Statistics

What is statistics ?

⇒ It is a branch of Mathmatics which helps to collection, organige, analyse the data.

Why use statistics ?

1. Summarigation ⇒ which is mean, medium, mode, Count.

2. Test assumption ⇒

3. Visualegation ⇒

⇒ To find what really is going on.
⇒ To spot the pattern & Problem.
⇒ To smart decision using data.
⇒ To simplify & understand large data.
⇒ For making futurce predict^{n} or estimates,

When to use statistics ?

⇒ When we need to understand large data Sets.
⇒ when we want to analyze the results of a Survey or experiment.
⇒ when we need to make decisions based on trends or patterns.

Where to use statistics?

1. Education ⇒ Analysing student result, school Performance
2. Business ⇒ Market analysis, customer behaviour, Profit/loss analysis,
3. ~~Economic planing~~
3. government ⇒ economic planing.
4. Medical field ⇒ Analysing drug effectiveness, patient recovery date.
5. Sports ⇒ Player performance analysis, team statistics.

How to use statistics?

1. Data collection ⇒ using surveys, experiments.
2. Data organization ⇒ using table, charts, grouping.
3. Data Analysis ⇒ using averages, %, Standard deviation, correlation.
4. Data interpretation ⇒ Understanding the meaning of the results.
5. Condusion/Desision Making ⇒ Taking actions based on the results.

Types of statistics

① Descriptive

② Inferential.

① Descriptive Statistics :- ⇒ Summary

⇒ They are some method for summerizing & understanding the key aspect of the data.

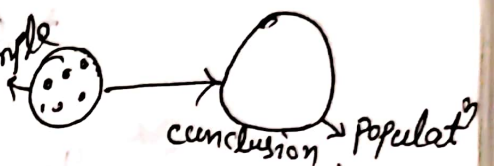
⇒ Summarizing data (Mean, Median, mode, graphs, tables etc.)

⇒ data ke baremain kuch information mile.

② Inferential Statistics :- ⇒ conclusions.

⇒ Drawing conclusions about ~~the future~~ or a largere population based on data.

⇒ Using data we can make some conclusions using some techniques



Types of Variable

① Quantitative
Numerical
- Discrete int
- continuous float.

② Qualitative
Categorical
- Nominal
- Ordinal.

① Qualintitative Data ⇒ (Numerical data)

⇒ This type of data consists of numbers and can be measured.

⇒ Use graph : line, histogram.

Discrete (whole no., int) ⇒ countable data, usually whole numbers.

Ex No. of student, No of bank ac, no. of cars.

Continous (fractional, decimal) ⇒ Mesurable data that can take on any value within a range.

Ex Height → 79.1, weight → 50.7, temp., time.

② Qualitative data ⇒ (categorical data) → labels.

⇒ This type of data describes non-numeric information.

⇒ graph → Bar, Pie.

Nominal ⇒ Data with categories that have no order.

⇒ Order does not matter.

Ex ⇒ gender, colors, religion, blood type, subject ...

Ordinal ⇒ Data with categories that have a meaningfull order, but the difference between values are not measurable

⇒ orders matters.

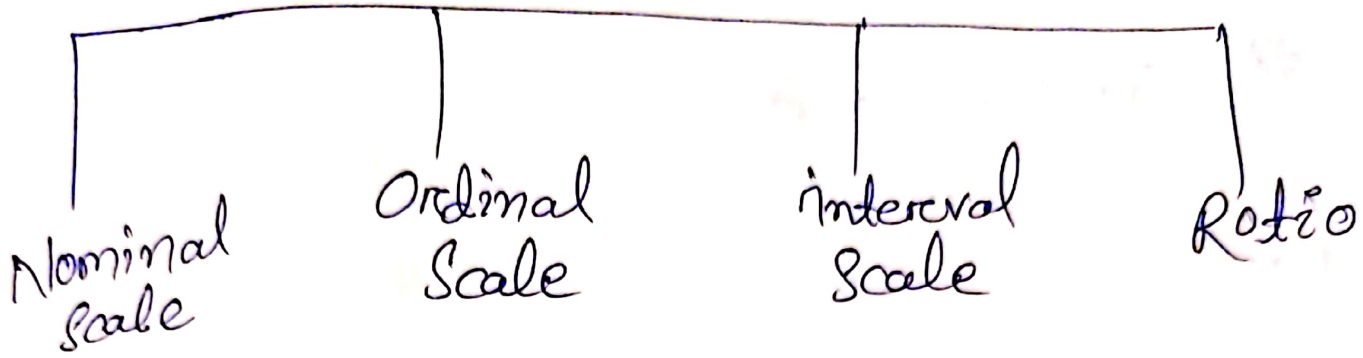
Ex →
- feedback → Good, bad, Avg
- size → small, med., large
- Educat^{n} → High school, Bachelors, masters

Scales of Measurements

⇒ It defines how the data is classified compared & used mathmatically.

Properties
- → Identity (label)
- → order (ranking)
- → Equal interval
- → True zero

Scales of Measurement



① Nominal Scale :-

⇒ Naming data

⇒ Label on category data with no or

Ex gender, Region, country

② Ordinal Scale

⇒ Lable on category with orde

Ex Customer feedback
Educatn level.

③ Interval Scale

⇒ Numeric scale with equal intervals, between values.

⇒ No true zero value.

⇒ No true zero matlab addition, substraction kar sakte hai

Ex ⇒ Temp, calender. (*, / nahi kar sakte hai)

⇒ Zero doesnot mean absence

⇒ Mean, median, gender deviation find kar sakte ho.

⇒ Histogram, Boxplot depends on data.
Co-relation find kar sakte hai ho, 2 numerical ke bich main se find kar sakte hai,

④ Ratio

⇒ Equal interval & true zero value.

⇒ yahape Zero ka matlab zero hi hai, zero ka matlab nothing.

Ex ⇒ Height, weight, money

⇒ All mathematical operation can be perform.

⇒ yaha pe line chat, box plot, histogram sab bana sakte hai ho.

⇒ co-relation bhi find kar sakte ho.

Population & Sample

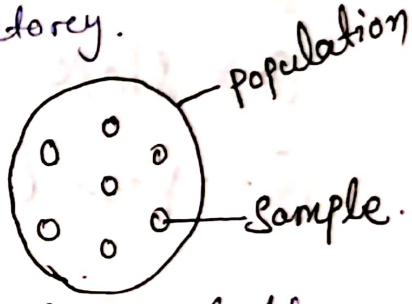
Population ⇒ whole data set.

⇒ A populat^n is the entire group of individual items or data points that you study & draw conclusions from,

⇒ It is the complete set of data.

OR ⇒ • All students in a university.
• All citizens of a country.
• Every Product from a factory.

⇒ Populat^n represent (N)



Sample

⇒ A sample is a subset of population, selected for the actual study or analysis,

⇒ sample represent (n)

What is data

⇒ Facts or pieces of information than can be measured.

Sampling Techniques

① Simple Random Sampling.
② Stratified Sampling.
③ Systematic sampling.
④ Convenience Sampling.

Descriptive Statistics

⇒ Descriptive statistics are the methods for summarizing & understanding key aspects of data.

⇒ key aspects means the most impotant parts, features, or characteristics of something.

Types of DS

① ⇒ Central Tendency → Mean, Median, mode

② ⇒ Dispersion → range, variance, SD

③ ⇒ Shape → Skewness, Kurtosis

④ ⇒ Visual → diff of charts.

*What is descriptive statistics? ⇒ Summarization

⇒ DS is the branch of statistics that focuses on summarizing & organizing data so it can be easily understood.

*Why used? ⇒ find pattern/info

⇒ Summarize large datasets in a simple, meaningful way.

⇒ Identify patterns or trends

⇒ Make data easier to interpret.

⇒ Present data visually for reports or presentations.

* When is DS used ? ⇒ before decision Making.

⇒ Before analysis to explore and understand the dataset

⇒ In reports and presentations to summarize result.

⇒ In research to describe characteristics of a Populat^{n} or sample.

⇒ In business to track performance (e.g, avg sales, customer age)

* Where is DS used ? ⇒ every where.

⇒ In education (e.g, average student scores)

⇒ In business (e.g, monthly sales performance)

⇒ In healthcare (e.g, patient age distribut^{n})

① Measure of central Tendency

1. # Measure of central Tendency

⇒ Measure of central tendency describes the "center" of the data.

⇒ They provide summary of the data with a single value that represents the data as a whole.

① **Mean** ⇒ average

⇒ Refers to the center of distribution.

Used when:

Data is numeric and evenly distributed.

Population

Formula:

$$M = \frac{\sum_{i=1}^{N} x_i}{N}$$

M = mu

N = Total no. data points in Pop.

$\sum$ = Summation

x_i = particular data points

i = row no.

Sample

Formula:

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

n = Total no. of data points in sample.

Ex

(1, 1, 2, 2, 3, 3, 4, 5, 5, 6)

$\bar{x} = (1+1+2+2+3+3+4+5+5+6) \div 5$

$= 9\frac{32}{10} = 3.2$

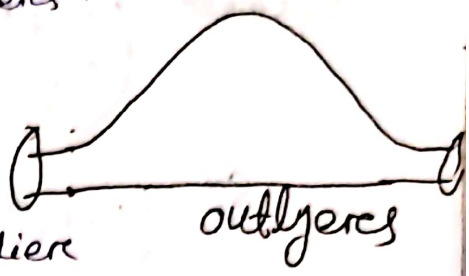
Outliers

⇒ a data point(s) who doesn't follow the pattern of data / or are extreme points.

⇒ Mean is sensitive of outliers.

Ex

{1, 1, 2, 2, 3, 3, 4, 5, 5, 6, 100}
↳ outlier



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

② Median ⇒ Middle value.

⇒ The median is a measure of central tendency the represents the middle value in a sorted list of numbers.

Even:

If the number of values is even, the median is the average of the two middle numbers.

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

Ex {1, 1, 2, 2, 3, 3, 4, 5, 5, 6}

$$\frac{10^{th}}{2} + \left(\frac{10}{2}+1\right)^{th} = \frac{5^{th} + 6^{th}}{2}$$

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

Ex
List: 10, 2, 5, 8
Sorted: 2, 5, 8, 10
Median = $\frac{5+8}{2}$
= 6.5

Odd:

⇒ If the number of values is odd, the median is the middle number.

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

Ex = {1, 1, 2, 2, 3, (3), 4, 5, 5, 6, 100}

$\left(\frac{11+1}{2}\right)^{th} = 6^{th} = 3$

Ex
List = 3, 1, 7
Sorted = 1, 3, 7
Median = 3

③ **Mode** ⇒ Most frequent

⇒ The mode is the number that appears most frequently in a set of values.

Ex Data set:

3, 7, 3, 2, 9, 3, 5

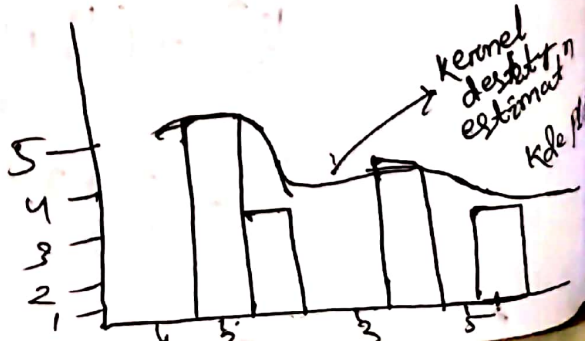
Here, 3 appears three times more than any other number.
So, the mode is 3.

Why use Mode?

⇒ It's helpful in understanding the most common value.

* [1, 1, 2, 3, 5, 1, 1, 3, 1, 3, 5, 1]

1 = 6 3 = 3
2 = 1 5 = 2



2 # Measure of Dispersion

⇒ Measure of dispersion are statistical tools used to describe the spread or variability of a data set.

⇒ They show how much the data values differ from the average or mean.

• ⟶

① Range

⇒ Difference between the highest and lowest values.

Formula:

Range = Maximum − minimum.

⇒ Simple but sensitive to outliers.

Ex

data: 5, 7, 8, 9, 100

Range = 100 − 5 = 95

yahan 100 ek outlier hai (baki sab value choti hai), aur iski wajah se range bahut bada ho gaya isliye range ek simple but outlier-sensitive measure of dispersion hai.

⇒ isliye hum data ko sahi se samajh ne keliye variance ka use karte hai.

② Variance

⇒ It measures how far each data point is from the mean.

High variance → Spread ←→ (More data are far from mean)

Low variance → Spread ←→ (More data around mean)

Population

$$\sigma^2 = \sum_{i=1}^{N} \frac{(x_i - \mu)^2}{N}$$

Sample

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

③ Standard deviation ⇒ (unit same easily comparable)

⇒ Standard deviation is measure of how spread out the values in a data set around the mean (avg).

⇒ It tells you how much individual data points differ from the mean value.

key points:

⇒ A low standard deviation means data points are close to the mean.

⇒ A high standard deviation means data points are spread out over a wider range.

3. # IQR → Inter Quartile Range

Percentile

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

$$\text{percentile} = \frac{I(\alpha) - 1}{n} \times 100$$

($I(\alpha)$ → Index of (α); n → Total observation)

$$\frac{\text{percentile}}{100} \times n = I(\alpha) - 1$$

$$\left(\frac{\text{percentil} \times n}{100}\right) + 1 = I(\alpha)$$

Percentile of rank?

10 → value

$$\frac{\text{no. of values below } x}{n} \times 100$$

$$= \frac{16}{80} \times 100^{5} = 80$$

Five Number Summary

1. Minimum ⟶ Q_0
2. 25% ⟶ First Quartile (Q_1)
3. Midian ⟶ 50 percentile Q_2
4. 75%. ⟶ Third Quartile (Q_3)
5. Maximum ⟶ Q_4

Dataset = 2, 2, 3, 4, 5, 5, 5, 6, 7, 8, 8, 8, 8, 8, 9, 9, 10, 11, 11, 12

↓ (ascending)

$n = 20$

$Q_1 \rightarrow$ 25% tile $= \left(\frac{25}{100} \times 20\right) + 1 = 6$th index

→ 5

$Q_3 \rightarrow$ 75% tile $= \left(\frac{75}{100} \times 20\right) + 1 = 15 + 1$

$= 16$th index

→ 9

IQR = Inter quartile range

$= Q_3 - Q_1 = 9 - 5 = 4$

Min = Lower fence $\Rightarrow Q - 1.5(IQR)$

$= 5 - 1.5(4)$

$= 5 - 6 = -1$

Max = Higher fence $= Q_3 + 1.5(IQR)$

$= 9 + 1.5(4) = 9 + 6$

$= 15$

$x < -1 \rightarrow$ outlier

$x > 15 \rightarrow$ Outlier

Box plot

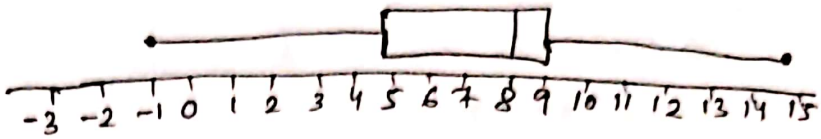
Min = -1

25% = 5

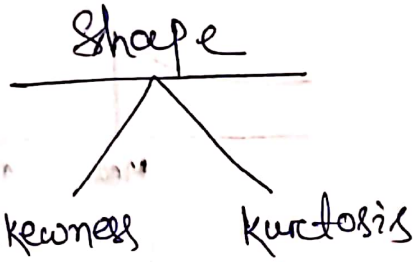
Median = 8

75% = 9

Max = 15



Shape



Skewness

~~Tail is more towards right or~~

⇒ It measures asymmetry of a distribution around mean.

* Symmetric means → value → around mean → evenly spread.

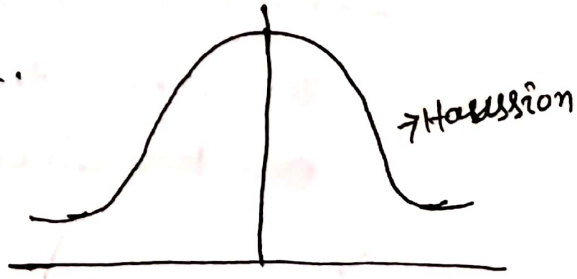
Why?

⇒ Understand the shape of data.

Ex

= • Imagine people earn between 20K – 50K.

• A few people earn 2 lakhs +



Mean = median = Mode

Symetric

Types of Skew

① **+ve Skew**

⇒ Tail is more towards right or +ve

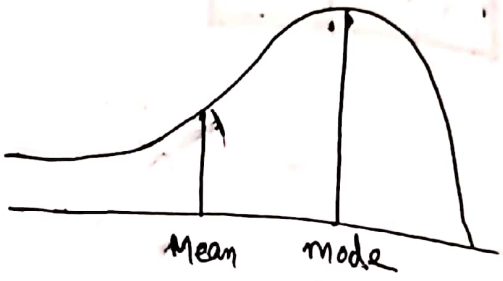
⇒ Mean > median > Mode

⇒ Skewness > 0

② **−ve Skew**

⇒ Tail on the left side longer.

⇒ Mean < median < Mode



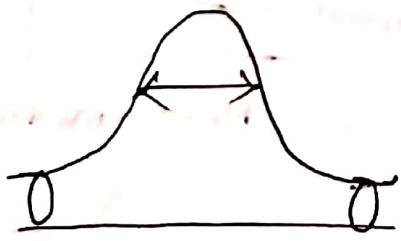
$$\text{Skewness} = \frac{n}{(n-1)(n-2)} \Sigma \left[\frac{(x_i - \bar{x})^3}{s} \right]$$

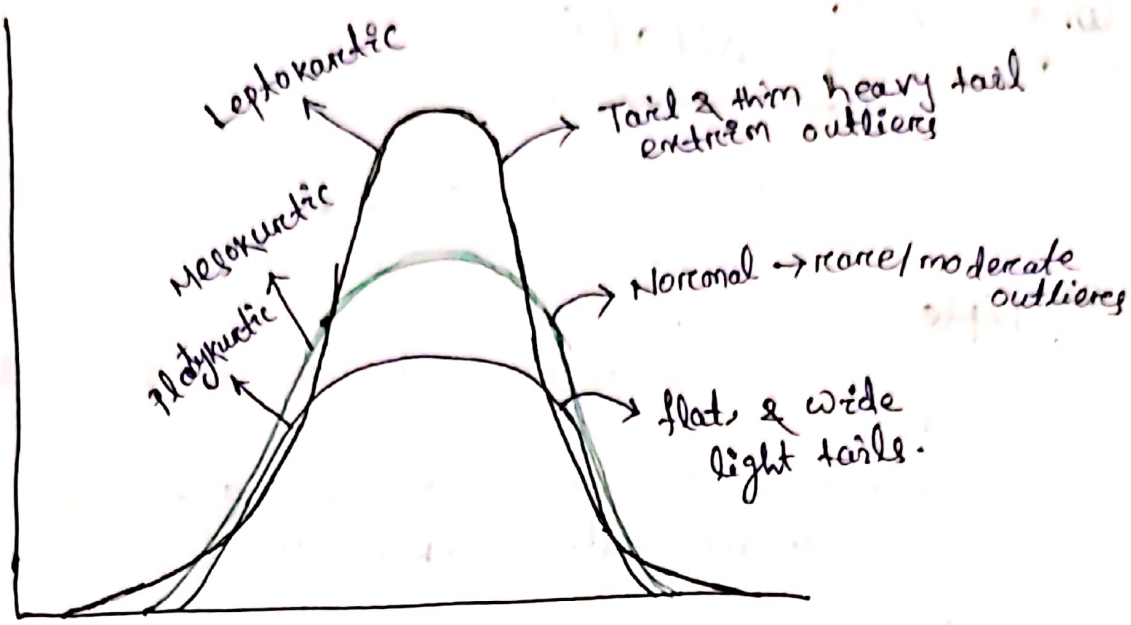
x_i – data pt
$\bar{x}$ – mean
s – std
n – no of observat^n.

∴ Kurtosis

⇒ Kurtosis is a statistical measure that describes the "tailedness" or "peakedness" ~~or peakedness~~ of a probability distribution.

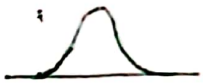
⇒





Platykurtic

1. Mesokurtic
   - ⇒ Distribution with normal kurtosis;
   - ⇒ Moderate tails, similar to the normal distribut^{n}.



2. Leptokurtic
   - ⇒ Distribution with high kurtosis (kurtosis > 0)
   - ⇒ Heavy tails and a sharp peak
   - ⇒ most extreme outliers.
   - ⇒ Higher probability of extreme values.

3. Playty kurtic
   - ⇒ Distribution with low kurtosis (kurtosis < 0)
   - ⇒ Light tails & a flatter peak – fewer outliers.

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

$K > 0 \rightarrow$ lepto
$K < 0 \rightarrow$ Platy
$K = 0 \rightarrow$ Meso.

Why?

⇒ To detect outliers.

⇒ To understand risk in data.

Ex Lepto

① Stock returns → more chance of big profit or loss.

② Most students score 50 but a few get 0 or 100 → extreme value → high kurtosis.

Platy

① fewer extreme answers.

② Scores are more evenly spread between 40-60 low kurtosis.

4. # Visual

⇒ Statistics Visuals are graphical represent-at°s of data used to summarize, interpret and communicate/understand information clearly and effectively.

⇒ Jisse hume patterns, outliers, distribut° pata chalta hai,

① Histogram

② Bar chart

③ Pie chart

④ Box plot

⑤ Line chart

⑥ Scatter plot.

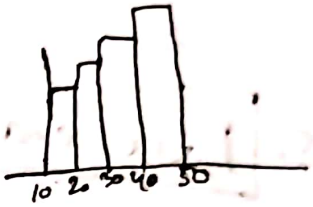
① Histogram ⇒ Univariate

⇒ Histogram is the graphical representation of numeric variable / quantitative data.

⇒ Use for continuous data & groups data into range.

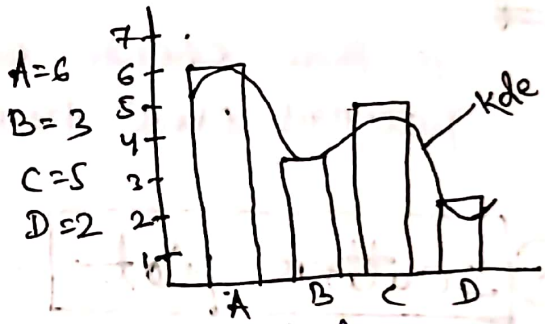
Why?

⇒ To understand how data is spread out, patterns like skewness, central tendency or outliers.



② Bar chart ⇒ Bivariate

⇒ Bar chart represents a categorical data, and the length or height of the bar shows the value or frequency of that category.

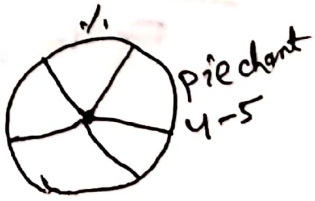


Why?

⇒ Easy to understand, compare different categories or traits.

③ Pie chart

⇒ A pie chart is a circular chart divided into slices, where each slice represents a percentage of a whole data.



Why?

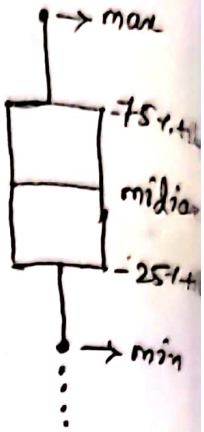
⇒ Want to visualize percentage.

⇒ easy to see which category is biggest or smallest,

4. **Box plot** ⇒ univariate.

⇒ Box plot shows the spread and skewness of numerical data.

⇒ using five-number summary.



Why?

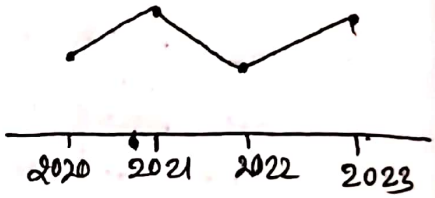
⇒ understand the spread and center of the data, To outlier detection.

5. **Line chart** ⇒ Bivariate.

⇒ Line chart shows trends over time by connecting data points with a line.

Why?

To show change or trends patterns over time.



6. **Scatter plot** ⇒ Bivariate analysis tools.

⇒ Scatter plot displays the co-relationship between two numerical variables.

Why?

⇒ To detect trends or outliers.

Distribution (Probability Distribution)

⇒ The possible values that a variable can take & how frequently they can occure.

* $y \rightarrow$ actual outcom

$y \rightarrow$ one of the possible outcome, (0-1)

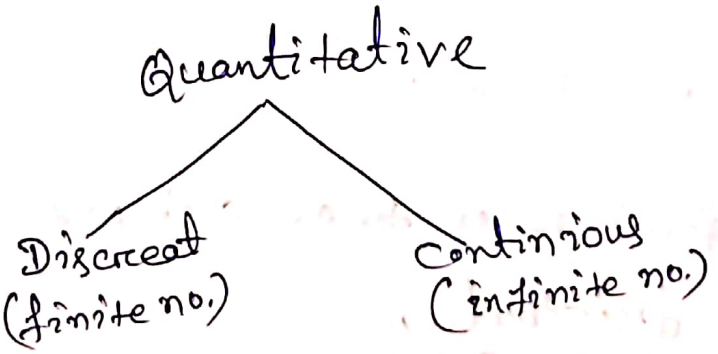
Ex

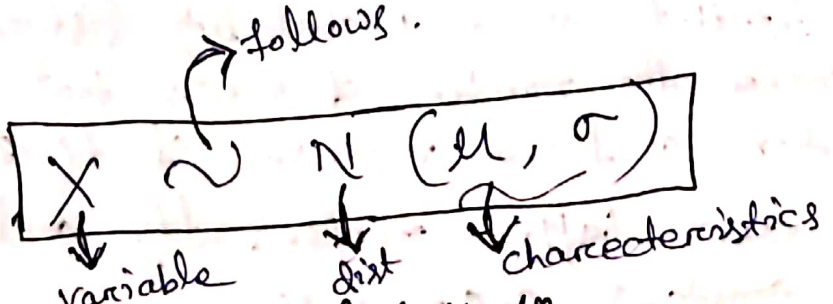
Coin < H – ½ / T – ½

Dice ⚀ ⚁ ⚂ ⚃ ⚄ ⚅ n = 6, 1/6, 1 2 3 4 5 6

* P(∈) = P(y=3) = 1/6 (∈ → event)

*





Why do we need distribut^{n}

⇒ where it +

Discrete distribution ⇒ Values that can be counted (not measured)

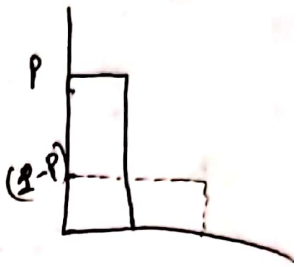
1. Bernoulli distribution

⇒ It models a single experiment with two possible outcomes.

⇒ 1 trial & outcomes.

⇒ Success ($x=1$), failure ($x=0$)

$X \sim Bern(P)$



* P = Probability of success

1-P = probability of failure [Binary outcome 0 or 1]

2. Binomial distribution

⇒ It is a sequence of identical Bernoulli event.

$$x \sim B(n, P)$$

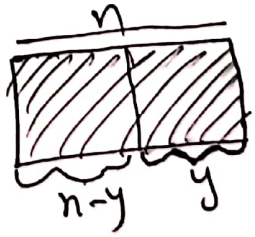
⇒ fixed no of trial each with Bernoulli result.

Probability concept:

- P (desired outcome) = p
- P (alternate outcome) = 1-p

$y \rightarrow P(y) = p$

$P(n-y) = 1-p$



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

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

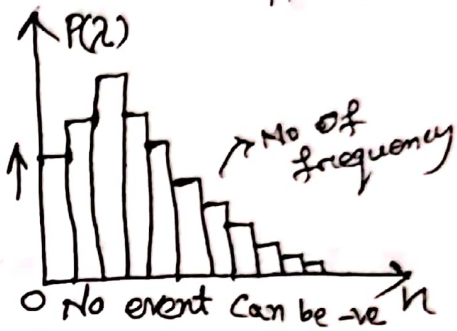
3. Poisson distribution

⇒ Poisson distribution is a probability distribution that describes the number of events that occure within a fixed interval of time.

⇒ Count of events in fixed time, event happen indipendently, rarely.

$$X \sim Po(\lambda)$$

(λ → interval)



Ex: • each day 4 question ans.. asked, 1 day 7 question asked.

$$P(y=7) = \frac{\lambda^y e^{-\lambda}}{y!} \qquad e = \text{Euler's no.}$$

$$= \frac{4^7 \times e^{-4}}{7!} = \frac{4^7 \times 2.72^{-4}}{7!}$$

Contineous distribution

⇒ It show how different values occur in data set & describes the over all pattern of the data.

1. Normal Distribution (Gaussian Distribution)

⇒ A normal distribution is a continuous probability distribution, that is symmetric around its mean, forming a bell-shaped curve.

It is one of the most important distributions in statistics because many natural and social phenomena follow it.

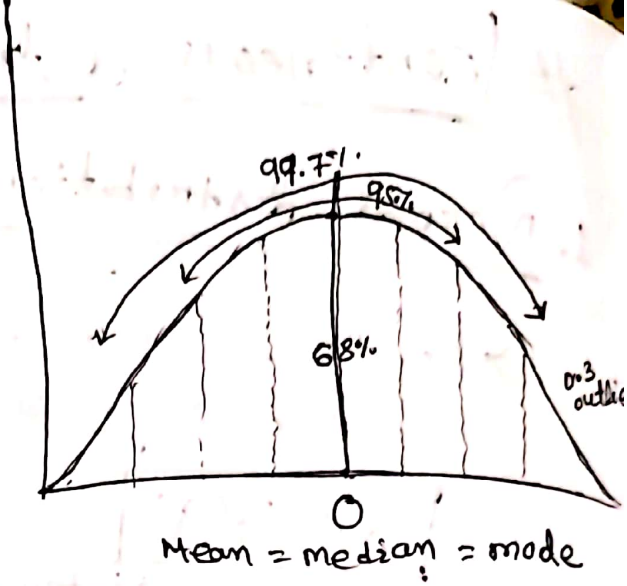
Empirical rule

$\mu \pm 1\sigma \longrightarrow 68\%$

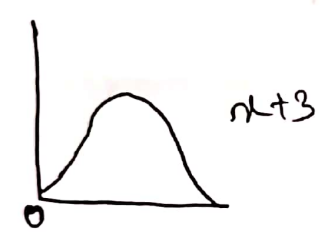
$\mu \pm 2\sigma \longrightarrow 95\%$

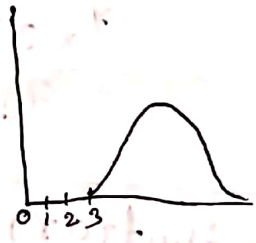
$\mu \pm 3\sigma \longrightarrow 99.7\%$

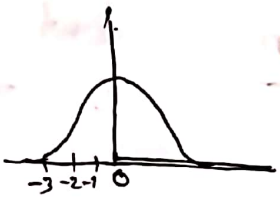
$0.03 \longrightarrow$ outliers

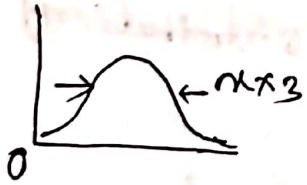


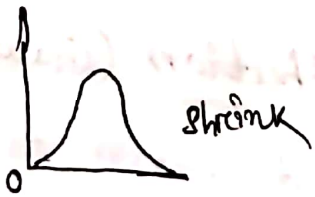
* Distribut^{n} ke upar transformat^{n} bhi kar sakte hain; value add ya subtract kare. to distribut^{n} shift hoga.

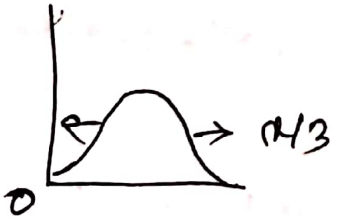


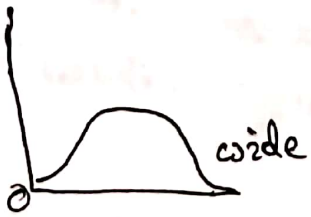












Special kind of transformation jisko hum Z-score kehte hain.

*

Standardized ⇒ Compaisce data to other / data point + Other.

$$Z\ score = \frac{x - \mu}{\sigma}$$ (x → data point) ⇒ Mean(μ) = 0, SD(σ) = 1

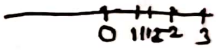
* Used to standardize a data point so you can compaisce it with other.

Mean = 0
+ve value = above avg
-ve value = below avg.

Ex
μ = 70
σ = 10
SND = 85
n > 30

$Z = \frac{85 - 70}{10} = \frac{15}{10} = 1.5$

M = 0 & SD = 1



2. Student t-distribution

what ⇒ Bell shape distribut" similare to ND, But wider & flater.

⇒ use small sample

Sample SD → known

Pop SD → unknown

Why

⇒ Sample jab small hota hai, use hum inforence leke ana hai.

⇒ n < 30 ho tab hum isko use karetain hai.

Inforence means darewing a conclusion based on evidence.

3. Chi-Square distribut^n $(\chi)^2$

⇒ The Chi-Square distribution is the distribut^n of a sum of the squares.

⇒ It is used in hypothesis testing, confidence interval estimat^n for categorical data.

Use · Chi-Square Test of independence (eg. gender Vs voting preference)

• The sketch shows a right skewed curve.

$$df = (r-1) \times (c-1)$$

r = no. of row
c = no. of column

Hypothesis testing

⇒ Hypothesis Testing is a statistical method used to make decisions or inferences about a Popula^n based on sample data.

⇒ It is a process. where you try to gather the evidence to reject the null hypothesis and to prove that alternate hypothesis is true (accept)

- Hypothesis → kind of a statement
- Two type ← null Hypothesis
  alternate hypothesis.

Null hypothesis (H_0)

⇒ It is a kind of statement which says nothing new is happening.

⇒ There is no significant relationship between two variable.

Alternate hypothesis (H_1 / H_a)

⇒ There is significant relationship between two variables.

- 50 - point estimate ⇒ avg of all matches (sample)
- 50 +-5 [+-5 → margin of error]
- [45, 55] → confidence interval.

confidence level ⇒ $(1-\alpha)$

α = significance level

= 0.05 = 5%

$\alpha = 0.05$

confidence level $1-\alpha$

$= 1 - 0.05 = 0.95 = 95\%$

- confidence level ⇒ 90% = 0.9 = [0.9, 0.95, 0.99]

[40 - 60] = 0.99 ⇒ Precision less, large range [type - 2]

[45 - 55] = 0.95 ⇒ Moderate = mostly used

[48 - 52] = 0.90 ⇒ Precision more, range less ⇒ confidence less [type 1]

- 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 are accepting Ho when it is actually false.

- Loss of significance ⇒
  It is the probability that we are taking the risk to reject Ho when it is actually true, (type 1), [1 - confidence level], [0.05]

Sampling Techniques

Types of Sampling

① Simple Random Sampling :-
- Every member has an equal chance of selection.
- Example :- Drawing names from a hat.

② Statified Sampling :-
- population is divided into subgroups and random samples are taken from each subgroups.

Example :- Age group
10 - 20
21 - 30 } male/female.
31 - 40

③ Systematic Sampling :-
- every Kth (like 3rd, 5th, 10th) member is selected from a list.

Example :- choose 7th, 14th, 21th, 28th person in parttern queue.
- Selecting "nth" number follows pattern

Z-TEST

A Z-test is a statistical test used to determine whether there is a significant difference between sample and population means (or between two sample means),

When the population variance is known and the sample size is large. (typically $n \geq 30$)

Formula:-

pop

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

Sample

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

When?

- The sample size is large ($n \geq 30$)
- The populat^n standard deviation (σ) is known.
- Random sampling [normalized]
- The data is approximately normally distributed (especially important for small samples)

* μ = population mean
  σ = populat^n Std dev.
  $\bar{x}$ = sample mean
  n = sample size.

P-value	Kya karna hai	Matlab
$P \leq 0.05$	Ho ko reject karo	Ha ko support karo
$P > 0.05$	Ho ko reject nahi kar sakta	Ha ko support nahi milta

T-TEST

T-test is a statistical test used to determine whether there is a significant difference between the means of two groups.

It commonly used when:

- The sample size is small ($n < 30$)
- The population is unknown, ~~(~~ Sd = unknown

① One Sample t-test

Compares the sample mean to a known value (often a population mean).

Ex: Checking if a class's average score differs from the national average.

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

df = degree of freedom

$$df = n - 1$$

② Two Independent Sample t-test.

Compaires the means of two independent groups.

Ex: Comparing test scores of two different classes

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

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\left(\frac{s_1}{\sqrt{n_1}} - \frac{s_2}{\sqrt{n_2}}\right)}$$

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

③ Paired t-test (dependent)

Compairces the oneans of two ocelated groups.

Eg: Pre-test and post-test scores for the same students.

Formula:

$$t = \frac{\bar{d}}{Sd/\sqrt{n}}$$

$\bar{d}$ → mean of the differences

Sd → Standard deviat^n of the differen

Chi-Square-Test

The chi-square test (χ^2 test) is a statistical method used to test ocelat^n between categorical variables.

It helps answer questions like:

- Is this distribut^n what we expected?
- Are two categorical variable ocelated ore independent?

Why?

To compare observed data (what u collected) with expected data (what u theoretically expect)

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

O = Observed frequency
E = Expected frequency

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

$$df = (r-1) \times (c-1)$$

r = orow total
C = column total

Example :- Educat^n background & Job satisfactory

Education	Satisfied	Not Satisfied	Total
High school	50	70	120
Graduation	90	60	150
Post gradcat^n	20	10	30
	160	140	300

Of = observed frequency

Educat^n	Satisfied	Not satisfied.
High School	$\frac{120 \times 160}{300} = 64$	$\frac{120 \times 140}{300} = 56$
Graduation	$\frac{150 \times 160}{300} = 80$	$\frac{150 \times 140}{300} = 70$
Post graduation	$\frac{160 \times 30}{300} = 16$	$\frac{140 \times 30}{300} = 14$

ER = Expected frequency or Result

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

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

$$G/S = \frac{100}{80} = \boxed{1.25}$$

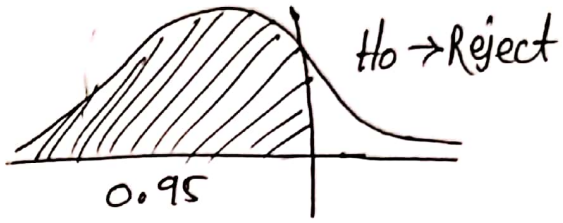
$$G/NS = \frac{100}{70} = \boxed{1.43}$$

$$PG/S = \frac{16}{16} = \boxed{1}$$

$$PG/NS = \frac{16}{14} = \boxed{1.14}$$

$x^2 = 3.06 + 3.5 + 1 + 1.14 + 1.25 + 1.43$

$\boxed{x^2 = 11.38}$



$df = (r-1) \times (c-1)$

$= (3-1) \times (2-1)$

$= 2 \times 1 = 2$

$\boxed{df = 2}$