not both).

Types of One-tailed Tests:-

* Right tailed test
* Left tailed test.

* Right tailed Test:-

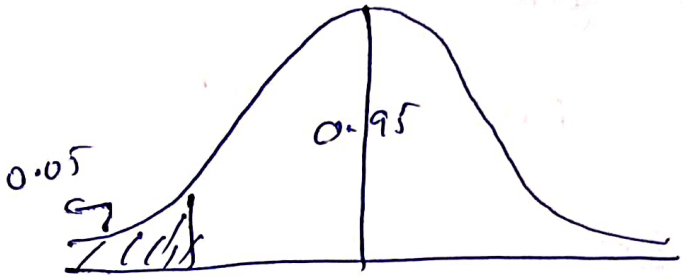
→ Testing if the Sample mean is Greater than the Hypothesited mean.



Onetailed Test (Right)

* Left tailed Test:-

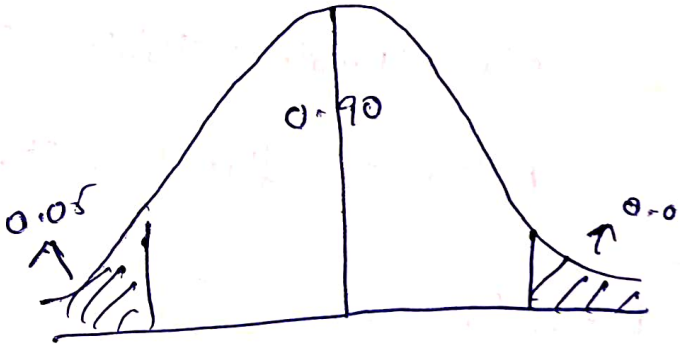
→ Testing if the sample mean is lurthan the Hypothesited mean.



Left tailed Test

Two tailed Test (Non-Directional Test)

* A two-tailed Test is used when we are testing for any Significant difference (Either an increase or Decrease).



Two tailed test

* We check if the Sample mean is Significatly different from the hypothesized mean, without Specifying direction.

Example:-

* A Railway claims the average delay is 10 min after introducing a new system.

Hypothesis:-

* Null Hypothesis (H_0):- The average delay 10 min
* Alternative Hypothesis (H_1):- The average delay is not 10 min (could be more or less)

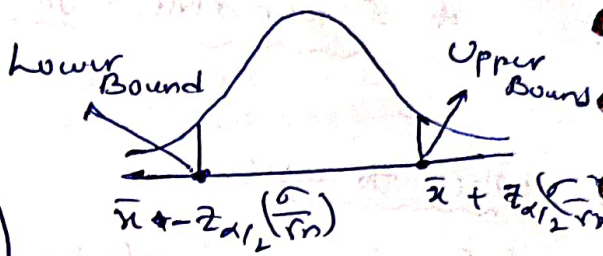
When to use:-

* If we are testing for any change, with out specifying wheather it will be higher or lower.

Confidence Interval

* It is a range of values that likely contains the true population parameter (Eg., mean or proportion) with a certain level of confidence.

* It helps in Estimating the uncertainty of a sample statistic.



formula :-

$$CI = \bar{x} \pm z_{\alpha/2}^{x}\left(\frac{\sigma}{\sqrt{n}}\right)$$

Here, $\bar{x}$ = Sample mean

z = z score (based on confidence level)

σ = population standard deviation

n = sample size

Example:-

1. Railway Monitoring - Estimating average train delay with confidence interval

2. Manufacturing - Checking if the average weight of a ~~population~~ product is within an acceptable range

Point Estimate

The value of any statistic that Estimates the value of a parameter.

Ex:- Inferential Stats

Sample mean ⟶ population mean

$\bar{x}$ ⟶ μ

$\bar{x} = 2.9$ $\quad$ $\mu = 3$

Margin of Error.

Confidence Intervals

point Estimate ± Margin of Error

$$\bar{x} \pm Z_{\alpha/2}\left[\frac{\sigma}{\sqrt{n}}\right]$$

⟶ Standard Error

T-test :-

* Used to compare the means of two groups when the sample size is small ($n \geq 30$) and population variance is unknown.

Types of T-test :

→ One Sample T-Test :- ($\bar{x}$) (μ)

* Compare the mean of a sample to a known population mean.

Formula :-

$$t = \frac{\bar{x} - \mu}{\frac{s}{\sqrt{n}}}$$

Ex :- The company has a sample of 30 Employes and wants to check if their average productivity score(75) differs from the industry standard. (70)

Hypothesis :

* H_0 : There is no significant difference ($\mu = 70$)
* H_1 : ~~Altern~~ The Employee's productivity is different from 70 ($\mu \neq 70$)

where $\bar{x} = 75$ (Sample mean)

$\mu = 70$ (population mean)

s : Sample Standard deviation

$n = 30$ (Sample size).

→ Two Sample (Independent) Test:-

* Compare the means of two independent Groups.

Formula:- $$t = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}}$$

Ex:- The company wants to compare IT employees & HR Employees, to see their productivity levels are different

Hypothesis:-

* H_0 : There is no significant difference ($\mu_1 = \mu_2$)
* H_1 : Productivity levels are different ($\mu_1 \neq \mu_2$)

Here, $\bar{X}_1, \bar{X}_2$ = Sample means of both groups.

S_1^2, S_2^2 = Sample Variances

n_1, n_2 = Sample sizes.

Paired T-test (Dependent)

* Compare the means of the Same Group before & after treatment.

Formula

$$t = \frac{\bar{D}}{\frac{s_D}{\sqrt{n}}}$$

Ex:- The company tests if Employee's productivity improved after a training program.

Hypothesis:-

* H_0 : No improvement after training ($\mu_{before} = \mu_{after}$)
* H_1 : productivity improved ($\mu_{before} \neq \mu_{after}$)

Here,

$\bar{D}$ = Mean of the differences (Before - After)

s_D = Standard deviation of differences

n = Number of pairs.

Scenario ($n<30$)	Test to use
Comparing one group to known population mean (population variance is unknown)	One-Sample T-test
Comparing two independent groups	Two Sample T test (Independent Test)
comparing same group before & after Treatment	Paired T-test

Z-Test :-

* Used to compare the means of two groups when the sample size is large ($n > 30$) and the population Variance is known.
* Used to test population proportions.

Types of Z-Test :-

→ One Sample Z-Test :-

Compare the mean of a sample to a known population mean when the population variance (σ^2) is known and sample size is large ($n > 30$)

formula, $$Z = \frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}}$$

→ Two Sample Z-Test :-

Compare the means of two independent groups when population Variance is known & sample size is large

formula, $$Z = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

Scenario ($n > 30$)	Test to use
Comparing one groups to known population mean (Population Variance is known)	One Sample Z-test
comparing two independent groups	Two Sample Z-test

Anova Test:-

* Used to compare means of three or more Groups to check if at-least one group is significantly different.

Types of Anova:-

→ One way Anova:-

* Compare the means of more than two independent groups to determine if atleast one group is significantly different.

Formula, $$f = \frac{\text{Between-Group Variance}}{\text{within Group Variance}}$$

* Between Group Variance – Measures how much the group means different from the overall mean.

$$MSB = \frac{\sum n_i (\bar{x}_i - \bar{x})^2}{K-1}$$

* within Group Variance – Means how much the individual data point different from their respective Group mean.

$$MSW = \frac{\sum (x_{ij} - \bar{x}_i)^2}{N-K}$$

If F-Statistic is greater than the Critical Value from the F-table we reject H_0.

→ Two-way Anova (without Interaction):-

* Compares the means of groups across two independent Categorical Variables, without considering their interaction

Formula, $F = \frac{\text{Variance Due to Factor A or B}}{\text{Error Variance}}$

Ex:- The company now wants to check two factors.
1. Work Environment (Remote, Hybrid, Office)
2. Experience level (Junior, Mid-level, Senior)

→ They want to analyze wheather productivity is affected by work Environment, Experience level or Both.

→ Two way Anova (with Interaction Effect) :-

* In addition to analyzing the individual Effects of two factors, it also checks if the two factors interact with Each other.

formula, $F = \frac{\text{Variance Due to Interaction}}{\text{Error Variance}}$

Ex:- The company now wants to check wheather Experience level influences the Effect of work Environment on productivity

→ They want to analyze
1. main Effect of work Environment
2. Main Effect of Experience level
3. Interaction Effect b/w work Environment & Experience level

Scenario	Test to Use
Comparing one categorical factor with more than two groups	one-way Anova
comparing two categorical factors independently	Two-way Anova (without interactions)
Checking if two categorical factors interact with each other	Two-way Anova (with interactions)

Chi-Square Test :-

The Chi-Square (χ^2) test is used for categorical data to check whether there is a significant association b/w two variables or whether the observed data follows an expected distribution.

1. Chi-Square Goodness of fit test :-

To check if the observed frequency distribution of a categorical variable matches an expected distribution.

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

where,

O_i = Observed frequency

E_i = Expected frequency

2. Chi-Square Test for independence :-

To check if two categorical variables are independent or related

$$\chi^2 = \Sigma \frac{(O_{ij} - E_{ij})^2}{E_{ij}} \quad ; \quad E_{ij} = \frac{\text{Row total} \times \text{Column Total}}{\text{Grand total}}$$

Scenario	Test to use
Checking if observed data follows an expected distribution	Chi-square Goodness of fit
Checking if two categorical variable are related	Chi-Square Test for independence

Scenario	Test to Use
Comparing means of two groups, Small Sample size	T-test
Comparing means of two groups, Large sample size	Z-test
Comparing means of 3+ groups	ANOVA test
Testing relationship blw two categorical Variables	Chi-Square test

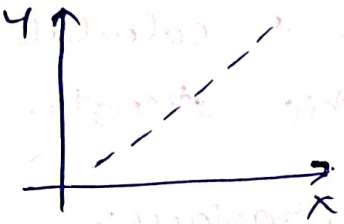
Covariance:-

* Covariance is a statistical measure that indicates the direction of the linear Relationship between two Variables.

Types:-

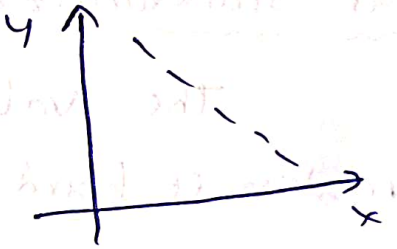
* **Positive Covariance:-**

When one Variables increases, the other also tends to increase



* **Negative Covariance:-**

When one Variable increases, the other tends to decrease



* **Zero Covariance:-**

No relationship between the Variables.

Formula,

For population,

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

For Sample,

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

Advantages of covariance:-

* **Indicate Relationship Direction:-**

  Shows wheather two variables move together (positive or Negative relationship)

* **Useful in Portfolio Mangement:-**

  Helps in diversifying investments by analyzing how asset prices move together.

* **Foundation for correlation:-**

  Used to calculate correlation, which normalizes the relationship strength.

Disadvantages of covariance:-

* **No Standardized Scale:-**

  The value depends on the units of the variables, making it hard to interpret.

* **Doesn't Show Strength:-**

  A large covariance doesn't always mean a strong relationship.

* **Sensitive to Outliers:-**

  Extreme values can distort the covariance calculation.

Correlation:-

* Correlation measures both the direction & Strength of the relationship between two variables.

* It is a standardized version of covariance, making it easier to interpret.

Formula :-

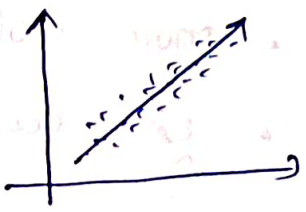
$$r = \frac{Cov(X,Y)}{\sigma_X \sigma_Y}$$

* It is derived from covariance and ranges betwin -1 & 1.

* Unlike covariance, which only indicates the direction of the relationship, correlation provides a standardize measure.

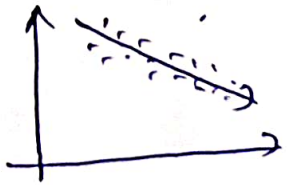
- Positive correlation (Close to +1) :-

  As one variable increases, the other variable also increase.



- Negative correlation (close to -1) :-

  As one variable increase, the other variable also decrease



- Zero correlation :-

  There is no linear relationship between the variables.

Types of Correlation:-

* Pearson correlation (linear Correlation):-
  - Measure the linear relationship between two variables.
  - Value ranges from −1 to +1
  - Ex:- If temperature increases, ice cream sales also increase.

* Spearman's Rank correlation:-
  - Measure the monotonic relationship (not necessarily linear)
  - works on ranked data instead of actual values.
  - Ex:- If a student studies more, their exam rank improves.

* Kendall's Tau Correlation:-
  - Measure the association between two ranked variables.
  - More robust for small datasets
  - Ex:- Customer satisfaction ranking Vs product sales rankings.

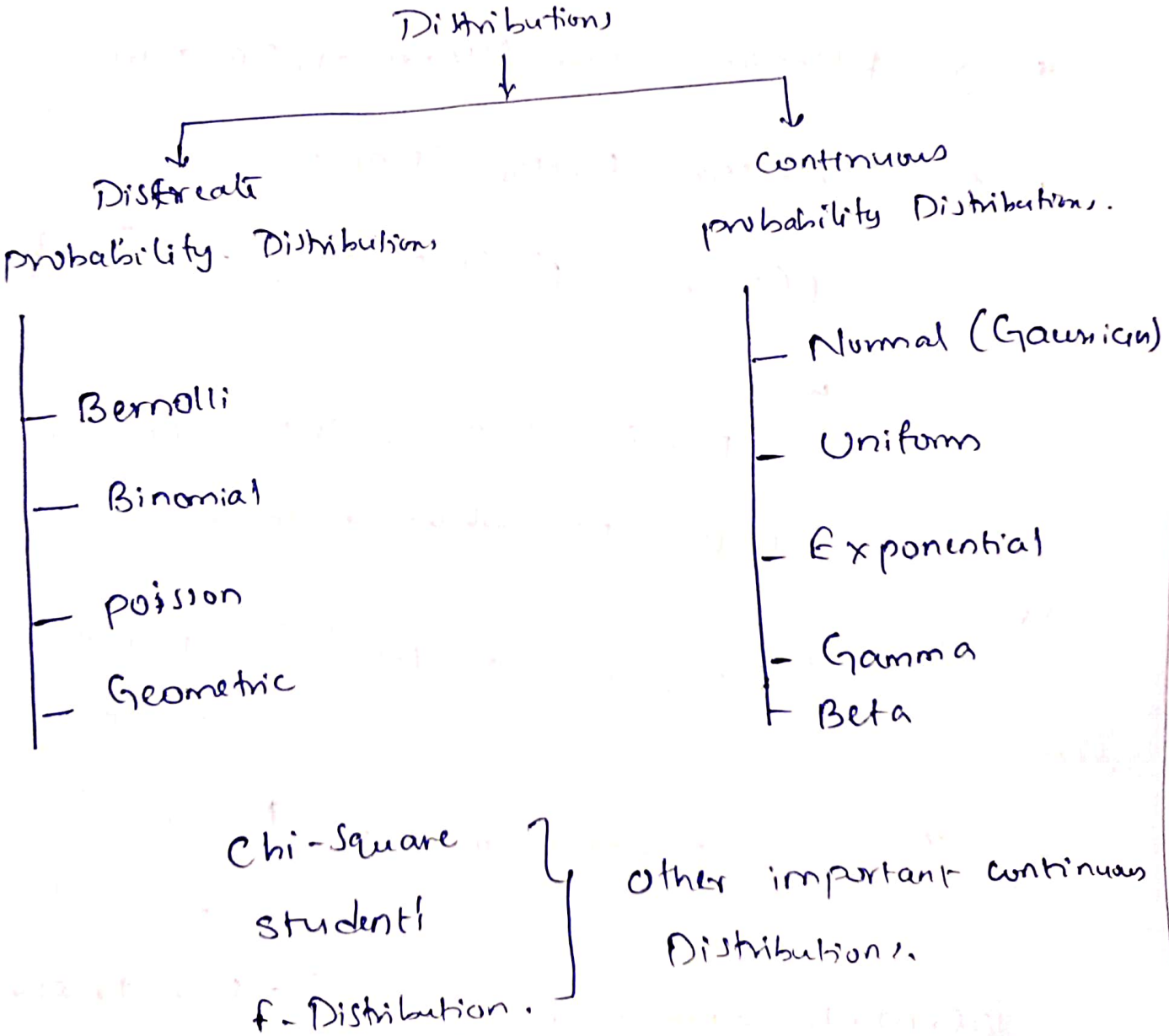
Distributions!-

→ It refers to how values of a dataset or a random variable are spread or arranged.

→ It shows the possible values and their frequency of occurance.

→ It helps in understanding patterns, making predictions & applying statistical tests.

Types of Distributions:-



1. Discreate probability Distributions:-

These are used for Variables that take Countable Values.

Ex:- Number of Succes, number of Events.

a. Bernoulli Distributions:- (Single trai) [Succes/Failure]

* Represents a Single trail with only two possible outcomes. Succes (1) or failure (0)

* Ex:- Flippling a coin (Heads = 1, Tails = 0)

* probability Mass function (PMF):

$$P(X=x) = p^x (1-p)^{1-x}, \quad x \in \{0,1\}.$$

where,

p - probability of succes. $(0 \le p \le 1)$.

1-p - probability of failure

if;

$x=0$, $P(X=0) = p^0 (1-p)^{1-0} = 1(1-p)$

$= 1-p.$

$x=1$, $P(X=1) = p^1 (1-p)^{1-1} = p(1)$

$= p.$

Bernoulli distribution, $P(X) = \begin{cases} 1-p, & \text{if } x=0 \\ p, & \text{if } x=1 \end{cases}$

6. Binomial Distributions:- (n - independent trails), count Success

→ The Binomial Distribution represents the probability of obtaining 'k' succes in 'n' independent trails,

→ where each trail is a Bernoulli trail with a success probability - P.

* Probability Mass function (PMF)

$$P(X=k) = {}^{n}C_{k}\, p^{k} (1-p)^{n-k}, \quad k = 0, 1, 2 \ldots n$$

where,

* ${}^{n}C_{k} = \frac{n!}{k!(n-k)!}$; a Binomial Coefficient, which representing the number of ways to choose k succes. from n trails.

* P = is the probability of succes in each trail

* 1-P = is the probability of failure in each trail

Ex:- Filip a fair coin (P : 0.5) → 3 times (n : 3)

calculate the probability of getting Exactly 2 heads (k : 2)

$$P(X=2) = \frac{3!}{2!(3-2)!} (0.5)^{2} (1-0.5)^{3-2} = \frac{3!}{2!(1)!} (0.25)(0.5)$$

$$= \frac{3}{1} \times 0.125$$

Exactly 2 heads in 3 filps ; 0.375 (or) 37% ; 0.375.

Poission Distribution:- (Time) count Event in a time / Space Interval

* This models the probability of observing a certain number of events in a fixed interval of time or space.
* Given that the event occurs at a constant average rate & independently of each other

→ probability mass function (PMF)

$$P(X = K) = \frac{\lambda^K e^{-\lambda}}{K!}, \quad K = 0, 1, 2, \ldots$$

where,

$X \rightarrow$ is the number of occurrences of an event

$\lambda \rightarrow$ (Lambda) is the average number of occurrence per time / space unit (also called the rate parameter).

$e \rightarrow$ Euler's Number (≈ 2.718)

$K! \rightarrow$ factorial of K, possible arrangements of K events

Ex:- Call center receives an average of 5 calls per hour ($\lambda = 5$), and we want to find the probability of receiving Exactly 3 calls in Hour ($K = 3$)

$$P(X=3) = \frac{5^3 e^{-5}}{3!} = \frac{125 \cdot e^{-5}}{6} = \frac{125 \times (2.718)^{-5}}{6}$$

$$= \frac{125 \times 0.0067}{6} \approx 0.1396$$

∴ Exactly 3 calls in an Hour is 13.96%.

3. Geometric :- (Number of trails until first Success)

→ The Number of Bernoulli trails needed to get the first Success in a Sequence of independent trails,

→ where Each trail has a Success probability P and a failure Probability (1-P).

* Probability Mass function (PMF) :

$$P(X=K) = (1-P)^{K-1} \cdot P \quad , \quad K = 1, 2, 3, \ldots$$

Where :

X - is the number of trails until the first Success occurs

P - probability of Success in a single trail.

1-P - probability of failure

K - The trail number where the first Success happen

Ex:- Flip a biased coin, where the probability of getting heads (Success) is P = 0.4. we want to find the probability that the first heads appear on the 3rd flip.

$$P(X=3) = (1-0.4)^{3-1}(0.4)$$

$$= (0.6)^2 \times 0.4$$

$$= 0.36 \times 0.4$$

$$= 0.144$$

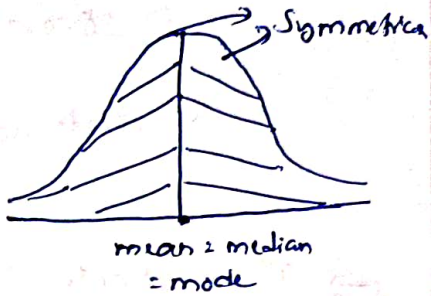
∴ Probability of getting the first heads on 3rd flip is 0.1444 (14.4%)

Continuous Probability Distributions:-

These describe variables that can take any value within a range (Ex:- Height, temperature, time).

a. Normal Distribution (Gaussian) :- (Natural phenomena)

* A bell-shaped curve where most values cluster around the mean.



Ex:- Heights of people, IQ scores.

Probability density function (PDF)

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}, \quad -\infty < x < \infty.$$

Normalizing factor ← $\frac{1}{\sigma\sqrt{2\pi}}$; $e^{-\frac{(x-\mu)^2}{2\sigma^2}}$ → Exponential part

where,

x - Random variable (value we evaluate the distribution at)

μ = Mean

σ - Standard deviation

σ^2 = variance

e = Euler's No., (≈ 2.718)

π = 3.1416.

Properties:-

1. Symmetrical : Centered at ~~0~~ with Equal probabilities on Both sides
2. Mean = Median = mode,
3. 68-95-99.7 Rule (Empirical)

Ex:- $\mu = 100$, $\sigma = 15$, $x = 120$

$$f(120) = \frac{1}{15(\sqrt{2\pi})} e^{-\frac{(120-100)^2}{2(15)^2}} = \frac{1}{37.67} e^{-0.8889}$$

$$\approx 0.0267$$

Probability density at $X = 120$, is ≈ 0.0267

b. Uniform distribution :- (Equal probability in a range)

* The Uniform distribution is a probability distribution where all outcomes are equally likely within a given range

* It is divided into:
  - → Discreate Uniform Distribution (finite set of outcomes)
  - → continuous Uniform Distribution (infinite set of outcomes with in an interval)

→ Discreate Uniform Distribution :- (All probabilities are equal)

PMF → $P(X=x) = \frac{1}{n}$, $x = x_1, x_2, \ldots x_n$

Ex :- Rolling a fair coin.

$x = \{1,2,3,4,5,6\}$, so $n = 6$

$P(x=3) = \frac{1}{6} = 0.1667$ (for rolling a 3).

→ Continuous Uniform Distribution :-

→ In which all values in a given range have equal probability.

PDF → $$f(x) = \begin{cases} \frac{1}{b-a}, & a \le x \le b \\ 0, & \text{otherwise} \end{cases}$$

Here, a, b - minimum & maximum values

$\frac{1}{b-a}$ - Height of the distribution is constant

Cummulative distribution function (CDF):-

$$F(x): \begin{cases} 0 & , x < a \\ \frac{x-a}{b-a} & , a \le x \le b \\ 1 & , x > b. \end{cases}$$

Ex1.

Bus arrives randomly b/w 10 Am & 10:30 Am.

If you arrive at a random time in the interval, the waiting time (X) is uniformly distributed btw

$a = 0$ (minutes, 10:00 Am)

$b = 30$ (minutes, 10:30 Am)

PDF, $f(x) = \frac{1}{30-0} = \frac{1}{30}$, $0 \le x \le 30$

Find the Probability that wait less than 10 mins.

CDF, $P(X \le 10)$:

$$P(X \le 10): \frac{10-0}{30-0} = \frac{10}{30} = 0.3333$$

waiting less than 10 min = 33.33%

C. Exponential Distribution:- [waiting times [time until the next event]]

→ Used to model the time until an event occurs in a poisson process.

→ It is commonly used to represent waiting times. Such the time b/w customer arrivals at a service ~~station~~ center or the time until a machine fail.

PDF, $$f(x) = \begin{cases} \lambda e^{-\lambda x} & , x \geq 0 \\ 0 & , x < 0 \end{cases}$$

Here, CDF, $F(x) = P(X \leq x) = 1 - e^{-\lambda x}$, $x \geq 0$.

x - time until the event occurs

λ = rate parameter (inverse of mean, $\lambda = \frac{1}{\mu}$)

e = Euler's number (≈ 2.718).

Ex: Call center receives 5 calls per hour. on average The time b/w calls follows an Exponential distribution

$\lambda = 5$ calls per hour

Find the probability that the next call occur with in ~~(DF)~~ 10 min (1/6 hour)

CDF : $P(X \leq 1/6) = 1 - e^{-5(1/6)}$

$= 1 - 0.4352$

$= 0.5648$

$\therefore$ 56.48% Probability that the next call arrive with in 10 min

d. Gamma Distribution:- [Generalization of Exponential]

+ Gamma Distribution used to model the time required for multiple independent Events to occur in a poinon process.

+ It is a generalization of the Exponential Distribution, which models the time until a single Event occurs

PDF; →

$$f(x) = \begin{cases} \frac{\lambda^k x^{k-1} e^{-\lambda x}}{\Gamma(k)}, & x \geq 0 \\ 0, & x \leq 0 \end{cases}$$

CDF, →

$$F(x) = P(X \leq x) = \frac{\gamma(k, \lambda x)}{\Gamma(k)}$$

$$\gamma(k, \lambda x) = \int_0^x t^{k-1} e^{-t} \, dt.$$

Ex:- Call center 5 call per hour on average, three calls arrive follows gamma distribution.

$K = 3, \quad \lambda = 5$

Find the Probability that three calls arrive within 30 min (1/2 hours)

$$P(X \leq 1/2) = \frac{\gamma(3, 5(1/2))}{\Gamma(3)}$$

$$= 0.875$$

87.5 % probability that third call arrive with in 30 min.

Beta Distribution:- [Probability distributions in Bayesian stats]

→ It is used to model probabilities and proportions, often in Bayesian statistics & machine learning

PDF, $$f(x;\alpha,\beta) = \begin{cases} \frac{x^{\alpha-1}(1-x)^{\beta-1}}{B(\alpha,\beta)} & , 0 \le x \le 1 \\ 0 & , \text{otherwise} \end{cases}$$

- α, β – Shape parameters (both must be greater than 0)

$B(\alpha,\beta)$ = Beta function, which normalizing the distribution

$$B(\alpha,\beta) = \int_0^1 t^{\alpha-1}(1-t)^{\beta-1}\,dt.$$

$$B(\alpha,\beta) = \frac{\Gamma(\alpha)\,\Gamma(\beta)}{\Gamma(\alpha+\beta)} \quad ; \quad \Gamma(n) = (n-1)! \text{ if } n \text{ is integer.}$$

CDF,

$$F(x) = I_x(\alpha,\beta) = \frac{\int_0^x t^{\alpha-1}(1-t)^{\beta-1}\,dt}{B(\alpha,\beta)}.$$