Statistics

Date __/__/____

What?:- It is a branch of Mathematics which helps to collect, organize, analyze the data.

Why:- Use Cases:-

1. Summarization
2. Test Assumption
3. Visualization

→ To find what really is going on
→ To spot the pattern & Problem
→ To smart decision using data

⇒ **Types of Stats:**

1. **Descriptive Stats:-** To describe and summarize the main features of a dataset, such as central tendency, dispersion and shape.

→ Measures of central Tendency:- Mean, Median, Mode
→ Measure of Dispersion:- Standard deviation, variance.

Date __/__/____

→ Frequency Distributions:- Organizing data into intervals and counting the occurrences within each interval.

→ Data visualization:- Charts, graphs, and tables to represent data visually.

2. Inferential Stats:- To use sample data to make inferences about a larger population or to test hypotheses.

→ Hypothesis Testing:- Using statistical tests to determine the likelihood of an outcome based on sample data.

→ Regression Analysis:- Examining the relationship between variables to predict one variable from another.

→ Confidence Intervals:- Estimating the range within which a population parameter is likely to fall.

Date __/__/____

Types of Data:-

1. Qualitative (Categorical) Data:

It is used to describe qualities or attributes of something. It's often used to group or label data into categories.

- Nominal Data: Categories have no inherent order or ranking (E.g. colors, gender)

- Ordinal Data:- Categories have meaningful order or ranking (E.g. Student grades (A, B, C), satisfaction levels (very satisfied, satisfied, neutral, dissatisfied, very dissatisfied)).

2. Quantitative (Numerical) Data:-

It is used to measure or count something. It provides numerical values that can be manipulated mathematically.

Date __/__/____

→ Discrete Data:- Values can be counted and have distinct, separate values (E.g. number of students in a class, number of cars in a parking lot).

→ Continuous Data:- Values can take on any value within a range and can be measured with more precision (E.g. height, weight, temp).

⇒ Further Considerations:-

→ Ratio Data:- A type of quantitative data where there is a true zero point.

Scales of Measurements:- It refers to the ways variables or data are categorized, measured, and interpreted. They define the nature of the information within the values assigned to variables.

⇒ Properties:- Identity (Label), Order (Ranking), Equal Interval, True zero.

Date __/__/____

⇒ There are four main scales of measurement:

1. Nominal Scale (Categorical) :- Pie, Bar Chart

→ Data are grouped into categories that have no order or ranking.

→ Example:- Gender (Male, Female)
Blood Type (A, B, AB, O)
Eye color (Blue, Brown, Green)

→ Key Characteristics:-

1. Labels or names only
2. No numerical or quantitative meaning
3. Cannot perform mathematical operations

2. Ordinal Scale:- Data are grouped into categories that have a meaningful order, but the intervals between values are not equal or known. (Bar Chart)

→ Example:- Class Rank (1st, 2nd, 3rd)
Satisfaction Ratings (Satisfied, Neutral, Dissatisfied)

→ Key Characteristics:- Ordered
Differences between values are not measurable

Date __/__/____

3. Interval Scale:- Numerical data with equal intervals between values, but no true zero point.
(Histogram, Boxplot)

=> Examples: Temperature in Celsius or Fahrenheit
IQ Scores
Calender dates

=> Z-test, T-Test Correlation

=> Key Characteristics:-

1. Order and Equal Spacing
2. Can add and subtract values
3. No true zero (0 doesn't mean "none")

4. Ratio:- Numerical data with equal intervals and a true zero point.

=> Examples:- Height, Weight, Age, Income

=> Key Characteristics:-

1. All mathematical operations are meaningful (Addition, Subtraction, Multiplication, Division)

2. True zero means absence of the quantity

Date __/__/____

⇒ Summary Table of Scale Measurements:-

Scale	Order	Equal Interval	True Zero	Example
Nominal	No	No	No	Gender, Eye Color
Ordinal	Yes	No	No	Rank, Satisfaction
Interval	Yes	Yes	No	Temperature (°C or °F)
Ratio	Yes	Yes	Yes	Height, Age, Income

⇒ Population & Sample :- It refers to the entire group of individuals, items, or data that you're interested in studying.

Examples:- 1) All citizens of a country
2) Every manufactured item from a factory

Date __/__/____

Characteristics: 1.) Often large or infinite
2.) Describe by parameters
(e.g., population mean μ,
standard deviation σ)

⇒ Sample: It's a subset of the population
that is actually observed
or analyzed.

- It's used when studying the whole
population is impractical,
costly, or impossible.

Characteristics: ~~1.) Should be represen~~

1.) Should be representative of the
population
2.) Described by statistics
(e.g., sample mean $\bar{x}$,
sample standard deviation s)

3. Used to make inferences about the population

Example: Population: All adult women in
the country.
Sample: 1000 randomly selected adult women

Date __/__/____

Descriptive Statistics :- It is a branch of statistics that involves organizing, summarizing, and presenting data in an informative way.

⇒ Key functions :-

1. Summarize Data :- Helps reduce large amounts of data into a more understandable form.

2. Identify Patterns :- Reveals trends, central values and variability.

⇒ Types of Descriptive Statistics :-

1. Measure of Central Tendency :-

- Describe the center or average of data.
- Examples: mean (average)
  Median (middle value)
  Mode (most frequent value)

Date __/__/____

2. Measure of Disporsion (Spreed):-

- Shows how spreed out the data is.
- Example:- Range (Diff. betn max and min)
  Variance (avg. squared deviation from the mean)
  Standard Deviation (square root of variance)

3. Measure of Position:-

- Describe the relative standing of data points.
- Examples:- Percentiles, Quartiles

4. Data Visualization:-

- Tools to visually present the data.
- Examples:- Histogram, Pie charts
  Bar charts, Box Plots

Example:- If you have test scores for 100 students.
- Mean score: 75
- Median Score: 78
- Standard Deviation: 8
- Histogram: Shows how scores are distributed

Date __/__/____

What is a Measure of Central Tendency?

- It is a single value that represents the center point or typical value of a dataset.

→ The Three main measures:-

1. Mean (Arithmatic Average)

Formula: Mean = $\frac{\text{sum of all values}}{\text{Number of values}}$

2. Median (Middle value):- The middle number when the data is ordered from smallest to largest.

- If there's an even number of values, take the avg of the two middle numbers.

- Ex. 3, 5, 7 → median = 5
  2, 4, 6, 8 → median = (4+6)/2 = 5

Date __/__/____

3. Mode (Most frequent value):- The value that appears most often in the dataset
- A dataset can have no mode, one mode, or multiple modes.

- Ex. 2, 3, 3, 5, 7 → Mode = 3

→ When to use each:

Measure	Best Used when...
Mean	Data is symmetric and has no extreme outliers.
Median	Data is skewed or contains outliers.
Mode	You need the most common item (e.g. categorical data)

Date __/__/____

Measure of Dispersion :-

- It refers to statistical tools that describe how spread out or varied a set of data values is.

⇒ Common Measures of Disportion:

1.) Range :- The difference between the highest and lowest values in a dataset.

- Formula: Range = Maximum − Minimum

2.) Variance :- The average of the squared differences from the mean.

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

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

Date __/__/____

3) Standard Deviation :- The square root of the variance.
It gives a measure of spread in the same units as the data.

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

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

4.) Interquartile Range (IQR):

- The range of the middle 50% of values.

- Formula : $IQR = Q_3 - Q_1$

Note :- Where Q_3 is the third quartile and Q_1 is the first quartile.

⇒ Why Is Disperion Important?

1. To understand data variability.
2. It indicates consistency or risk.
3. It allows comparison between different datasets.

Date __/__/____

⇒ Inter Quartile Range (IQR):-

- It is the difference between the third quartile (Q3) and the first quartile (Q1).

$$IQR = Q_3 - Q_1$$

- Q_1 (First Quartile): The 25th percentile (The 25% of data is below this value).

- Q_3 (Third Quartile): The 75th percentile (The 75% of data is below this value).

→ Why Use IQR?

1. It is not affected by outliers, unlike the range.

2. It gives a better idea of the spread of the central values in a dataset.

Date __/__/____

⇒ Example: 2, 4, 5, 7, 8, 10, 12, 15, 18

- Q1 (25th percentile) :- 4.5
- Q3 (75th percentile) :- 13.5

- IQR = 13.5 - 4.5 = 9

⇒ IQR & Outliers: Outliers are often defined as values that fall:

Below Q1 - 1.5 x IQR or
Above Q3 + 1.5 x IQR

⇒ Five Number Summary:- It is a quick and informative way to describe a set of data.

1. Minimum:- The smallest data point in the dataset.

2. First Quartile (Q1):- The median of the lower half of the data (25th percentile).

Date __/__/____

3. Median (Q2):- The middle value of the data (50th percentile).

4. Third Quartile (Q3):- The median of the upper half of the data (75th percentile).

5. Maximum:- The largest data point in the dataset.

Example:- 4, 7, 8, 12, 15, 18, 21, 25, 28

- Minimum = 4
- Q1 = 8
- Median = 15
- Q3 = 25
- Maximum = 28

⇒ Uses:- 1.) Summarizes the distribution of data

2) Helps in creating box plots

3) Identifies spread, center, and outliers in data.

Date __/__/____

=> Shape :- It refers to the visual pattern or distribution of data when it's plotted on a graph, usually as a histogram, box plot, or density curve.

=> Types of Shapes :-

1. Symmetrical (Normal Distribution) :-

- The data is evenly distributed on both sides of the mean.

- Bell-shaped curve.
- Mean = Median = Mode
- Example : Heights, test scores

2. Skewed Right (+ve) :- The tail is longer on the right side.

- Mean > Median > Mode
- Example: Income Distribution

3. Skewed Left (-ve) :- The tail is longer on the left side.

- Mean < Median < Mode
- Ex:- Age at retirement.

Date __/__/____

⇒ Kurtosis: It measures the "tailedness" or peakedness of a data distribution.

⇒ Types of Kurtosis:-

1. Mesokurtic (K = 3)

- The distribution has moderate tails.
- Shape: Symmetrical, bell-shaped curve
- Exa:- Standard Normal Distribution

2. Leptokurtic (K > 3) :-

- Tails are heavier than normal - more extreme outliers.
- The peak is sharper and taller.
- Shape: High peak and fear tails
- Exa:- Stock market returns

Date __/__/____

3. Platykurtic (Kurto k<3):-

- Tails are lighter than normal - fewer outliers

- The peak is flatter and broader.
- Shape: Low peak, thin tails.
- Ex:- Uniform distribution.

⇒ ~~Visuals~~

⇒ Visuals:- Visuals are graphical representations of data that help you understand, analyze and communicate patterns, trends and relationships in a dataset

1. Bar Chart:-

- Shows counts or frequencies for categorical data.
- Bars can be vertical or horizontal.
- Ex:- Number of students in different departments.

Date __/__/____

2. Histogram:-

- Similar to a bar chart but for continuous data.
- Shows the distribution of numerical data in intervals.
- Example:- Exam Score Distribution

3. Pie Chart:-

- Circular chart showing proportions or percentages.
- Best for showing parts of a whole.
- Ex:- Market share of companies.

4. Box Plot (Box-and-Whisker plot):-

- Shows the Five-Number summary: min, Q1, median, Q3, max.
- Helps visualize spread, center and outliers.
- Ex:- Compare salaries in two departments.

Date __/__/____

5. Line Graph:-

- Shows trends over time
- Often used for time series data
- Eg:- Monthly sales over a year

6. Scatter Plot:-

- Display the relationship between two quantitative variables.
- Useful for identifying correlation or clusters.
- Ex. Height vs. weight.

Distribution:- It shows the frequency (or probability) of different outcomes or values in a dataset or population.

⇒ Why is Distribution Important?

- Helps understand the data pattern
- Used in making predictions
- Helps choose the right statistical method
- Essential in hypothesis testing

Date __/__/____

Discrete Distribution:- A discrete distribution assigns probabilities to specific values of a discrete random variable.

$$\sum P(x) = 1$$

⇒ Key features:-

- Values are countable
- Each value has an associated probability
- Represented using probability mass functions (PMF)

⇒ Why :-

- Helps in modeling real-world count data
- Essential in decision making, quality control, risk assessment, etc.

Date __/__/____

Bernoulli Distribution :- It is the simplest discrete probability distribution is statistics.

- A random variable X Follows a Bernoulli distribution if:

$P(x=1) = p$ and $P(x=0) = 1-p$

where:

- $X \in \{0, 1\}$
- p = probability of success (1)
- $1-p$ = probability of failure (0)
- $0 \leq p \leq 1$

⇒ Example :-

- Tossing a coin (Head = 1, Tail = 0)
- Pass / Fail in a test

Binomial Distribution :- It is a discrete probability distribution that describes the number of success in a fixed number of independent trials. Two possible outcomes: Success or Failure.

Date __/__/____

$X \sim \text{Binomial}(n, p)$

⇒ Binomial Probability Formula:

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

Where: n = number of trials
k = number of successes
p = probability of success

⇒ Conditions for Binomial Distribution:

1. Fixed number of trials n.
2. Only two outcomes: success or failure
3. Each trial is independent
4. Probability of success p is constant

⇒ Shape of Binomial Distribution:-

- Symmetric when $p = 0.5$
- Skewed when p is closer to 0 or 1
- As n increases, it approximates the normal distribution

⇒ Applications: Quality control, Survey Analysis, Marketing

Date __/__/____

Poisson Destribution :-

- It is a discrete probability distribution that expresses the probability of a given number of events occuring in a fixed interval of time or space.

- The events happen independently.

- The events occur with a constant average rate (λ, "lambda").

- Formula :- $P(y) = \frac{\lambda^y - e^{-\lambda}}{y!}$

⇒ When to Use :- When you are modeling the number of times an event happens in a fixed interval.

- Number of calls received by a call center per hour.

- Number of typing errors per page.

Date __/__/____

Continuous Distribution :- It is a probability distribution in which the set of possible outcomes forms a continuous range of values.

⇒ Normal Distribution :-

The Normal Distribution (also called the Gaussian distribution) is a continuous probability distribution that is symmetrical and bell-shaped, representing how data tends to cluster around a central mean.

- Formula: $f(x) = \frac{1}{\sigma\sqrt{2\pi}} \cdot e^{-\frac{(x-\mu)^2}{2\sigma^2}}$

where: μ = mean
σ = standard deviation
x = value of the variable

- Why It's Important:-

- Many natural phenomena (like height, blood pressure) follow a normal dist.

Date __/__/____

- It's the foundation for many statistical data tests (like z-test, confidence intervals etc.).

⇒ Key Features:-

Property	Description
Shape	Symmetrical bell curve
Mean = Median = Mode	All are the same of the distribution
Spread	Controlled by standard deviation (σ)
Total Area Under curve	Equals 1 (100% probability)
Empirical Rule (68-95-99.7)	68% of data within 1σ, 95% of data within 2σ, 99.7% of data within 3σ of the mean (μ)

Date __/__/____

Standard Normal Distribution:-

- It is a special case of the normal distribution where:

Mean (μ) = 0
Standard Deviation (σ) = 1

It's also called Z-Score

Key Features:

Property	Value / Description
Mean (μ)	0
Standard Deviation	1
Symmetry	Perfectly symmetrical around 0
Total Area under curve	1 (100% probability
Use	To find probabilities and percentiles using Z-scores

Date __/__/____

⇒ What is a Z-score?

A Z-score tells you how many standard deviation a data point is from the mean:

$$Z = \frac{X - \mu}{\sigma}$$

Where:

X = Original value
μ = mean of the distribution
σ = standard deviation

⇒ Why It's Useful:

- Standardizes different normal distributions to a common scale.

- Helps compare data from different sources

- Used in hypothesis testing, confidence intervals and more

Date __/__/____

Sampling Techniques:- Sampling techniques in statistics are methods used to select a subset (sample) from a larger population to analyze and make inferences about the entire population.

⇒ Why Sampling is used?

1. Studying the entire population is impractical or impossible.
2. It saves time, money and resources.
3. Provides quick insights and reliable estimates.

⇒ Types of Sampling:-

1. Simple Random Sampling:- It is a method of selecting a sample of size n from a population of size N in such a way that every possible sample has an equal ~~probability~~ probability of being chosen.

Date __/__/____

⇒ When to use Simple Random Sampling:

- When the population is homogeneous.
- When you have a complete list of the population.
- For surveys, experiments, or descriptive studies.

2. Stratified Sampling:- It involves dividing a population into distinct groups (strata) based on a specific characteristic (like age, gender, income, etc.) and then a random sample is taken from each group.

3. Systematic Sampling:- In systematic sampling, the first element is selected randomly, and the rest are chosen at regular intervals (k) from the ordered population list.

⇒ When to Use:- 1) When a complete list is available.

2) When the population is uniformly distributed.

Date __/__/____

4. Convenience Sampling:- It involves selecting samples that are easiest to reach or readily available to the researcher.

⇒ When to Use:-

- When time or budget is limited.
- For initial testing of ideas.
- When precision is less critical.

Z-Test :-

- It is a hypothesis test that uses the Z-distribution (standard normal distribution) to test whether the observed data significantly deviates from what is expected under the null hypothesis.

⇒ Formula:-

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

Date __/__/____

Where:- $\bar{x}$ = sample mean
μ = population mean
σ = population standard deviation
n = sample size.

⇒ Types of Z-Tests:

Type	Purpose
One-sample	Compare a sample mean to a known population mean.
Two-sample	Compares means of two independent samples.
For proportions	Compare sample proportion to a population proportion, or between two proportion.

⇒ Use Cases:- 1. Quality Control
2. Medical Trials
3. Marketing surveys

Date __/__/____

T-Test :- A T-Test assesses whether the means of two groups (or a sample and population) are statistically different from each other using the t-distribution.

⇒ Formula :-

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

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

⇒ Types of T-Test :-

1. One Sample T-Test :-

- It compares the sample mean to a known population mean.

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

$$df = n - 1$$

Date __/__/____

2. Two- Independent T- Test:-

- It compares the means of two independent groups.

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

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

3. Paired T- Test:- It compares means from the same group at different times.

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

$\bar{x}d \Rightarrow$ mean diff
$sd =$ std diff.

Date __/__/____

Chi-Square Test :- It evaluates whether the differences between observed and expected frequencies in a categorical dataset are due to chance or indicate a significant relationship.

⇒ Types of Chi-Square Test :-

Type	Purpose
Chi-Square Test of Independence	Checks if two categorical variables are related (association).
Chi-Square Goodness-of-Fit Test	Tests whether a single categorical variable follows a specified distribution.

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

where: O = Observed frequency
E = Expected frequency
χ^2 = Chi-square statistic