STATISTICS

What is statistics

It is a branch of mathematics which helps to collect organize . analyze this data

USE Cases:

=> Summarization (Avg. mean, median, ...)
Helps to summarizing large datasets in a meaningful way
=> Test assumption.
=> to find what really is going on.
=> to spot pattern and problems.
=> to make smart decision with /using data.

DIKW pyramid

① Data : unorganized, unprocessed raw values, no context, no meaning, just facts.

Ex John Red 25 25/12/2025
(meaningless)

② Information : . organized data
- context, label

- Processed and organized data that is meaningful and useful

Date : ..... Page No. : .....

Ex: Name : John
Age : 25
Fav. Col. : Red
Reports, tables, charts.

(3) knowledge :- Interpreted Information
- Pattern
- Apply logic

Insight derived from analyzing information, often involving understanding relationships and pattern.

(4) wisdom : actionable insights (Applied knowledge)

types of stats.

(1) Descriptive
(2) inferential

(1) Descriptive : Methods to organize and summarize data using numerical and graphical techniques.

(mean, median, mode, Range, SD., charts (pie, Bar, Box Plots, Histogram)

Ex: calculating the average score of a class of a student.

(2) Inferential : techniques to make Predictions about a larger Population based on a sample of data.
(T-tests, chi-square, understanding relationships b/w variables)

Ex: Predicting election results based on a poll of a few thousand people.

Types of Data

- Quantitative (numerical)
- Qualitative (categorical)

(1) Quantitative Data : represents numerical values and can be measured or counted.

maths involve types graph - line, histogram

(1) Discrete Data
- finite values. countable (int)

Ex: No. of students in class, no. of employees.

(2) Continuous Data
- Decimals, fractional no. (float)

Ex: Height, weight, temp. time.

Date : ..... Page No. : .....

② Qualitative Data: represents characteristics or labels that describe attributes or qualities

No maths. Graph. Bar pie

- what type of data?
- category

types: (i) Nominal Data: categories without logical order

- No order

Ex colors, gender.

② ordinal Data: categories with a meaningful order or ranking but the difference between categories is not measurable

- Order matters.

Ex satisfaction Ratings (low, median, high) feedback, education level (high school)

Scales of measurements.

It defines how the data is classified compared and used mathematically.

Properties :- lube identity (lubeling)

- order (ranking)
- Equal Interval
- true zero

there are four main scale:

① Nominal : Represents data that can be categorized but not ranked.

- Naming data
- cannot perform arithmetic operations
- counting grouping

Ex Gender, region, country.

② ordinal : Represents data with a meaningful order.

- label or category with order
- counting

Ex Customer satisfaction (satisfied, Neutral, Dissatisfied).
Education level (High school, Bachelor's, Master's)

③ Interval Scale : Represents ordered data with equal intervals between values but no true zero point.

- Add, sub possible, but no multipli, div
- zero does not mean absence.

Ex Temperature (0°C does not mean absence of temp),
calender.

(4) Ratio: Represents numerical data with equal intervals and true zero point.

- Add., sub, multi, Div, are meaningful
- true zero point indicates the absence of the attribute

Ex Height, weight, Age, Income.

Summary

	Scale	type	Order	equal interval	true zero	Example
(1)	Nominal	categorical	NO	NO	NO	colours, genders
(2)	ordinal	categorical	yes	NO	NO	RANKS Ratings
(3)	Interval	Numerical	yes	yes	NO	Temp. calender
(4)	Ratio	Numerical	yes	yes	yes	Height weight, Age Income

Nominal Data :- mode, frequency, Percentage
- Bar, pie
- chi-square.

Ordinal data : Rank based measures
- Bar Charts
- U-test.

Interval scale: mean, median, SD
Histrogram, Boxplot
Z-test, T-test, Annova, correlation

Ratio: All mathematical operations
- lineplot, Box, Histogram
- correlation, Annova

Population: The entire group of individuals items or data want to study or draw conclusion about size. usually large. ex- all students in University.

Sample: A subset of population selected for actual observation.
ex 200 students randomly choosen from university.

Descriptive stats.

DS are the methods for summarizing key aspects / characteristics of data

- Descriptive statistic is a branch of statics that focuses on summarizing and organizing data to make it easily understandable.

Key components of DS

1. Central Tendency - mean, median, mode
2. Discrision - Range, variance, SD, IQR (Interquartile Range)
3. Shape - skewness, kurtosis
4. Visuals - Bar, Histogram

what - Summarization
why - To find pattern
when - Before decision making
where - Everywhere

1. measure of central Tendency:

measure of CT describes the 'center' of dataset.

- They provide the summary of the data with a single value that represents the whole dataset.

i) mean:
- The average of all data points
- Referes to the center of distribution

Population	Simple
μ - POP avg	$\bar{x}$: sample avg
$\mu = \frac{\Sigma x_i}{N}$	$\bar{x} = \frac{\Sigma x_i}{n}$

Ex $\{1,1,2,2,3,3,4,5,5,6\} = \frac{32}{10} = 3.2$

$\{1,1,2,2,3,3,4,5,5,6,\underset{\text{outlier}}{100}\} = \frac{132}{10} = 12$

mean is sensitive to outlier

② median:

The ~~midd~~ median is middle value of the data when

① sort

② n - even, odd

n-Even - $\frac{\left[\frac{n}{2}\right]^{th} + \left[\frac{n}{2}+1\right]^{th}}{2}$ n-odd - $\left[\frac{n+1}{2}\right]^{th}$

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

$= \frac{\left[\frac{10}{2}\right]^{th} + \left[\frac{10}{2}+1\right]^{th}}{2}$ $= \left[\frac{11+1}{2}\right]^{th}$

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

$= \frac{3+3}{2}$ $= 3$

$= 3$

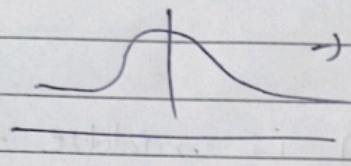
Outlier :- A datapoint(s) that doesn't follow the pattern or data trend of the data or are extreme point

Ex $\{1,1,2,2,3,3,4,5,5,6,\underline{100}\}$
outlier.

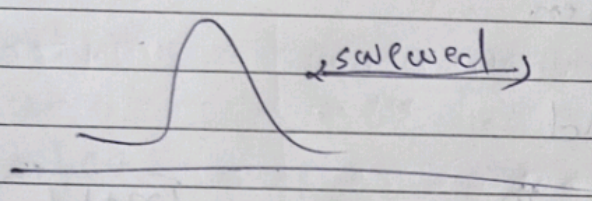
Date : ..... Page No. : .....

Mode :

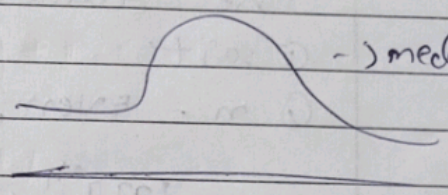
The most frequent value in the dataset



→ normal Kaussian → mean



← skewed →



→ median

(2) measure of Dispersion :-

It describes the spread of the data points.

- MD indicates how much individual data points differ from the central value

ex [5, 1, 1, 1, 2] - avg · 2 } same
[2, 2, 2, 2, 2] → avg · 2

MD checks whether the data is different or same

i) Range : The difference between the maximum and minimum value (MAX-MIN)

(2) Variance : It measures how far each data point in the dataset is from the mean

High variance - more spread
low variance - less spread

Date : ..... Page No. : .....

Population

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

Sample

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

Bessel's correction

(3) standard Deviation:-

SD is square root of variance, representing the average distance from the mean

Population

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

$$\sigma = \sqrt{\frac{\Sigma (x_i - \bar{x})^2}{N}}$$

Sample

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

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

I check actual data we don't compare square data with real/actual data.

(4) IQR (Interquartile Range):

The range between first (25%) and third (75%) quartiles.

- IQR measures the spread of the middle 50% of data.

Percentile? - It is value below which a certain percentage of observation lie.

Date : ..... Page No. : .....

Eg: [2, 2, 3, 4, 5, 5, 5, 6, 7, 8, 8, 8, 8, 8, 9, 9, 10, 11, 11, 12]

$n = 20$

percentage of 10 $= \frac{\text{no. of values below 10}}{n} \times 100$

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

$= 80\%$

25 %tile $= \left[\frac{\text{percentile}}{100} \times n\right] + 1$

$= \left[\frac{25}{100} \times 20\right] + 1$

$= [5] + 1$

$= 6^{th}$ index

$= 5$ - value

75 %tile $= \left[\frac{\text{percentile}}{100} \times n\right] + 1$

$= \left[\frac{75}{100} \times 20\right] + 1$

$= [15 + 1]$

$= 16^{th}$ index

$= 9$ - value

Five Number summary

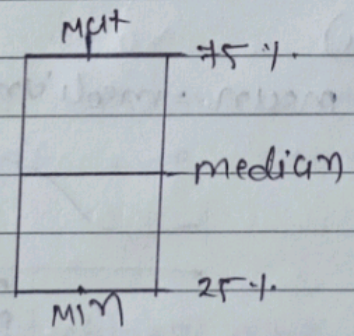
1. minimum
2. Q_1 - 25%
3. median
4. Q_3 - 75%
5. maximum

$IQR = Q_3 - Q_1$

Ex $= 9 - 5$
$= 4$

minimum — low-fence $= Q_1 - 1.5\ IQR$
$= 5 - 1.5(4)$
$= -1$

maximum - upper-fence $= Q_3 + 1.5\ IQR$
$= 9 + 1.5(4)$
$= 15$



→ Box plot

3. Shapes (Skewness and kurtosis)

⇒ ~~Sk~~ Skewness measures the anosymmetry of the distribution of data. It is used to understand the shape of data to choose appropriate statistical test.

Types:

Skewness value = 0. shape - symmetrical.
meaning : Data is evenly spread on both sides
(mean = median)

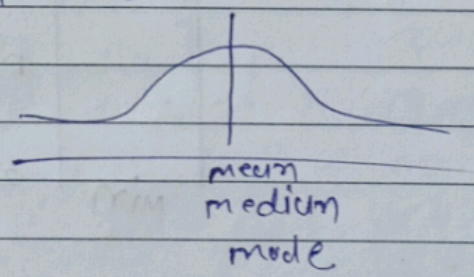
Skewness value > 0, shape - Positively skewed
meaning : tail is longer on the right (high outliers)
(mean > median)

Skewness value < 0, shape - Negatively skewed
meaning : tail is longer on the left (low outliers)
(mean < median)

Ex Income distribution is usually right-skewed a few people earn much more than most.

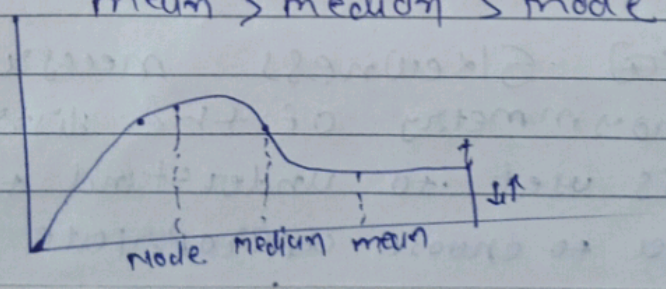
① zero skew

mean = median = mode



② Positive skew

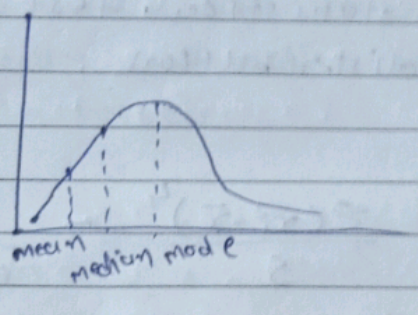
mean > median > mode



Date : ..... Page No. : .....

(3) Negative skew

mean < median < mode



Formula

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

x_i = data point $\bar{x}$ = mean

s : std deviation n . no of observation

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

=> Kurtosis

Kurtosis measures the "tailedness" or "sharpness of the peak" of the data distribution

- It shows whether the data has heavy or light tails compared to normal distributio

why?

Kurtosis helps to identify outliers or extreme volume in data.

where?

It is used in finance, ML, & DS etc.

Date : ..... Page No. : .....

Theory:

When assessing the normality of data before applying statistical tests that assume normal distribution

Formula:

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

$\bar{x}$ = mean
x_i = datapoint
s = std. deviation
n = no. of observation

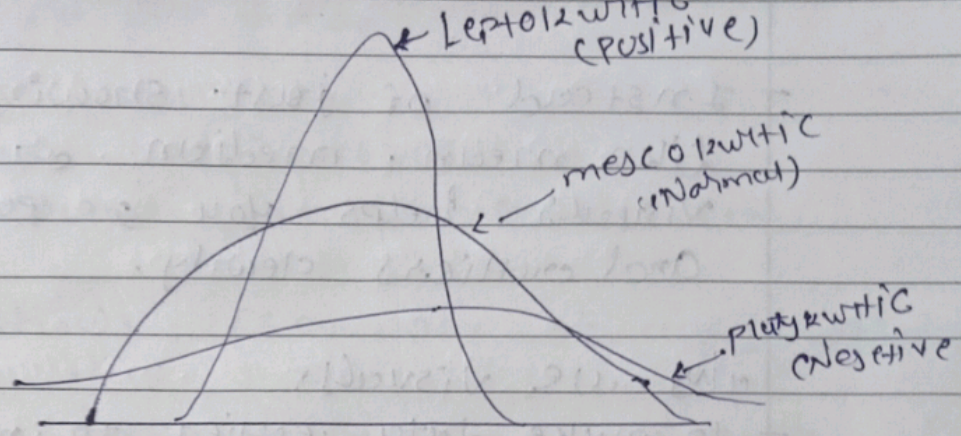
Types:

Type	Kurtosis value	Description
① mesokurtic	≈ 0	Normal distribution (moderate peak)
② Leptokurtic	>0	sharp peak and heavy tails (more outliers)
③ Platykurtic	<0	flat peak and light tails (less outliers)

Date : ..... Page No. : .....

Example:

Stock returns are often leptokurtic usually small changes but big gains or losses.



Summary of Skewness and kurtosis:

	Skewness	kurtosis
what	It measures asymmetry	It measures tailedness
why	to detect bias or imbalance	to detect risk of extreme outlier
	= skew ()	= kurt ()

Date : ..... Page No. : .....

Visuals

Visuals are charts or graphs used to visually summarize, understand, communicate or explain the data

- Instead of just showing numbers like mean, median or std deviation visuals helps you see patterns, trends, and outliers clearly.

why use visuals:
- to make data easier to interpret
- to spot trends and pattern quickly
- to detect outliers or skewed distributions
- to communicate insights clearly.

Histogram (univarient)
- It shows the frequency of data range

what? - A bar chart that show how data is distributed into intervals (bins)

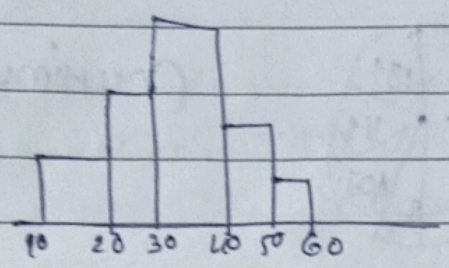
why? - understand frequency distribution shape → normal, skewed.

where? - Business, eduction, age, income, sales

How? - Group data → range → plot frequency

Date : ..... Page No. : .....

Ex



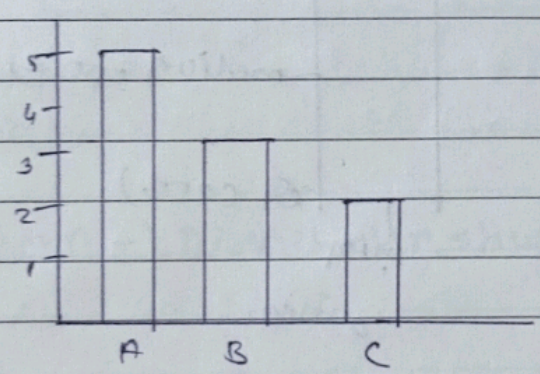
Marks - (10-20)
(20-30)
(30-40)

Bar chart (Bivarient)
- compare quantities across items

what - shows frequency of categorical data

why? - to compare different categories

where? - survey, results, sales



A = 5
B = 3
C = 2

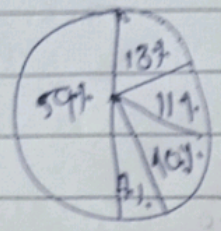
Pie chart (may be univarient bivarient)

Proportions of categories

what? - A circular chart divided into slices to show proportion.

why? shows proportion or %.

where? market shares, Budgets.



(Maximum 4-5 category is used)

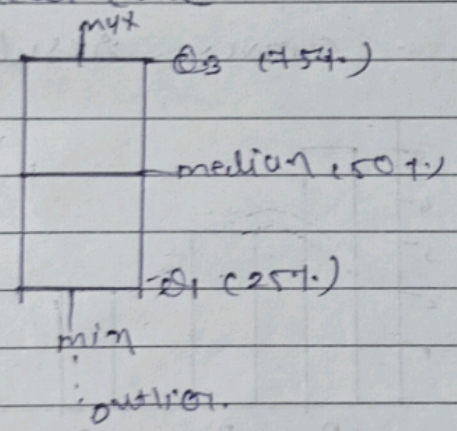
Box plot (univarient)

continuous numeric data

- median, spread, outliers, quartiles
- to detect outliers.

why? spot, spread, median, outlier

what? DA, Healthcare

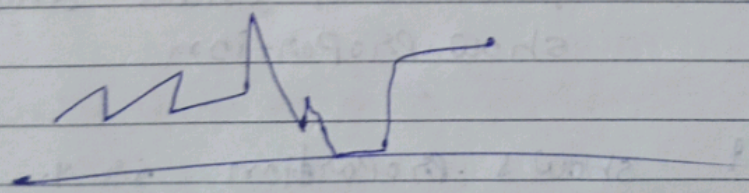


line chart (Bivarient)

what? show trends over the time

why? to track change over the time

where? stock prices weather



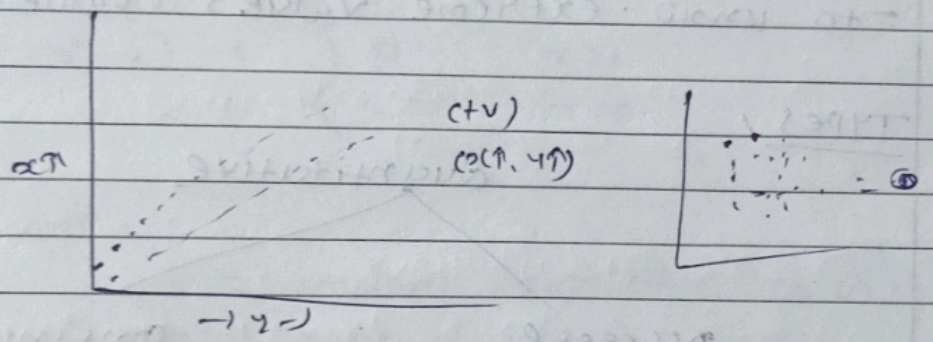
Date : ..... Page No. : .....

Scatter plot

what? It shows the relationships between two numeric variables

why? to explore relationships or correlation between two variables.

where? DA, Business Cadvertising spend vs sales



summary

Histogram -) Distribution shape -) continuous data, frequency & intervals

Bar -) comparing categories -) categorical data -) frequency or value

pie -) proportions -) cate / numerical data -) part of whole

Box -) spread & outliers -) continuous data -) range, median, quartiles

line -) trends overtime -) time-series -) pattern

Scatter -) relationships b/w variable -) two numeric data -) correlation / pattern

Distribution

The possible values that a variable can ~~t~~ and how frequently they can occur.

Why distribution is used?

- to know the center tendency (CT) of data
- to know how data is spread
- what is the shape of data · skewed/symmetric
- to know · extreme values (outliers)

Types.

Quantitative

Discrete (finite no) | continuous (infinite no)

X variable | $\sim$ follows | N dist. | (μ, σ) characteristics

Discrete Distribution (finite no)

(i) discrete distribution is a type of probability distribution that deals with countable finite outcomes (like whole number)

Types
1. Bernolli Distribution
2. Binomial Distribution
3. Poisson Distribution

Bernoulli Distribution

The bernoulli distribution is used for experiments with exactly two possible outcomes

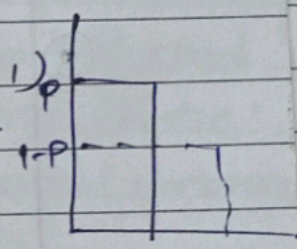
It has one trail and two outcomes (success, failure.

example - toss, answer (correct or incorrect)
(Head, tail)

$$P(X=x) = \begin{cases} p & ; x=1 \\ 1-p & ; x=0 \end{cases}$$

where.

x = random variable (0 or 1)
p = Probability of success
$1-p$ = Probability of failure



$X \sim Bern(p)$

Binomial Distribution

It is a sequence of identical Bernoulli events

The number of successes in a fixed number of independent Bernoulli trials where each trial has only two outcomes - success or failure.

$$P(y) = p$$
$$P(n-y) = (1-p)^{n-y}$$

$$P(y) = {}^{n}C_{y} \cdot p^{y} (1-p)^{n-y}$$

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

where

p - desired outcome
$1-p$ = alternate outcome
n = no. of trials.

ex Tossing a coin 10 times
$X \sim B(n, p)$

③ Poisson Distribution

It shows how many times an event occurs in a fixed interval of time or space. given that the events happen independently and at a constant rate

ex If a student ask 4 question in every lecture. but today he/she asked 7 question in lecture then what is probability that he/she can ask 7 question

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

where

λ = avg. no. of events (mean)
x = number of events
e = euler's number ≈ 2.718

Continuous Distribution (infinite no)

A continuous distribution describes the probabilities of outcomes for a continuous random variable

Ex Height of people, temp. over a day etc.

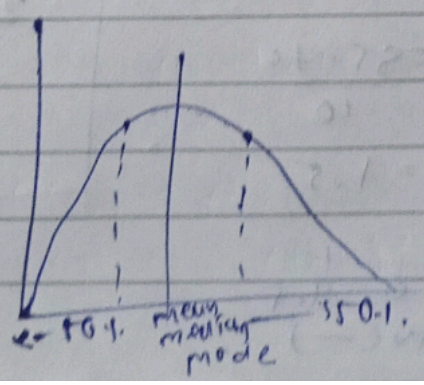
Types
1. normal Distribution
2. student's - T Distribution
3. chi - square Distribution

1. Normal Distribution

A normal distribution is also called Gaussian distribution is "bell shaped and symmetrical probability distribution where most data points are clustered around the mean

$$X \sim N(\underset{\text{mean}}{\mu}, \underset{\text{variance}}{\sigma^2})$$

Properties / Feature
1. Graph - symmetry
2. majority of data - near mean
3. Less amt of data - far from mean
4. outlier - less



Date : ..... Page No. : .....

Real life data follows normal Dist.

mean ± std - 68% (data lies within 1 std for mean)
mean ± 2 std = 95% (")
mean ± 3 std = 99.7% (")
0.03% data are outliers.

* Z-Score
A z-score tells you how many std (σ) a data point is from the mean of the dataset.

z-score is special kind of data

$$z = \frac{x - \mu}{\sigma}$$

x = value in dataset
μ = mean
σ = std

z-score is used when $n > 30$

* standard normal Distribution
It is used to standardize data point so you can compare it with others

mean = 0
+ve = above avg.
-ve = below avg.

$$z = \frac{x - \mu}{\sigma}$$

Eg. μ = 70
σ = 10
x data point

$$z = \frac{85 - 70}{10} = 1.5$$

0 1.5
+ve →

$y \sim N(\mu, \sigma^2)$ - ND
$z \sim N(0, 1)$ - SND

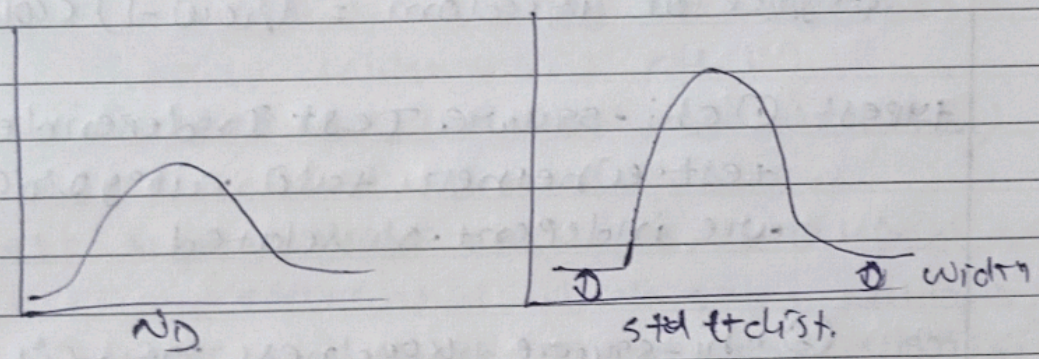
where $z = \frac{y - \mu}{\sigma}$

Date : Page No. :

(2) Student's t-distribution

the student's t-distribution test is used to compare the means of two groups to determine if they are significantly different from each other.

- It is bell-shaped distribution similar to Normal distribution but wider or flatter.



when to use?

(1) sample size is small ($n < 30$)

(2) Population standard deviation is unknown

(3) chi-square distribution

It is non-parametric statistical test used to determine wheter there is a significant relation between categorical variables or whether observed frequencies differ from expected frequencies

- Is there a significant difference between what we expect and what we observe?
- Only used in hypothesis testing

when to use? - (1) The data is categorical

(2) test of independence: Are two categorical variables related?

Date : ..... Page No. : .....

③ Comparing observed frequecies and expected frequencies.

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

where O - Observed frequency
E - Expected frequency

degree of freedom = (row - 1) (col - 1)

types: ① chi-square Test Independence
test whether two categorical variables are indepent. or related

② chi-square goodness of fit test
test wheter single categorical variable matches an expected distribution.

why to use: ① to analyze categorical data
(e.g) - yes/No, gender, color, etc)

② to explore relationship between categorical variables in table.

③ to test if diff. betn expected and observed data are statistical.

Hypothesis testing.

Hypothesis testing is a process where you try to gather the evidences to reject the null hypothesis and to prove that alternate hypothesis is true.

- Hypothesis is a kind of statement

there is two types of Hypothesis
(1) Null Hypothesis (H_0)
(2) Alternate Hypothesis (H_1 / H_a)

(1) Null Hypothesis (H_0) - It is a kind of statement which says nothing new is happing There is no significant relationship between two variables

(2) Alternate Hypothesis (H_0 / H_a) - It stats that there is a relationship between two variables

confidence Interval: It is a range of value that's used to estimate an unknown population parameter (like mean, or proportion with a certain level of ~~sign~~ confidence

confidence level: It is the percentage of times you expect a result to fall within a certain range if you repeated the experiment many times

Date : ..... Page No. : .....

Common confidence levels - 90%, 95%, 99%.
99% - Precision less, large range (type-2)

95% - moderate - mostly used

90% - Precision more, range less
(confidence less) (type-1)

TYPE of Error

TYPE-1 : It is a type of error where you are rejecting Ho when it is actually true.

TYPE-2 : It is a type of error where you accept Ho when it is actually false

Loss of significance : - It is probability but we are taking the risk of rejecting Ho when it is actually true. (type-1 error commonly 0.05)

P-value - Probability of getting test results at least as extreme as the observed ones

Test-statistic - A value used to decide whether to reject Ho

Date : ..... Page No. : .....

Conclusion - Reject H_0 or fail to reject H_0 based on P-value vs α.

Steps of Hypothesis testing:

1. State the Hypothesis - H_0, H_1
2. Set significance level (α) - 0.05
3. Choose the test - (Z-test, t-test, chi-square test etc)
4. calculate test statistic and P-value
5. make Decision
   If P-value $\leq \alpha$ -> reject H_0
   If P-value $> \alpha$ -> fail to reject H_0

Sampling techniques.

Sampling techniques are methods used to select a subset (sample) from a larger group (population) for analysis

Types : 1. simple random sampling
2. stratified sampling
3. systematic sampling
4. convenience sampling

1. Simple random sampling : Every item has an equal chance of being selected

2. Stratified sampling: Divide the population into strata (group) and randomly sample from each group

③ Systematic sampling: select every n^{th} element from the list after a random starting point ~~etc.~~

Eg (every 10th person in queue)

④ Convenience sampling: select individuals who are easiest to reach

Eg (surveying people at the mall)

Z-test

A Z-test is a statistical method used to determine whether there is a significant difference between sample and population means when the population variance is known and the sample size is large ($n > 30$)

- Z-test is used when the sample size is large and the population standard deviation is known

formula - Population: $\boxed{Z = \frac{\bar{x} - \mu}{\sigma}}$ Sample: $\boxed{Z = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}}}$

when to use Z-test?

① Data is normally distributed
② population standard deviation (σ) is known
③ the sample size is larger ($n \geq 30$)

Types:

① one - sample z-test
comparing a sample mean to a known population mean

$$\boxed{z = \frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}}}$$

where $\bar{x}$ = sample mean
μ = population mean
σ = population standard deviation
n = sample size

② two sample z-test
comparing the means of two indepent samples

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

where
$\bar{x}_1, \bar{x}_2$ - sample means
σ_1, σ_2 - population std
n_1, n_2 - sample sizes

③ z-test for proportion
comparing sample propotion with population proportion

Date : ..... Page No. : .....

$$z = \frac{p - P}{\sqrt{\frac{P(1-P)}{n}}}$$

p = sample proportion
P = population proportion
n = sample size

T-test

A t-test is a statistical method used to determine whether there is a significant difference between means of two groups [when population std is unknown and/or sample size is small (n<30)].

- T test is used when population std is unknown and sample size is small (n<30)

when do we use T-test?

1. you're comparing means (either between one sample and a population or between two groups)
2. The population std is unknown
3. The data is approximately normally distributed (for small sample)

Types ① one-sample T-test

compare sample mean to population mean

Date : ..... Page No. : .....

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

degree of freedom = $\boxed{(n-1)}$

(2) two - independent sample T-Test
Compares the mean of two independent groups

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

degree of freedom $$\frac{\left[\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right]^2}{\frac{\left[\frac{s_1^2}{n_1}\right]^2}{n-1} + \frac{\left[\frac{s_2^2}{n_2}\right]^2}{n-2}}$$

(3) Paired t-test
compares the means of two-related groups or measurements
ex before and after treatment

formula $$\frac{\bar{x}d}{sd/\sqrt{n}}$$

$\bar{x}d$ - mean difference
sd - std difference