Module 3 Applied Statistics in Excel

- What is Statistics ...?

→ Statistics is the branch of mathematics that deals with collecting, organizing, analyzing, interpreting and present data.

→ It helps make sense of large amount of data and draw conclusion or make decisions based on their data.

- Use cases :- Summerization [avg, median]
  Test assumption
  to Find what really is going on
  to spot pattern & problems
  to small decision with using data

* DIKW Pyramid Levels :-

[Raw Facts]

1. Data :- Unprocessed, raw symbol or fact without context.
   Example : "98.6", "101.4", "99.1"
   Use :- Need to be processed to have meaning

[Processed Data]

2. Information :- Data that is organized or structured to have meaning
   Example : "Patients temprature : "98.6°F, 101.4°F
   Use : • Helps identify what is happening

[Interpreted Information]

3. Knowledge :- Information combined with experience or rules to draw conclusions.

Example :- "A temperature over 100.4°F usually indicates Fever.

Use :- Helps understand why something is happening.

[Applied Knowledge]

4. Wisdom :- The ability to make sound judgement and decision based on knowledge

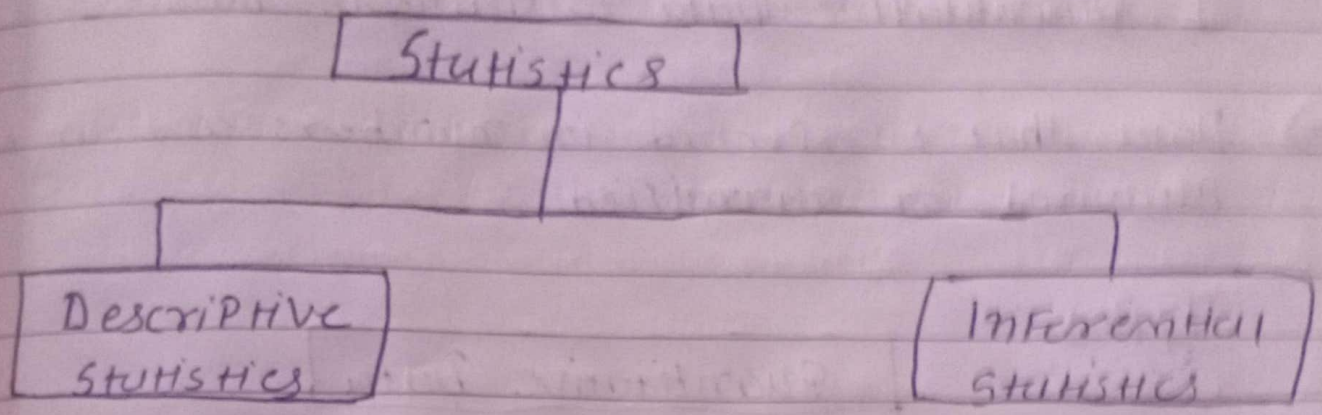
Example :- Prescribe antipyretics and monitor the Patient with 101.F for possible infection

Use :- Guides what should be done.

Summary Table :-

Level	Description	Example
① Data	Raw Facts	101.4 F
② Information	Structured Data	Patient temp is 101.4 F
③ Knowledge	Contexualized infer^n	101.4 F indicates Fever
④ Wisdom	Informed Desicison	Advise medicine & Observe Patient.

There are two types of Statistics



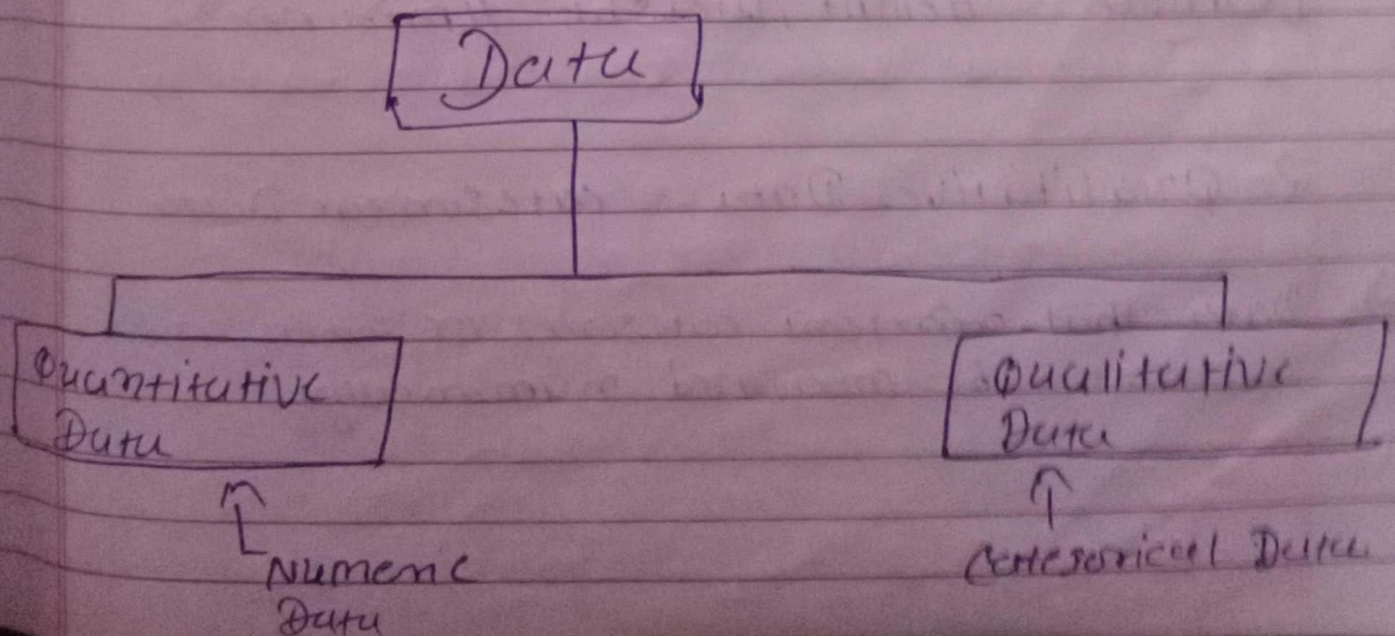
(1) Descriptive Statistics :- Describe & summarize the data.

Purpose :- To Present raw data in meaningful way
To identify patterns or trends.

(2) Inferential Statistics :- Make Predictions or generalization about population based on the sample.

Purpose :- To draw conclusion and make decisions under uncertainity.

There are two types of Data.



Graph :- Line
Histogram

1. Quantitative Data :- Numeric Data.

→ Data that is expressed in numbers and can be measured or quantified.

Types :-

- Quantitative Data
  - Discrete Data
  - Continous Data

(i) Discreate Data :- Data that can take only specific values. (usually counts)

Example :- Number of children in family, number of student in classroom

(ii) Continous Data :- Data that can be take any value within range (mesurments)

Example :- Height, weight temperature

2. Qualitative Data - catesorical Data.

→ Data that represent catesories or labels and cannot be measured numerically.

Graph: Bar
Pie

Types:

Qualitative Data

- Nominal Data
- Ordinal Data

(i) Nominal Data :- categories without any order or ranking.

Example :- Gender (male, Female)
marital status (single, married, Divorced)

(ii) Ordinal Data :- categories with meaningfull order or ranking but the diffrence between rank are not uniform

Example :- Education level : [High school, Bcholors, master]
survey rating [Poor, Fair, Good, excellent

* Scales of mesurment :-

→ To the diffrent ways in which variables or data can be classified and mesured

- Nominal scale
- ordinal scale
- Interval scale
- Ratio scale.

(categorical)

1. Nominal scale :- labels or names with no order or ranking

- Characteristics :- only classification is possible
  cannot perform mathematical operations

Example : Gender : Male, Female
Blood Type : A, B, AB
(chi. square) Colors : Red, Green, Blue

2. ordinal scale :- (ordered categories)

- categories with a meaningful order, but intervals between them are not uniform or known.

Characteristics :- can rank order items
- cannot do arithmatic operations
(U Test)

Examples : customer satisfaction :- Poor, Fair, Good
Education level :- High school, Bachelors, master
class : 1st, 2nd, 3rd.

3. Interval scale :- [ordered + Equal Intervals]

- ordered data with equal intervals but no true zero.

Characteristics :- Difference are meaningful
Ratio are not meaningful

Z Test, T, Test, anova, correlation

Example :- Temperatures in celsius or Fahrenheit
Data : 2000 2020.

4. Ratio Scale :- (Interval + True zero)

- like intervals scale, but has a true zero point

Characteristics :- Allows all arithmatic operation
Ratio are meaningful

Example Height, weight, age
income, distance, time.

- correlation, Anova

* Population :- The entire group of individuals items, or data want to study or draw conclusion about.

Size :- Usully large.

Example :- All student in University
All voters in country

* Sample :- A subset of the population selected for actual observation or analysis.

- 200 students randomly chosen from university
1000 Votes survey from across the country

* Descriptive Statistics :- is used to summarizing and organizing data.

→ Central Tendency :- mean, median, mode

→ Dispersion → variance, standard deviation

→ Shape :- Skewness & Kurtosis

→ visuals :- Bar, Histogram

What → Summarization

Why → Find patterns

When → Decision making

Where → Every where

How → Graphs

* Measure of Central Tendency :-

→ Measure of CT describe the center of Dataset they provide summary of the data with single value that represent the dataset as a whole.

$\boxed{\text{1. Mean}}$:- avg : Refer center of distribution

Population	Sample
μ	$\bar{x}$
$\mu = \frac{\Sigma x_i}{N}$	$\bar{x} = \frac{\Sigma x_i}{n}$

Ex. $\{1, 1, 2, 2, 3, 3, 4, 5, 5, 6\}$

$$= \frac{32}{10} = \boxed{3.2}$$

$\{1, 1, 2, 2, 3, 3, 4, 5, 5, 6, 100\}$

$$\frac{132}{11} = \boxed{12}$$

→ Mean is sensitive to outlier

outlier :- a datapoint who doesn't follows the

outlier :- A datapoint who doesn't follow the pattern or trend of data / or extreme points

$\boxed{2. \text{ Median}}$:- The middle value when data is ordered

$n \rightarrow$ odd
$n \rightarrow$ even

$\{1, 1, 2, 2, 3, 3, 4, 5, 5, 6\}$

Even :- $$\frac{\left(\frac{n}{2}\right)^{th} + \left(\frac{n}{2} + 1\right)^{th}}{2}$$

$$\frac{\left(\frac{10}{2}\right) + \left(\frac{10}{2}+1\right)}{2}$$

$$\frac{5^{th} + 6^{th}}{2} = \frac{3+3}{2} = \boxed{3}$$

odd :- $\left(\frac{n+1}{2}\right)^{th}$

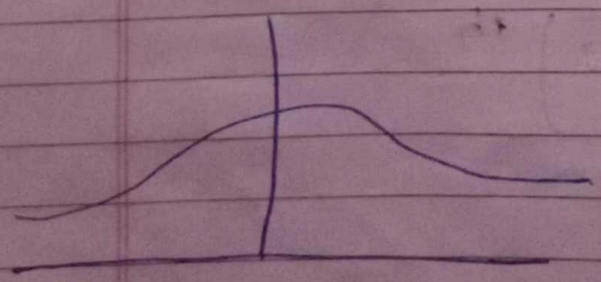
$\{1, 1, 2, 2, 3, 3, 4, 5, 5, 6, 100\}$

$= \left(\frac{11+1}{2}\right)^{th}$

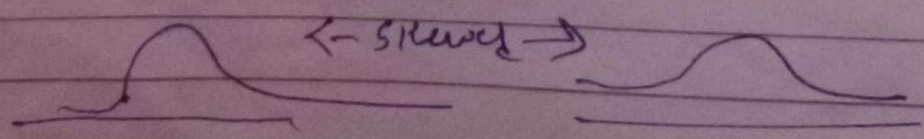
$= \frac{12}{2} = 6^{th} = \boxed{3}$

3. Mode :- The value that appear most frequently.

Types :- Unimodal :- one mode
Bimodal :- Two mode
Multimodal :- more than two modes
No mode :- All values occur equally



Normal → mean



* Measure of Dispersion :-

[5, 1, 1, 1, 2] → 2

[2, 2, 2, 2, 2] → 2

Same ?

MD describe the spread or variability of dataset. they indicate how much individual data point differ from the central value.

1. Variance :- It measure how far each datapoint in the dataset is from the mean

- Average of the squared diffrance from the mean.

High variance → more spread ←——→

low variance → less spread ←——→

Population

$$\sigma^2 = \frac{\Sigma (x_i - \mu)^2}{N}$$

Sample

$$s^2 = \frac{\Sigma (x_i - \bar{x})^2}{\underbrace{n-1}_{\text{Bessel's correlation}}}$$

2. Range :- Diffrance btw the highest and lowest values.

= MAX − MIN

3. Standard Deviation :-

→ Square root of variance It shows how much data varies from the mean.

Population

$\sigma = \sqrt{\text{variance}}$

$= \sqrt{\frac{(x_i - \mu)^2}{N}}$

Sample

$s = \sqrt{\text{variance}}$

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

IF SD = 0 → all data values are same

IF SD = high → data is widely spread

4. IQR : IQR stands for Interquartile range.

→ It is measure of statistical dispersion, or how spread out the values in a dataset are.

→ IQR measure the range between the first quartile (Q_1) the third quartile (Q_3)

$$IQR = Q_3 - Q_1$$

Q_1 = 25th Percentile

Q_3 : 75th Percentile

IQR :- middle 50% data

Uses :-

→ Detecting outlier :- Below $Q_1 - 1.5 \times IQR$

Above $Q_3 + 1.5 \times IQR$

Understanding spread :- Helps assess the variability in the central portion of the Data.

Percentage :- Way to express number as a fraction of 100.

$$\text{Percentage} = \left(\frac{\text{Part}}{\text{Whole}}\right) \times 100$$

$$= \frac{45}{60} \times 100 = \boxed{75\%}$$

Percentile :- indicate the relative standing of value within dataset. tells you the percentage of data values below particular point

→ Used in standarized tests, growth charts statistical analysis

[45, 50, 55, 60, 65, 70, 75, 80, 85, 90]

* Find student who scored 75th marks

$$= \left(\frac{\text{number of value below 75}}{\text{total number values}}\right) \times 100$$

$$= \frac{6}{10} \times 100 = 60^{th} \text{ Percentile.}$$

Q_1 = average 3rd & 4th values $= \frac{55+60}{2} = 57.5$

Q_3 = average 7th & 8th values $\frac{75+80}{2} = 77.5$

IQR: $Q_3 - Q_1 = 77.5 - 57.5 = 20$

* Skewness: is statistical measure that describe the asymmetry of distribution values in dataset. It tells whether data is skewed to left or right than.

What

Why: Understand the shape of data

→ Detect outliers and non-normality to dataset bias or imbalance.

→ Decide whether to transform data.

→ Choose appropriate statistical tests

When:- Before applying statistical tests or models assume normality.

→ When you notic outliers or strange distribution in your data

Types of skewness :-

1. Positive skew :- Tail on the right side longer

→ Most data points are concentrated on the left

→ mean > median > mode.

2. Negative skew :- Tail on the left side longer

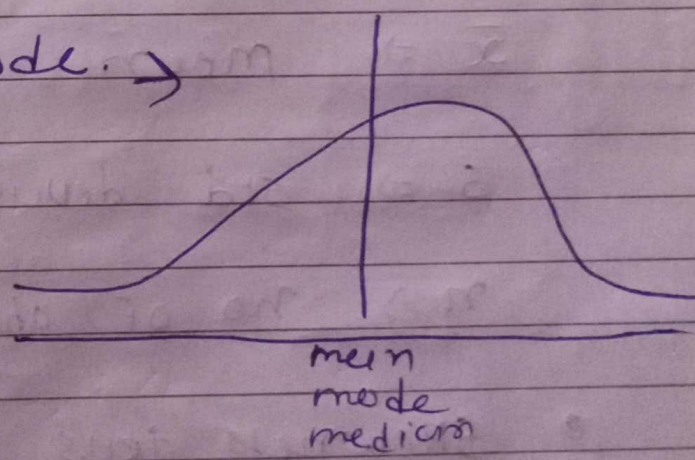
- Most data points are concentrated on the Right

→ mean < median < mode.

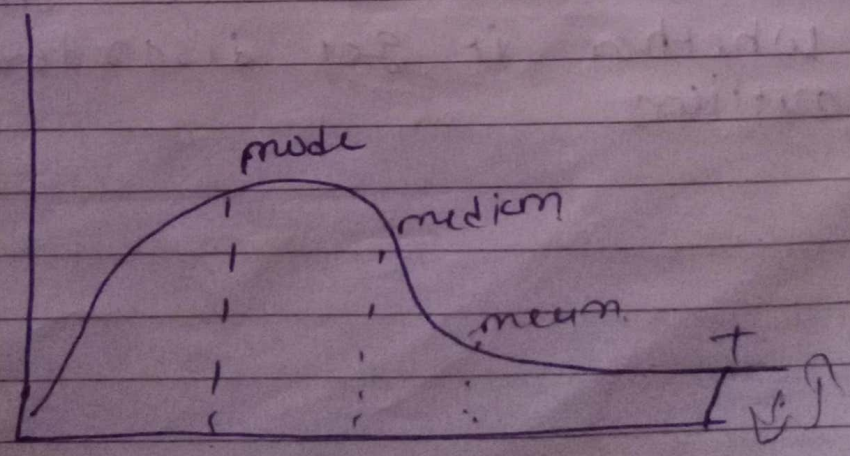
3. Zero skew (symmetrical Distribution)

- Data is evenly distributed around the mean

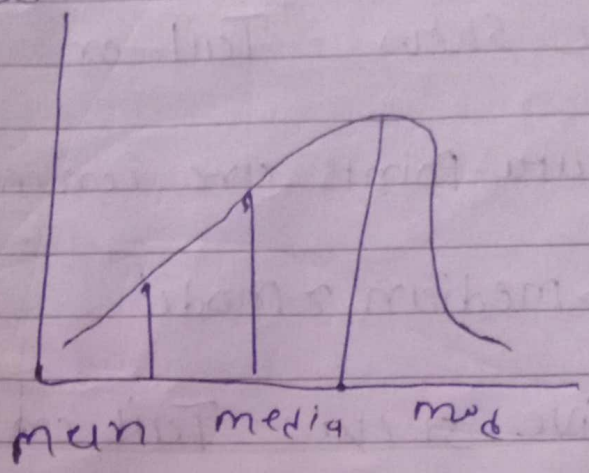
- mean = median = mode. →



Postive skew :-



Negative skew :-



Formula :-

$$\boxed{\frac{n}{(n-1)(n-2)} \sum \left[\frac{(x_i - \bar{x})^3}{s} \right]}$$

$x_i \Rightarrow$ Data Point.

$\bar{x} \Rightarrow$ mean

$s \Rightarrow$ std - deviation

$n \Rightarrow$ no of observation

- Skewness tells us above the mean whether it get disordered or not by outliers.

* Kurtosis :-

What :- It is statistical measure that describe the tailedness of data distrubution

→ It show whether the data has heavy or light tail compared to normal Distribution. Focus on the extremes rather to center.

Why :- Helps identify outliers or extreme values in data.

- Affect the risks in financial model, Useful in assesing whether dataset is appropriate for certain tests.

Where :- Finance, ML & DS, Psychology and for survey test data.

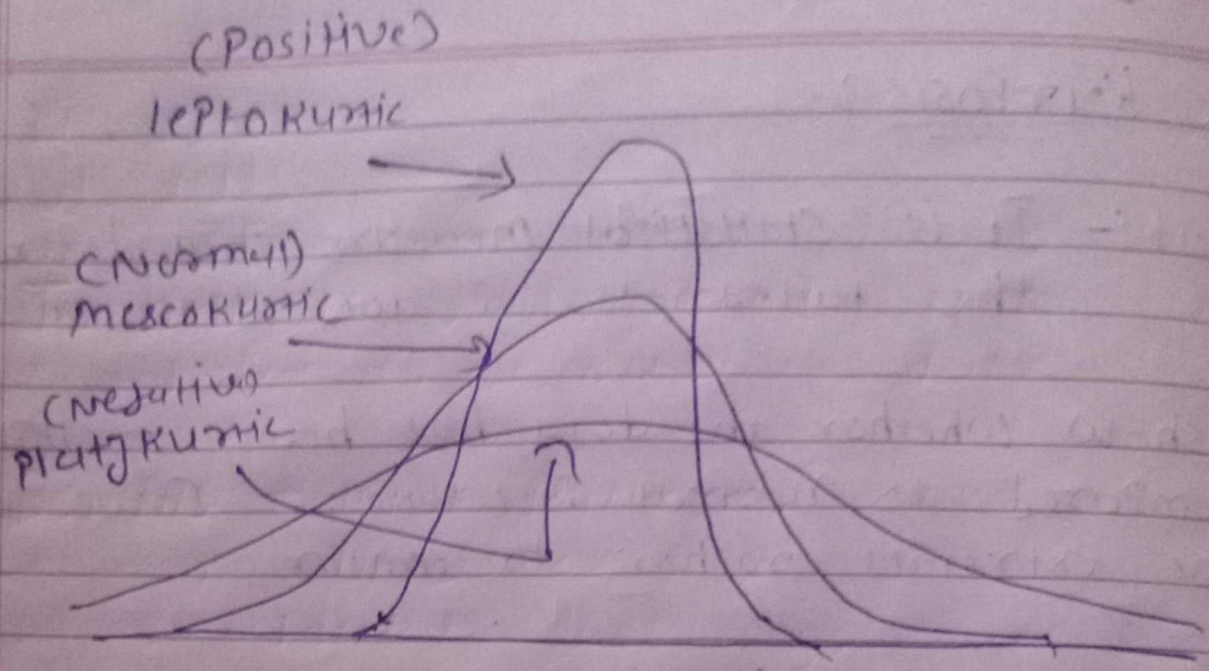
When :- When assessing the normality of data Before applying statistical tests that assume normal distribution.

How :- $$\frac{n(n+1)}{(n-1)(n-2)(n-3)} \sum \frac{(x_i - \bar{x})^4}{s} - \frac{3(n-1)^2}{(n-2)(n-3)}$$

$\bar{x}$ = mean

s = Standard deviation

n = number of observation



① Normal Distribution :- Kurtosis = 0.

② Leptokurtic : Kurtosis > 0 heavy tails
extreme outliers + Tall, thin peak

~~③ Leptokurtic :-~~

③ Platykurtic : Kurtosis < 0. Flat wide
Shape + ~~thin~~ light tail.
Few outliers.

④ Mesokurtic : Kurtosis = 0. like the
Normal, moderate tails.

Skewness	Kurtosis
→ measure asymmetry of the data.	measure tailedness or peakness
→ Direction of Data skew [left, Right, symmetric]	Data has heavy or light tails.
→ Types :- Positive Negative zero.	Types :- leptokurtic platykurtic mesokurtic
→ shows bias in data distribution Affects central tendency use case → Data visualization normal testing. transformation decision	→ show concentration in tails. Affect risk and outliers. usecase :- outlier detection Financial risk model tail analysis.
→ Focus :- mean and symmetry	Tail and peak of the distribution
→	

Visuals :-

→ To help us: Understand, summarize, Interpret, communicate data.

Benifits :- Spot Patterns,
Detect outliers
Identify distribution, central tendency

① Histogram :-

What :- It looks like bar chart but it shows Frequency distribution of continuous data.

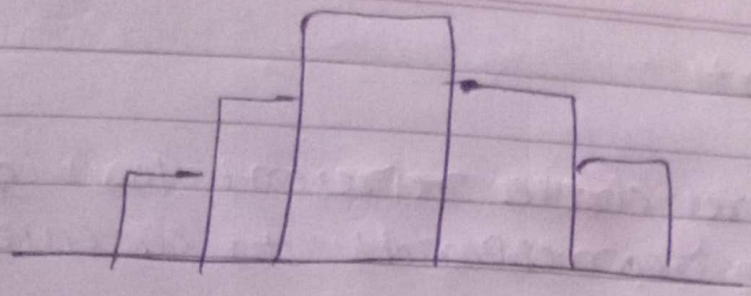
→ Each bar represent how many value Fall within a specific range.

Why :- Understand shape of data distribution visualize Frequency, detect patterns outliers and data spread.

Where :- Education, Healthcare, Finance, DS & DA.

When :- Dealing with large set of numerical data Data is univariate.

How :- Divide data into bins and plot Frequency of each bin.



② Bar chart :-

What :- The length / height of each bar is proportional to the value / frequency of the category.

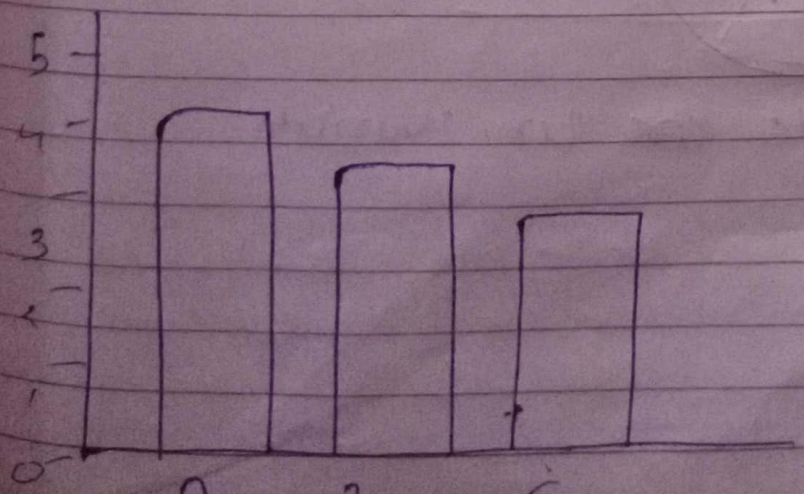
→ Bar are seprated not touching unlike histogram

Why :- Compare different categories clearly. to visualize trends in discrete data. Identify which categories are most / least common.

Where :- Business, Education, surveys, DA.

When :- Working with categorical or discrete data. Want to compare frequencies or values of different groups.

How :- Identify categories. Plot :- X-axis : category
Y-axis : frequency



(3) Pie chart :-

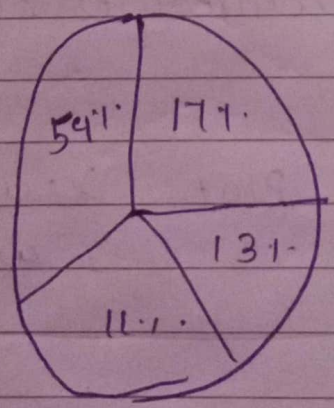
What :- The entire circle represent 100% of the data
Each slice corresponds to a category's contribution to the total

Why :- To show Proportion or Percentage at glance
to emphsize how each part contribute to the whole. Useful when want to compare sizes

Where :- Business, surveys, Education

When :- Want to emphasize Percentage-based
there are few categories (5 or fewer)
Precise comparision are not required Just normal idea

How :- Collect categorical data with associated values or Percentage
Divide circle into slice based on % contribution



→ It is Bivariate or uni-variate.

-Univariate

(4) Box Plot :- Box and Whisker Plot is graphical representation based on Five summary statistics :- minimum
First quartile
median
Third quartile

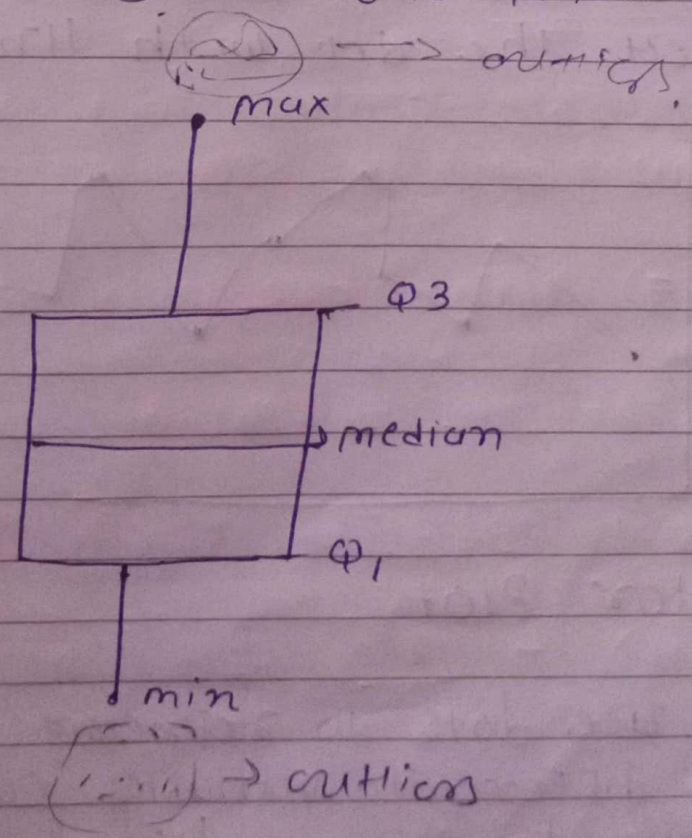
It shows outlier and spread

Why :- visually summarize the distribution of data
identify skewness, spread and outliers.
compare distribution across multiple groups

Where :- DS & ML, Healthcare, Education.

When :- Want to summarize numerical data with quartile
Comparing distribution across multiple categories.
For univariate or grouped numerical data.

How :- Calculate Q_1, Q_2, Q_3 IQR and Plot box whiskers accordingly.



(5) Line chart :- Bivariate

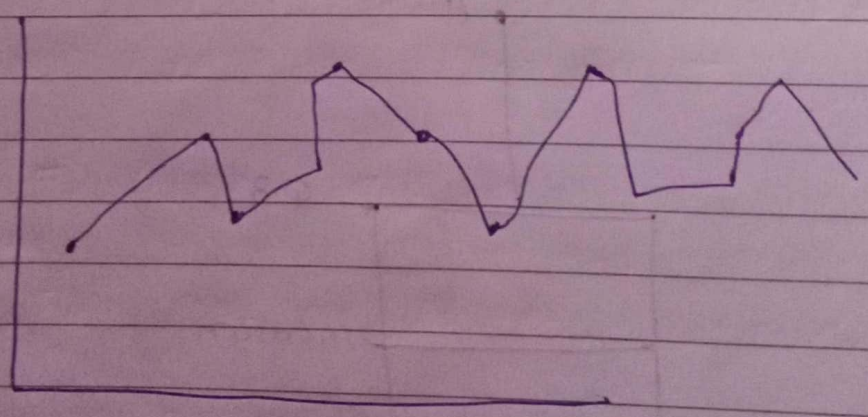
What :- that display data Points connected by straight lines typically show trends x-axis, y axis

Why :- track change/trends over continuous interval
Compare multiple data set same Period
visulize patterns growth, decline, peaks

Where :- Finance, Weather, Health

When :- Collected over a continuous Period (daily, monthly, Yearly..)
to identify trends or predict future values
to compare multiple time series

How :- Data Points for each x-y pair.
connect the dot with lines to show trend



(6) Scatter Plot :-

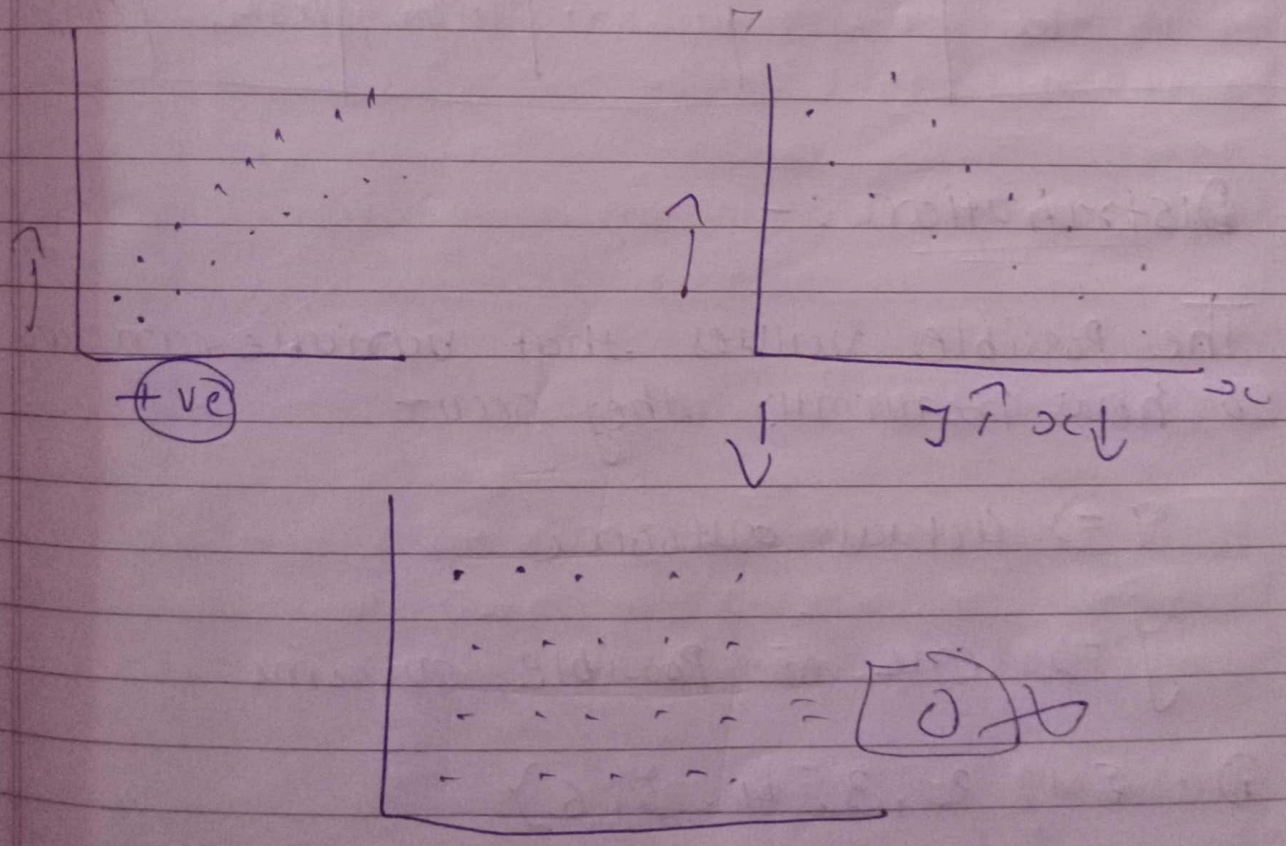
What :- That use dots to represent the value of two different variables. each dot represent a single data Point defined by two coordinates (x, y).

Why :- visualize correlation between two continuous variables. identify Patterns, clusters, outliers whether a linear or non-linear trend exist

When linear regression, study hours vs test scores
ASC vs BP.

When looking for Patterns, trends or correlation
detecting clusters, groups or outliers
two continuously variable want to examine their relationship

How: Plot dots for each Pair on 2D graph



Features	Histogram	Bar chart	Pie chart	Box chart	Line chart	Scatter
Data type	Numerical	Categorical	Categorical	Numerical	Time Series	2 Continuous
Purpose	Distribution	Group comp	Proportion	Spread & outliers	Trends	Correlation
Outlier shown	X	X	X	✓	X	✓
Best For	Frequency Distribution	Comparing groups	Showing % share	Detailed Distribution	Time based Data	Relationship Analysis

* Distribution :-

→ The possible values that variable can take & how frequently they occur

$y \Rightarrow$ actual outcome

$Y \Rightarrow$ one of possible outcome

$D = \{1, 2, 3, 4, 5, 6\}$

$$P(4) = \frac{1}{6}$$

$$P(Y = 4) = \frac{1}{6}$$

* Discreate Distribution :- These distributions model the Probabilities of discreate random variables

→ A random variable is discreate if it has finite and countable number of Possible outcomes.

→ Types :- Bernoulli Distribution

Binomial Distribution

Poisson Distribution

* Continuous Distribution :- Continuous Data can take any value within a given range. The range may be finite or infinite.

→ The weight of girl can be any value from 54. or 54.5 or 54.36 kgs.

→ Types :- Uniform, Normal, log-Normal, Exponential } Distribution.

* Discreate Distribution :-

① Bernoulli Distribution :- This distribution is generated when we perform an experiment once and it has only two Possible outcomes success or failure.

$X \sim Ber(P)$ / $1-P$: Failure of Prob

→ P is the Probability of success and 1-P is Probability of Failure.

→ Example:- A toss, If our call is head and if toss results in head, you win else lose there no midway.

$$P(x=r) = {}^nC_r (P)^r (1-P)^{n-r}$$

Binomial!

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

→ n is the total number of trials.
P is the Probability of success in trial
r is the number of success after n trial

Binomial Distri

* Binomial Distribution:-

→ This distribution is generated when we perform an experiment repeatedly n times and plot the probability each time

→ Probability of r success.

Properties:- $X \sim B(nP)$

① Each trial is independent since the outcome of the previous toss doesn't determine or affect to outcome of current

→ The total number of trials is fixed.

→ Each trial is binary i.e has only two possible outcomes success and failure.

→ The probability of success is the same for all the trials.

Bernoulli Distribution	Binomial Distribution
→ Used when want to outcome of single trial of an event	Used when want outcome multiple trials of an event.
→ Represent $x \sim\sim (P)$ P is probability of success	It is denoted as $x \sim\sim$ Binomial (n, P) where n is number of trials
→ mean, $E[x] = P$	mean $= E(x) = nP$
→ variance:- $Var(x) = P(1-P)$	$Var(x) = nP(1-P)$
→ Example:- suppose the probability of passing exam 80% and fail 20%. then Bernouli distribution can be used to model the passing or failing in such an exam	Example:- suppose the probability of passing exam 80% fail 20%. then if find the probability that student will pass in exactly 4. out of 5. use Binomial

Poisson Distribution :-

→ Estimate the average number of calls per hour but we cannot determine the exact number and exact time at which there is call.

→ Possion Distribution is applicable in situation where events occurs at random points of time and space wherein our interest lies only in the number of occurrence of the event

Example

→ The number of emersency calls recorde at a hospital in a day.

→ The number of thefts reported in an area on a day.

Used as an approximation of the Binomial if n is large and p is small

$$P(X=x) = \frac{e^{-\lambda} \lambda^{x}}{x!}$$

λ = average of times the events occured.

x = number of events in time

e = Eulers' number = $e = 2.71828$

* Bernouli Distribution:-

What:- Random experiments with only two Possible outcome ① success ② Failure

When:- Used when modeling single trial or event with Binary outcomes.

Where: In Probability theory, DS & ML.

Why:- simple yet powerful. For modeling binary events forms the basis other distribution Binomial

How: $P(x=1) = P$. $P(x=0) = 1-P$.

* Binomial Distribution.

What:- is the sum of multiple independent Bernouli trials. It models the number of success in n trials.

When:- Used when Performing the same binary experiment multiple times.

Where: In QC, survey sampling and market analytics

Why:- Helps model real-world Process where outcomes are yes/no and repeated multiple times.

How: $P:(X=K) = \binom{n}{K} p^{K} (1-P)^{n-K}$

n= numbers

P= success of Probability

* Poisson Distribution :-

What :- the number of events occuring in fixed interval of time or space when events happen independently constant rate

When: Used when modeling rare or infrequent events over time / distance.

Where: telecommunication, traffic flow, server request.

Why :- Useful when event count is known but no the number of trials

How :- $P(X=k) = \frac{e^{-\lambda} \lambda^{x}}{x!}$

---

* NORMAL DISTRIBUTION :

What :-

→ Also called Gaussian distribution is continuos probability distribution that is symmetrical and bell-shaped

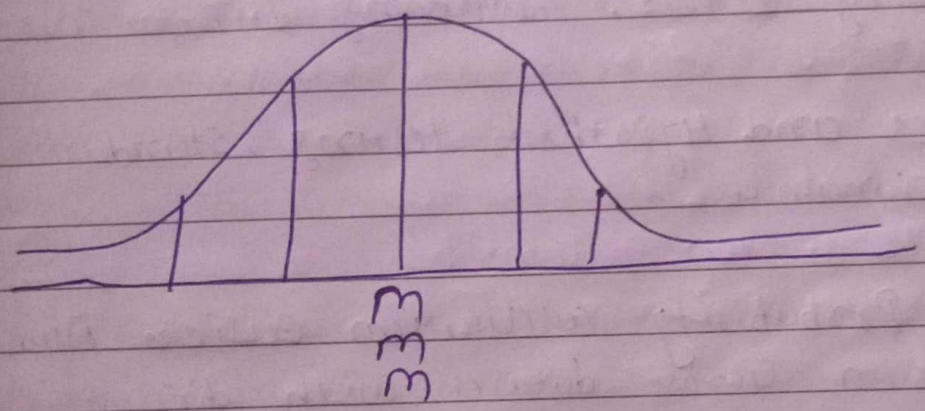
→ Most values cluster around the mean and fewer values away from the mean.

When: Dealing with naturally occuring continuous data like height, weight, exam scores

Where: Statistics and Data Analysis
ML, Finance, Economics

Why: It's mathematically tractable and has properties that simplify analysis.

- many statistical method like confidence intervals hypothesis testing assume normality



68% → $1\sigma \pm \mu$ $\quad \mu - \sigma \longleftrightarrow \mu + \sigma$

95% → $2\sigma \pm \mu$ $\quad \mu - 2\sigma \longleftrightarrow \mu + 2\sigma$

99.7% → $3\sigma \pm \mu$ $\quad \mu - 3\sigma \longleftrightarrow \mu + 3\sigma$

0.3 outliers

$$\frac{x - \mu}{\sigma} \Rightarrow \text{mean} = 0$$

standard deviation (σ) = 1

* STANDARD NORMAL DISTRIBUTION :-

What :- It is a special case of the Normal distribution where mean = 0
$\sigma = 1$
Also called 'Z-distribution' and values from this Z scores.

When :- Want to standardize the data.
- Performing Z-test, comparing different normal distribu

Where :- Statistics and Hypothesis testing, Standardize Data
P.C Finance.

Why :- make Probability calculation easier. Allows comparision across datasets with different mean different standard deviation.

How :- $Z = \frac{x - \mu}{\sigma}$

$Z \sim N(0,1)$

* Hypothesis Testing :-

-> It is statistical method used to make decision or inferance about population parameter based on sample data. It helps determine whether there enough evidence in sample to support specific claim about population.

1. Null Hypothesis :- It is the default assumption

- It state the there is no effect, no difrence or no change. assume H_0 is true until there is strong evidence against.

Ex :- A machine produce bulbs that last 1000 hours

$H_0 = \mu = 1000$.

2. Alternate Hypothesis :- It is the statement that trying to find evidence.

- It suggest that there is an effect, a difrence or a chance.

Types :- ① Two-tailed :- $H_1 : \mu \neq \mu_0$
(Parameter is not equal for value)

② Left tailed :- $H_1 : \mu < \mu_0$
(Parameter is less than for value)

③ Right-tailed : $H_1 : \mu > \mu_0$
(Parameter is greater than for value)

Confidence Interval :- (CI)

→ A CI is range of values derived from sample data that is likely to contain the true population parameter (like mean. population) with certain level of confidence.

→ A CI gives estimated range of valu which is likely to include an unknown population parameter. based on level (9

Ex:

A sample of 100 student has an average score of 70 with population sd of 10. Find 95% CI for population mean.

$\bar{x} = 70$

$\sigma = 10$

$n = 100$

$Z_{0.05} = 1.96$

$$CI = \bar{x} \pm Z_{s} \left(\frac{\sigma}{\sqrt{n}}\right)$$

$$= 70 + 1.96 \left(\frac{10}{\sqrt{100}}\right)$$

$$= 70 + 1.96$$

$$= \boxed{71.96}$$

Confidence Levels :-

→ A confidence level is the percentage that express how sure that confidence interval contain the true population parameter.

Confidence level	Significance level	Z-value
90%	0.10	1.645
95%	0.05	1.96
99%	0.01	2.576

90% :- Use narrower interval, less confidence. When precision is more important.

95% :- Standard in most studies.

99% :- More conservative wide interval used in critical application like medicine.

Loss of significance :-

→ loss of significance to the situation where a true difference or effect exist, but due to errors in data collection, calculation or method the test fails to detect it. Making the test less sensitive.

Causes :- Rounding errors
large sample error
Use inappropriate method.
Data entry or measurement error

Example

Suppose you calculate a test statistics and intermediate step are rounded.

- Exact t value : 2.015

- Rounded value : 2.0

leads to different P-value and possibly not reject Ho even if there significa

Types of ERROR :-

→ Two type of Error : ① Type - I (False Positive)

② Type - 2 (False Negative)

① Type I error :- Rejecting the null hypothesis when it is actually true.

- Denoted by α. , Reject Ho when Ho is true. False alarm (False Positive)

② Type II error :- Failing to reject the null hypothesis when it is actually false.

- Denoted by β, Fail to reject Ho when Ho is False. missed detection (False Negative

Rejection Region Method :-

→ The rejection region method is a classical way to perform hypothesis testing by comparing test statistic to critical value.

1. H_0 , H_a

2. α → significance level $\alpha = 0.05$

3. Assumption / value known

4. Decide the Test

5. value

6. Test conduct

7. Reject / Accept decision

8. Result.

1. State the hypotheses :-

H_0 :- Null Hypothesis

H_a :- Alternate Hypothesis

2. Set the significance level (α)

Common values : 0.05, 0.01, 0.10

3. Make Assumption: Sample is random
Data follows the correct distribution
Population Sd know or unknown.

4. Decide the Type of Test:

Z-Test: When Population σ is known
t-Test: " " " " unknown

5. Determine critical value:

Based on significance level and test type
use Z table.

6 conduct test: calculate Z or t.

7 make decision: If test statistic falls
in the rejection region reject H_0.
else fail to reject H_0.

8 Result: Interpret the Result in the
context Problem

Known σ. For Z score

Population: $\frac{x-\mu}{\sigma}$

Sample: $\frac{x-\mu}{\sigma/\sqrt{n}}$

For T. sample :- $\frac{x-\bar{x}}{s/\sqrt{n}}$ DF = $n-1$

Ex 50 Per Day. $\sigma = 5$. After training 30 emp to → 53 Per unit.

1. Ho :- $\mu = 50$

2. Ha :- $\mu > 50$

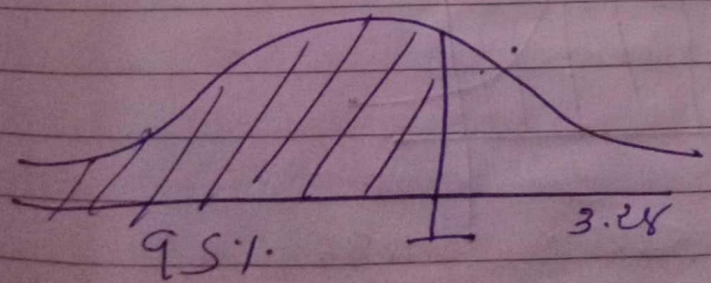
3. $\alpha = 0.05$

$\sigma = 5$. $n = 30$. random (normality)

4. Z Test :-

5. Z value $= \frac{x-\mu}{\sigma/\sqrt{n}} = \frac{53-50}{5/\sqrt{30}}$

$= \frac{3}{5} \times \sqrt{30} = \boxed{3.28}$



$\alpha = 0.05$
$= 1.65$

7. Ho → Reject.

$\mu = 50$ Pass $\quad \sigma = 4P$

$n = 40 \quad \bar{x} = 49$

1. $H_0 : \mu = 50 \quad H_u : \mu \neq 50$

2. $\alpha = 0.05$

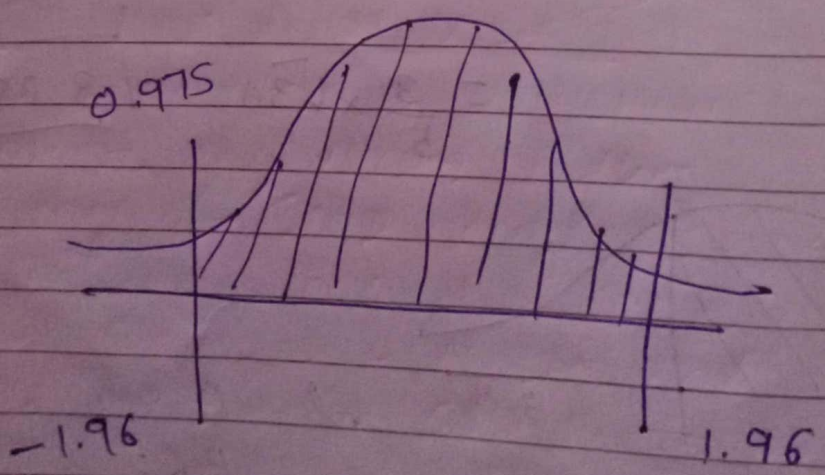
3.

4. Z. Test

5. $Z = \frac{x - \mu}{\sigma / \sqrt{n}} = \frac{49 - 50}{4 / \sqrt{40}}$

$= \frac{-1}{4} \times \sqrt{40}$

$\frac{-\sqrt{40}}{4} = \boxed{-1.58}$



7. Accept H_0

8. $\mu = 50$.

P value

$p \leq 0.01 \rightarrow$ very strong.

$.01 \leq p \leq 0.05 \rightarrow$ moderate

$p > 0.05 \rightarrow$ weak.

$p > .0.1 \rightarrow$ ✗

* Z-Test :-

→ The Z test is a statistical test used to determine whether there is significant difference between sample data and the population or between two independent sample datasets when the sample size is large and the population sd is known

→ The Z-test is particularly useful when you want to compare a sample mean to a population mean or compare two population proportions

→ Unlike the t-test which is used when the sample size is small and the population sd is unknown the Z-test is used when the sample size is large and the ~~population~~ population sd is known

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

Example

- Suppose the average weight of a population is 70 kg with $\sigma = 8$. A sample of 36 people has an average weight of 73 kg significance level at $\alpha = 0.05$?

Step 1 $H_0 : \mu = 70$
$H_a : \mu \neq 70$

Step 2 $\alpha = 0.05$, $Z_{0.025} = \pm 1.96$

Step 3 : $Z = \frac{73-70}{8/\sqrt{36}} = \frac{3}{1.33} = 2.25$

Step 4 : $2.25 > 1.96$ Reject H_0

T- Test :-

→ The t-test is a statistical test used to compare the mean of two groups to determine if they are statistically significant different from each other. It is commonly used when the sample size is small ($n < 30$) and population sd is unknown

→ The t-test helps to evalute whether the means of two datasets are different or if any observed difference is due to random chance.

Types of t-test :-

(i) one-sample t-test :- compare the mean of a single sample to known population mean

(ii) Independent two-sample t-test :- compare the means of two independent groups

(iii) Paired sample t-test :- compare the mean of two related groups or measurements

1. one sample t-test :-

$$t = \frac{\bar{x} - \mu}{s/\sqrt{n}} \qquad \boxed{DF = n-1}$$

2. Independent two sample t test :-

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{n_1} - \frac{s_2^2}{n_2}}} = \frac{\bar{x}_1 - \bar{x}_2}{\frac{s_1}{\sqrt{n_1}} - \frac{s_2}{\sqrt{n_2}}}$$

$$DF :- \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\frac{\left(\frac{s_1}{n_1}\right)^2}{n-1} + \frac{\left(\frac{s_2}{n_2}\right)^2}{n_2-1}}$$

3. Paired t-test :- $t = \frac{\bar{x}_d}{s_d/\sqrt{n}}$

$\bar{x}_d$: diff of mean

s_d :- " " sd

Chi-square (χ^2) Distribution :-

→ The chi-square distribution is a right skewed probability distribution used primarily in hypothesis testing and confidence interval estimation, especially when dealing with categorical data.

- only positive values.
- skewed right, but become more symmetric as degree freedom increase

$$\chi^2 = \frac{\Sigma(O-E)^2}{E}$$

O :- observed freq
E :- Expected freq

$$E = \frac{\text{Row total} \times \text{Col total}}{\text{Grand total}}$$

DF :- $(r-1) \times (c-1)$

Example

	IS	NS	Total
HS	50	70	120
G	90	60	150
PG	20	10	30
Total	160	140	300

DF

	S	NS
HS :	$\frac{120 \times 160}{300} = 64$	$\frac{120 \times 140}{300} = 56$
G :	$\frac{150 \times 100}{300} = 80$	$\frac{150 \times 140}{300} = 70$
PS :	$\frac{30 \times 160}{300} = 16$	$\frac{140 \times 30}{300} = 14$

E

$$\chi^2 = \frac{\Sigma (O-E)^2}{E}$$

$$HS/S = \frac{(50-64)^2}{64} = \frac{196}{64} = 3.06$$

$$HS/NS = \frac{(70-56)^2}{56} = \frac{196}{56} = 3.5$$

$$G/S = \frac{(90-80)^2}{80} = \frac{100}{80} = 1.25$$

$$G/NS = \frac{(60-70)^2}{70} = \frac{100}{70} = 1.43$$

$$PG/S = \frac{(20-16)^2}{16} = \frac{16}{16} = 1$$

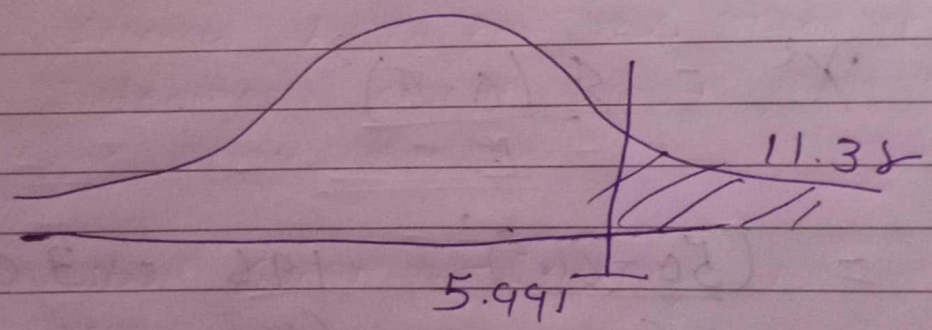
$$PG/NS = \frac{(10-14)^2}{14} = \frac{16}{14} = 1.14$$

$$\chi^2 = 1.14 + 1 + 1.43 + 1.15 + 3.5 + 3.06$$

$$\chi^2 = 11.38$$

DF: $(r-1)\ (c-1)$

$$(3-1)\ (2-1) = 2 \times 1 = \boxed{2}$$



Covariance :-

→ Covariance is a statistical measure the direction of the relationship btw two variables they change together.

- Covariance measure how two random variable vary together
- If the variable tend show similar behaviour the covariance is positive
- If one increase while other decrease the covariance is negative

$$COV = \frac{(x-\bar{x})(y-\bar{y})}{n}$$

$$= \frac{(x-\bar{x})^2}{n}$$

Features	Covariance	Correlation
Scale	unbounded ($-\infty$ to ∞)	Between -1 to
Interpretation	Direction only	Direction and strength
Units	Product of units $x \times y$	Unitless.

Types of sampling :-

1. Simple Random sampling :- Every individuals has an equal chance of being selected.

Ex :- Drawing names from hat.

2. Systematic sampling :- Select every k^{th} individual from list.

Ex: Every 10th customer on a list.

3. Stratified sampling :- Population is divided into subgroups based on characteristic and random samples are taken from each stratum

ensure representation from all key groups

4. cluster sampling:- Population is divided into clusters and entire clusters are randomly selected.

- Non-Probability Sampling:

1. Convenience Sampling:- sample is taken from a group that is easy to access

Ex: survey people from mall.

2. Judgemental or Purposive sampling:-
selected based on the researchers's judgement about who is most useful or representative.

3. Snowball sampling:- Existing participants refer new participants often used for hard to reach population.

4. Quota sampling:- Population is divided into groups and quota is set of each group participants are selected non randomly until the quo