Hamza Aslam Data Scientist

Notes (Statistics) for Data Science

What is Statistics?

Statistics is a branch of mathematics that deals with collecting,

Organizing, analyzing, interpreting, and presenting data.

It helps in understanding and making decisions based on data.

Types of statistics:

1. Descriptive Statistics

- **Definition:**
  Descriptive statistics ka kaam data ko summarize aur describe karna hota hai. Isme raw data ko meaningful aur simplified form mein present kiya jata hai.
- **Purpose:**
  - Data ka overview dena.

Trends aur patterns ko highlight karna.

1: Population vs sample:

2: central tendency-->mean, maiden, mode

3: Measure of variability

4: Outliers

5: Data visualization tools and plots

2. Inferential Statistics

- **Definition:**
  Inferential statistics ka kaam raw data ke basis par generalizations aur predictions karna hota hai. Iska focus sample data ke zariye puri population ke baare mein conclusions draw karna hota hai.
- **Purpose:**
  - Hypothesis testing karna.
  - Population ke baare mein predictions karna.

What is Data?

Data refers to raw facts, figures, or information collected from various sources,

Which can be analyzed and processed to derive meaningful insights.

It is the foundation for decision-making in fields like science,

Business, technology, and everyday life.

Age group sarway

Stratified Sampling

chotki bjana

Random Sampling

by professional sarway

Cluster Sampling:

Data:

1: Qualitative Data or Categorical Data:

--> Quality

--> No Number

_____> Nominal scale

_____> Ordinal scale

2: Quantitative Data or Numerical Data:

--> Quality

--> Only Number

--> Interval Scale

--> Ratio Scale

-->Arithmetic Data

Scales \ Level of Measurement:

1: Nominal Scale.

--> Male, Female 1, 0

--> Hamza, Majid

--> Eye color, Hair color

Name, labels, qualities but no number and no order

Statistics of Nominal:

Mode, Frequency, chi-square

2: Ordinal Scale:

All Nominal scale or rank and order

--> i agree, i strongly agree , i don't agree

--> First, Second, third

x	-	-
-	x	-
-	-	x

Statistics:

Median, Mode

3: Interval Scale:

--> Numbers Only whole Number

--> Equal difference between Numbers

--> Not True Zero

--> Absolute Zero

---> Temperature in C|F

---> Dates in

Statistics:

Mean, SD, ANOVA, Regression

4: Ratio Scale:

--> Number

--> True zero

--> Addition, - + * /

--> Height, weight, income, distance

Statistics:

Mean, Median, Mode, SD, Geometric Mean, Harmonic Mean

Discrete Data:

--> Integer 1, 2, 3, 4, 5, 6.....

jo Data ap intergar ki madaat sy lik saky or count kar saky or in

K darmyan equal khab ho Discrete Data kahlata a.

--> Car parking = 43

--> Number of children in your Family = 5

Continuous Data:

-->aysa data jo 1 range ki base sy lika ja saky 1.2, 1.3,……

--> Measureable Value: --> Height, Weight, Temperature

Binary Data:

--> Dogla Data

--> Yes or NO --> 0 | 1 --> True | False --> Head | Tails --> On | Off

Time Series Data: ss:mm:hh:dd:

--> The Data points collected\recorded at regular time interval

Is called time series data

--> Daily stack Market, Monthly Rainfall, Temperature

Categorical Data:

aysa data jis ma text use kia jata a

--> We can group data

-->Blood Group (A, AB, AB+, O)

Ordinal Categorical data:

clear order example: Rating,

Educational Level (High school, Under Graduation

Spatial Data:

The Data that has a geographical / spatial component

--> Location on Map

--> GIS

Multivariate Data:

--> aysa data jis ma Multiple data ho (Multiple columns, variables, Attributes

1 Univariate:-->jisme sirf ek variable ho. (e.g., [70, 80, 90, 85])

2 Bivariate:--> jisme do variables involved hon. (e.g., Marks vs. Study Hours)

3+ Multivariate: jisme 3 ya usse zyada variables ho. (Marks, Study Hours, Attendance).

Structured vs Unstructured Data:

Structured Data: SQL, Relational Database, Excels Sheets, Google Sheets

Unstructured Data: Does net fit convention method of storage data

--> Text file data, Multimedia data, webpages data, Audio data, Image data, and video data

Textual Data:

-->Pdf books, E-Book, Social Media post, Comment.....

Semi Structured Data:

--> A mix of structured and unstructured data

Example: Emails, Jason file

Boolean Data:

Data with only two possible value

Example: On, Off

Operationalization:

Example: Stress-->Heat beat

Self-reported felling

Contuses level

Proxy Measurement:

Variable (weight) directly cannot be measure then we find proxy

Example: Tree age, Class attendance

True and error store:

Continuous data.

X = T + E

Observed Value True value + Error

Types of errors:

1. Random errors

2. Systematic errors

Random errors:

Jab b ape 1 kam ko dobara sy kary to us ma koi na koi change a jay

Example: jasy ludo ka dana, weight = 70kg, 70.2kg, 70.5kg

Replicates measure (Mean)

Systematic errors:

Wing balance errors yahi instrument ma khrabi

Example: 1kg oil k packet ka jab weight kary to wo 2kg show ho

Calibration

Type 1 errors: (ON, OFF)

False Positive:

Jab doctor kisy male ko kahe you are pregnant

Example: Aisi cheez jo haqeeqat mein nahi hoti lakin ham usy Positive kehte hain. (jhuti tasalya) "False Hope"

Type 2 errors:

False Negative:

Agar doctor pregnancy test kary or kahe ki patient pregnant nahi hai, lekin wo asal mein pregnant ho, to yeh ek **False Negative** case hoga.

Data Science:

1: Better decision making

2: ML AI Models

3: Better understanding of risk management

4: Ethical Implement

Reliability: Claim is true

Reliability refers to the consistency of measures. Not a lot of difference.

A reliable tool or matured, When it yield the same result under consistent condition

Example: IPhone --> water Prof

Validity: Valid or Accurate

Accuracy of measure

Reliability and Validity:

Data is new oil.

1: Trust worthy

2: Sound decisions

3: Ethical Research

4: Effective Solution

Triangulation: Triangle (Reliability and validity)

Urban Pollution: (Lahore Pollution)

1: Health record --> disease

2: Air quality metrics

3: Satellite Images

Pillars:

1: Data Triangulation (Gathering data from difference scoured to answer the same question)

2: Methodological Triangulation (using difference method and technique to collect and analysis same data)

3: Theoretical Triangulation (difference theory use)

4: Investigate Triangulation (multiple research observe same study)

Help out:

1: credibility

2: Reduces Bias

3: Increase confidence on findings

4: Comprehensive view

Limitation:

1: Resource Intensive

2: Complex integration

3: Skill requirement

Surrogate endpoints: Final cannot be easily measured

Substitute measure for a clinical endpoint

Example: Heart attack (BP, cholesterol level)

1: Reassure short hand

2: Ethics

3: save money

Biasness:

Biasness woh condition hai jahan kisi external ya internal factor ki wajah se data aur uske results asal reality ko represent nahi karte.

It effects: Yeh results ki accuracy aur decision-making par negative impact dalta hai.

Example: Aapne green glasses pahne hain aur sochte hain ke sab kuch green hai.

It effects: Aapka data yeh batayega ke "har cheez green hai."Magar asal mein duniya colorful hai (green, blue, red, etc.).Aapki observation biased hai, kyunki aap ek distorted lens se dekh rahe hain.

Bias in Data:

1: Sampling Bias

2: Selection Bias

3: Conformation Bias

4: Survivorship Bias

5: Reporting Bias

Main Bias is called Information

1: Sampling Bias:

Jab data ka sample population ko theek tarah se represent nahi karta.

It effects: Pooray population ke liye galat conclusions nikale jate hain.
Example: Ek survey sirf college student's se karein, jabki target poori population hai.

2: Selection Bias:

Jab participants ko aise tarike se select kiya jata hai jo non-random ho aur specific groups ko favor karta ho.

It effects: Results skewed ho jate hain aur asal reality ko represent nahi karte.
Example: Ek experiment sirf healthy logon par kiya jaye, aur conclusion sab logon ke liye apply kiya jaye.

3: Conformation Bias:

Jab sirf us data ko note ya analyze kiya jata hai jo pehle se banayi hui sochon ko confirm karta ho.

It effects: Dusri important information ignore ho jati hai, jo better decision-making mein madadgar ho sakti thi.

Example: Ek doctor sirf un cases ko dekhta hai jo uski treatment ke positive results dikhate hain, aur failures ko ignore karta hai.

4: Survirovship Bias:

Jab sirf successful cases ko analyze kiya jata hai aur failures ko ignore kiya jata hai.

It effects: Success rate ko overestimate kiya jata hai.
Example: Startups ke analysis mein sirf successful companies ka data lena aur fail hone wali companies ko ignore karna.

5: Reporting Bias:

Jab results ya data ko selectively report kiya jata hai aur kuch cheezon ko intentionally suppress kiya jata hai.

It effects: Galat ya incomplete picture samajh aati hai.
Example: Ek medicine ke sirf positive results publish karna aur side effects ki information ko chhupa lena.

Ek Example Sab Types Ke Saath:

Agar ek university apne student’s ka data analyze karna chahti hai:

1. **Sampling Bias:** Sirf high-achieving student’s ka data lena aur average students ko exclude karna.
2. **Selection Bias:** Sirf ek department ke students ka data lena, jabki poori university ka analysis karna hai.
3. **Confirmation Bias:** Sirf usi data ko consider karna jo university ke "achhe results" ko support kare.
4. **Survivorship Bias:** Sirf un graduates ka data analyze karna jo successful careers mein hain, aur failures ko ignore karna.
5. **Reporting Bias:** Sirf positive findings report karna aur negative feedback chhupa lena.

Bias kasy kam ki jaty a:

1. Improve sampling Methods
2. Refined Research Designs
3. Data Collection Vigilance
4. Analytical Awareness (Pre Review)
5. Ongoing Education and Awareness (Books Reading, Courses kare, Critical thinking improve)

Bias Ko Kam Karne Ka Ek Holistic Approach:

1. **Planning:** Research aur data collection shuru karne se pehle, bias ko pehchanne ke tools aur techniques ready rakhein.
2. **Implementation:** Har stage pe unbiased methods ka use karna.
3. **Review:** Har result ko critically evaluate karein aur peer review ke liye bhejein.

Stastistics:

Statistics ek **science** hai jo data ke saath deal karti hai. Yeh ek systematic approach provide karti hai **data ko collect, analyze, interpret, present, aur organize** karne ke liye, takay humein logical aur meaningful insights mil sakein.

1. **Collecting**
    - Data ko gather karna, jisme surveys, experiments, observations, aur records ka use hota hai.
    - **Example:** Kisi school ke student's ke grades collect karna.
2. **Analyzing**
    - Data ko process karna aur patterns ya trends ko dhoondhna.
    - **Example:** Average calculate karna ya data distribution check karna.

3. **Interpreting**
    - Data ke results ko samajhna aur unka meaning nikalna.
    - **Example:** Yeh samajhna ke grades improve ho rahe hain ya nahi.
4. **Presenting**
    - Results ko clear aur understandable tarike se dikhana, jisme charts, graphs, aur tables ka use hota hai.
    - **Example:** Pie charts ya bar graphs banana.
5. **Organizing**
    - Data ko systematic aur structured form mein arrange karna.
    - **Example:** Data ko categories mein divide karna (e.g., students by class).

Statistics Ka Maqsad (Purpose):

- Decision-making ko support karna.
- Trends aur patterns ko identify karna.
- Future predictions karne mein madad karna.

Real-Life Examples of Statistics:

1. **Business:**
    - Sales trends analyze karna aur marketing strategies banani.
2. **Health:**
    - Vaccine trials ke data ka analysis aur conclusions.
3. **Sports:**
    - Player's ke performance stats analyze karna.
4. **Education:**
    - School aur university ke results ka evaluation.

A --> Decision Making

B --> Clarity and Precision

Data Science:

Data deiven decision making

Data Science ek interdisciplinary field hai jo **data** ke analysis, processing, aur interpretation ke zariye insights aur decisions banane par focus karti hai. Is field mein advanced tools aur techniques ka use hota hai jo data ke raw forms ko useful aur meaningful information mein convert karti hain.

1. **Data Collection:**
   - Data ko gather karna (e.g., surveys, sensors, APIs, web scraping).
   - **Example:** E-commerce website ke customer behavior ka data.
2. **Data Cleaning:**
   - Raw data ko process aur refine karna (missing values, duplicates, errors remove karna).
   - **Example:** Null entries aur incorrect formats ko fix karna.
3. **Data Exploration:**
   - Patterns aur trends ko samajhna using statistics aur visualization.
   - **Example:** Sales data ka analysis karna to check peak months.
4. **Data Analysis:**
   - Data ko analyze karne ke liye machine learning aur statistical algorithms ka use.
   - **Example:** Predictive modeling karna to forecast future sales.
5. **Data Visualization:**
   - Data ko readable aur understandable format mein present karna (charts, graphs).
   - **Example:** Bar charts ya heatmaps ke zariye trends dikhana.
6. **Deployment & Insights Sharing:**

- Final results ko stakeholders ke saath share karna aur actionable insights dena.
- **Example:** Marketing team ko recommend karna ki kis product ka promotion zaroori hai.

Core Components of Data Science:

1. **Statistics & Mathematics:**
   - Data ke patterns aur relationships samajhne ke liye zaroori hai.
2. **Programming:**
   - Tools: Python, R, SQL.
3. **Data Visualization:**
   - Tools: Tableau, Matplotlib, Seaborn, Power BI.
4. **Machine Learning:**
   - Algorithms jo future predictions aur automation mein help karte hain.
5. **Big Data:**
   - Large datasets ko process karna using tools like Hadoop aur Spark.

Examples of Data Science in Real Life:

1. **Healthcare:**
   - Patient data ka analysis karna diseases ko predict aur treat karne ke liye.
2. **Finance:**
   - Fraud detection aur credit risk analysis.
3. **E-commerce:**
   - Customer recommendations (e.g., "Customers also bought this").
4. **Entertainment:**

- Streaming platforms (e.g., Netflix, YouTube) par personalized recommendations.

Why Data Science is Important?

- **Decision-Making:** Accurate aur data-driven decisions ko support karta hai.
- **Problem Solving:** Complex problems ko efficiently solve karta hai.
- **Innovation:** New technologies aur products develop karne mein madad karta hai.

Computer→ Calculator

Programing→ Excel sheet

ML→ Normal Statistics Algorithms

Types of Statistics:

Statistics ki do main types hoti hain, jo data ko analyze karne ke liye alag-alag approaches aur methods provide karti hain.

1. Descriptive Statistics

- **Definition:**
  Descriptive statistics ka kaam data ko summarize aur describe karna hota hai. Isme raw data ko meaningful aur simplified form mein present kiya jata hai.

- **Purpose:**
  - Data ka overview dena.
  - Trends aur patterns ko highlight karna.
- **Techniques:**

1. **Measures of Central Tendency:**
   - Mean (average)
   - Median (middle value)
   - Mode (most frequent value)
2. **Measures of Dispersion:**
   - Range (maximum - minimum)
   - Variance
   - Standard Deviation
3. **Visualization:**
   - Graphs (Bar charts, Histograms)
   - Tables

- **Example:**
  - Average temperature of a city in a month.
  - Percentage of students passing in an exam.

2. Inferential Statistics

- **Definition:**
  Inferential statistics ka kaam raw data ke basis par generalizations aur predictions karna hota hai. Iska focus sample data ke zariye puri population ke baare mein conclusions draw karna hota hai.
- **Purpose:**
  - Hypothesis testing karna.
  - Population ke baare mein predictions karna.
- **Techniques:**

1. **Hypothesis Testing:**
   - T-test, z-test, chi-square test.
2. **Confidence Intervals:**
   - Range within which population parameter lies.

3. **Regression Analysis:**
    - Linear regression, logistic regression.

- **Example:**
    - Survey data ke basis par puri city ki voting preferences predict karna.
    - Medical trial ke sample se puri population par medicine ka effect estimate karna.

A: Paramaetric

B: Un-Paramatric

1: Descriptive Statistics:

Descriptive statistics ka kaam **data ko summarize aur describe karna** hota hai, taki raw data se meaningful information extract ki ja sake. Iska use **Exploratory Data Analysis (EDA)** ka first step hai, jo data ke patterns aur insights ko explore karne mein madad karta hai.

Key Components of Descriptive Statistics

Techniques in Descriptive Statistics

A. Summarize and Describe Data

- Summarization ka matlab hai raw data ko concise aur understandable form mein represent karna.

B. Measures of Central Tendency

- **Mean (Average):** Sab values ka sum divided by count.
- **Median:** Middle value after sorting.
- **Mode:** Most frequent value.

C. Measures of Variability

- **Range:** Max - Min values.
- **Variance:** Spread of data points from the mean.
- **Standard Deviation:** Square root of variance (how spread out values are).

D. Data Insights:

- **Visualization Tools:**
  - Histograms, Boxplots, Scatterplots.
- **Summary Statistics:**
  - Minimum, Maximum, Quartiles, Percentiles.

Purpose of Descriptive Statistics in EDA

1. **Understand the Dataset:**
   - Identify data's structure, key attributes, and potential issues.
2. **Generate Hypotheses:**
   - Prepare for inferential analysis by spotting trends or relationships.
3. **Provide Insights:**
   - Present data in a clear and interpretable format for decision-making.

Real-Life Example:

Dataset: Sales of a Retail Store

- **Mean:** Average sales per day = 500 units.
- **Median:** Middle sales value = 480 units.

- **Mode:** Most frequent sales = 450 units.
- **Range:** Highest sales (800) - Lowest sales (200) = 600 units.
- **Visualization:** A histogram showing sales distribution, identifying days with higher sales

2: Inferential Statistics:

Sample se population ke baare mein **inference lagana aur predictions karna** inferential statistics ka core idea hai.

Inferential Techniques for Predictions

A. Hypothesis Testing

- **Definition:**
  Ek systematic method to test whether a specific assumption (hypothesis) about a population is true.
- **Example:**
  - Null Hypothesis (H_0): Mean age of a population = 30 years.
  - Alternate Hypothesis (H_1): Mean age ≠ 30 years.
- **Steps:**
  1. Sample data analyze karo.
  2. Statistical test lagao (e.g., t-test, chi-square).
  3. Decision lo (accept/reject hypothesis).

B. Confidence Interval

- **Definition:**
  Range of values jisme ek population parameter (e.g., mean) honay ka high probability ho.
- **Example:**
  - "Population mean salary is likely between Rs50,000 and Rs55,000 with a 95% confidence level."

- **Interpretation:**
  - 95% ka matlab hai agar 100 samples liye gaye, to 95 samples ka mean iss range mein aayega.

C. Machine Learning (ML)

- **Role of ML:**
  Sample data ko use karke population ke baare mein predictive models banane ka process.
  - Example:
    1. Customer preferences ka prediction.
    2. Disease diagnosis models based on patient data.
- **Connection to Statistics:**
  ML algorithms ko train karne ke liye descriptive aur inferential statistics ka use hota hai.

- Important methods:
  1. **Hypothesis Testing:** Validity check of assumptions.
  2. **Confidence Intervals:** Reliable range estimation.
  3. **Machine Learning:** Advanced predictions based on sample data.

How to Use Statistics:

Statistics ka istemal **real-world problems ko samajhne aur solve karne** ke liye kiya jata hai. Yeh science data ko collect, analyze, aur interpret karne mein madad karti hai, jo decision-making ke liye critical hoti hai.

Real-World Examples of Statistics:

1. Business and Marketing

- **Application:**
  Sales trends aur customer preferences analyze karna.
- **Example:**
  - E-commerce platforms user data analyze karke personalized recommendations deti hain.

2. Healthcare

- **Application:**
  Patient data ko analyze karke diseases ka early detection aur treatment planning.
- **Example:**
  - COVID-19 case trends analyze karna aur vaccination campaigns plan karna.

3. Education

- **Application:**
  Student performance data analyze karna for better teaching methods.
- **Example:**
  - A school's test score distribution analyze karke weak areas identify karna.

4. Sports

- **Application:**
  Player performance aur game strategies analyze karna.
- **Example:**
  - Cricket teams ke players ke batting aur bowling averages analyze karna.

5. Finance and Banking

- **Application:**
  Risk assessment aur stock market trends ka analysis.
- **Example:**
  - Insurance companies customer data analyze karke premium rates set karti hain.

6. Government and Policy Making

- **Application:**
  Census aur survey data ke basis par policies banana.
- **Example:**
  - Unemployment rate analyze karke job creation programs plan karna.

7. Agriculture

- **Application:**
  Crop yield aur weather data analyze karke farming decisions lena.
- **Example:**
  - Rainfall aur soil quality ke basis par suitable crops grow karna.

8. Technology and AI

- **Application:**
  Machine Learning aur Artificial Intelligence ke models develop karna.
- **Example:**
  - Social media platforms statistical algorithms use karte hain to show personalized feeds.

Data Analysis:

Data Analysis ek systematic process hai jo data ko collect, organize, interpret, aur visualize karne ke liye kiya jata hai. Iska goal hai meaningful insights aur patterns identify karna jo **decision-making** aur problem-solving mein madad kar sake.

Steps of Data Analysis

1. **Data Collection:**
   - Relevant data ko gather karna.
   - Example: Surveys, databases, sensors, or APIs se data lena.
2. **Data Cleaning:**
   - Missing values, duplicates, aur outliers ko handle karna.
   - Example: Incomplete survey responses ko fill karna ya remove karna.
3. **Data Exploration (EDA):**
   - Data ko visualize aur summarize karna.
   - Tools: Scatter plots, histograms, bar charts, etc.
   - Example: Sales data ke trends aur patterns identify karna.
4. **Data Transformation:**
   - Data ko analysis ke liye process karna, e.g., normalize karna, aggregations karna.
   - Example: Monthly sales ko yearly trends mein convert karna.
5. **Data Modeling:**
   - Mathematical models ya machine learning techniques ka use karna.
   - Example: Predict karna ki agle month ki sales kitni hongi.
6. **Data Interpretation:**
   - Results ko analyze karna aur meaningful insights nikalna.
   - Example: Sales increase ka sabab identify karna, jaise promotional campaigns.

7. **Data Presentation:**
    - Results ko understandable format mein share karna.
    - Tools: Reports, dashboards, visualizations (e.g., Tableau, Power BI).

Importance of Data Analysis

1. **Decision-Making:**
    - Accurate aur data-driven decisions ke liye.
    - Example: A company decides to invest in a new market based on trends.
2. **Problem-Solving:**
    - Challenges aur bottlenecks identify karna.
    - Example: Low customer retention rate ka sabab samajhna aur uska solution dhundhna.
3. **Process Optimization:**
    - Efficiency aur productivity improve karna.
    - Example: E-commerce delivery times ka analysis aur optimization.
4. **Risk Management:**
    - Potential risks identify aur mitigate karna.
    - Example: Financial institutions fraudulent transactions detect karte hain.
5. **Future Predictions:**
    - Trends aur patterns ke basis par forecasting karna.
    - Example: Weather forecasting ya stock market predictions.

Types of Data Analysis

1. **Descriptive Analysis:**
    - Data ko summarize aur describe karna.

    - Example: Average sales per month.
2. **Diagnostic Analysis:**
    - Pichle results ka sabab analyze karna.
    - Example: Sales drop ka reason kya hai?
3. **Predictive Analysis:**
    - Future trends aur outcomes ko predict karna.
    - Example: Agle quarter ki sales forecast karna.
4. **Prescriptive Analysis:**
    - Solutions aur recommendations dena.
    - Example: Marketing budget ko allocate karne ka best tariqa kya hai?

Real-World Examples of Data Analysis

1. **E-commerce:**
    - Customer behavior analyze karke product recommendations dena.
    - Tools: Google Analytics, Tableau.
2. **Healthcare:**
    - Patient data analyze karke disease prediction models banana.
    - Tools: Python (Pandas, NumPy).
3. **Sports:**
    - Player performance aur team strategies optimize karna.
    - Example: Cricket mein data-driven field placements.
4. **Finance:**
    - Market trends analyze karna aur risk assess karna.
    - Example: Stock market predictions using time series analysis.
5. **Education:**
    - Student performance data analyze karna aur personalized learning plans develop karna.

Types of Data Analysis

1: Descriptive data analysis ----> sar dard

2: Diagnostic data analysis -----> wajah

3: Predictive data analysis (ML, Dep learing)-->Future

4: Prescriptive data analysism ----->Desprine

5: EDA --> Exploratory data analysis

6: Inferential analysis

7: Causal analysis

8: Mechanistic analysis

1: Descriptive Data Analysis

- **Purpose:**
  Data ko summarize aur describe karna.
- **Example:**
  Sales data ka average aur total calculate karna.
- **Real-Life Analogy:**
  "Sar dard ki symptom ko identify karna."

2: Diagnostic Data Analysis

- **Purpose:**
  Past events aur problems ki wajah analyze karna.

- **Example:**
  Sales decrease hone ka sabab analyze karna (e.g., low demand, bad marketing).
- **Real-Life Analogy:**
  "Sar dard ki wajah dhoondhna."

3: Predictive Data Analysis (ML, Deep Learning)

- **Purpose:**
  Future outcomes ko predict karna.
- **Example:**
  Machine Learning ka use karke agle mahine ki sales forecast karna.
- **Real-Life Analogy:**
  "Aane wale time mein kya hoga (future predict karna)."

4: Prescriptive Data Analysis

- **Purpose:**
  Solutions aur recommendations dena.
- **Example:**
  Marketing budget allocate karne ke best tariqe recommend karna.
- **Real-Life Analogy:**
  "Sar dard ke liye aspirin prescribe karna."

5: EDA (Exploratory Data Analysis)

- **Purpose:**
  Data ko explore karna aur trends, patterns, aur relationships ko samajhna.

- **Example:**
  Scatter plots aur histograms ka use karna to visualize data trends.
- **Real-Life Analogy:**
  "Data ke bare mein basic information ikattha karna aur insights lena."

6: Inferential Analysis

- **Purpose:**
  Sample data ke basis par population ke bare mein conclusions draw karna.
- **Example:**
  Election polls ke zariye ek region ki preferences predict karna.
- **Real-Life Analogy:**
  "Ek choti sample ke basis par bade group ke behavior ka andaza lagana."

7: Causal Analysis

- **Purpose:**
  Cause-and-effect relationships samajhna.
- **Example:**
  Smoking aur lung cancer ke beech ki causal relationship ko analyze karna.
- **Real-Life Analogy:**
  "Ek kaam hone ka doosre kaam par kya asar hota hai."

8: Mechanistic Analysis

- **Purpose:**
  Data ke underlying processes ko samajhna jo specific behavior ko control karte hain.
- **Example:**
  Medicine ke ek chemical component ka human body par asar study karna.
- **Real-Life Analogy:**
  "Har cheez ka internal mechanism samajhna jo system ko chalata hai."

Central tendency:

Central tendency ek statistical measure hai jo data ke "center" ya "average" value ko represent karta hai. Iska purpose data distribution ke ek representative point ko identify karna hai jo uske around rotate karta hai.

Why Central Tendency Matters?

1. Data ka ek **representative value** provide karta hai.
2. **Decision-making** mein help karta hai:
   - Example: Ek business ko apni average sales samajhne mein.
3. Different **distributions** ko compare karne mein.

Example: Class ke students ke marks ka average, Sweed ka diba jab open hota a to makya ka jurmat central tendency kahlta a

1: Mean (Average)

2: Median (center of data)

3: Mode (repeatedly value)

1: Mean (Average)

I): **Arithmetic mean** -->Sabhi values ka sum, aur un values ki total quantity se divide karna.

II): **Geometric mean** --> Sabhi values ka **product** lete hain aur uska nth root calculate karte hain, jahan n total numbers ki quantity hai.

III): **Harmonic mean** --> Values ke reciprocals ka mean, aur fir us reciprocal ko reverse karna. Zyada speed aur rates ke analysis ke liye use hota hai.

Reciprocal of the arithmetic mean (total number divided by sum of all no)

Iv): **Weighted mean** --> Jab dataset ki individual values ka importance alag-alag ho.

Example: Ek student ke grades aur credit hours (GPA), where each course has a different credit weight.

-->consider exam scores: [80, 90, 70] with weight (Percentage of

Total grade): [50%, 25%, 25%]

Weighted mean = (0.5x80 + 0.25x90 + 0.25x70) divider by (0.5+0.25+0.25)

v): **Truncated (or Trimmed) mean**--> involves removing a certain percentage of the

Smallest and largest values before calcullating the mean

Example :--> dataset= [1, 2, 3, 4, 5, 6]

Truncated mean = (2+3+4+5) divided by total num

2: Median (center of data)

Median ek aisi value hai jo **sorted dataset** ko exactly **middle** mein divide karti hai.

- Agar total numbers odd hain, to middle value median hoti hai.
- Agar total numbers even hain, to beech ke dono numbers ka average median hota hai.

Important Condition: Data ko sort karna zaroori hai (ascending ya descending order mein).

3: Mode (repeatedly value)

Mode ek aisi value hai jo dataset mein sabse zyada baar repeat hoti hai. Agar ek se zyada values barabar frequency ke saath repeat hoti hain, to dataset **multimodal** ho sakta hai.

Use in E-comerce and Market trands and SEO and Education

Python Functions for Mode Analysis:

1. **`.mode()`** → Mode calculate karne ke liye.
2. **`.unique()`** → Unique values ki list dikhata hai.
3. **`.nunique()`** → Unique values ki count return karta hai.
4. **`.value_counts()`** → Har value ki frequency batata hai.

Importance:

1. **Simplicity (Mean):**
   - Mode ko samajhna aur calculate karna asaan hai.
2. **Comparison (Mean):**

- Mode ko mean aur median ke saath compare kar ke data ke distribution ka analysis kiya jaa sakta hai.

3. **Decision Making:**
   - Frequently occurring value ke basis par informed decisions liye ja sakte hain.
   - Example: Sabse zyada bikne wala product identify karna.
4. **Foundation for Advanced Statistical Analysis:**
   - Mode advanced statistical tools ka base hota hai.

Limitations:

Sensitive to outliers

Variability (Spread / Dispersion)

Variability, jo ke **dispersion** bhi kaha jata hai, ek dataset ki heartbeats hai. Iska matlab hai ke data ka **spread** ya **diversity** kitna hai, jo ek dataset ko samajhne mein madad karta hai.

Beyond Average

Jab hum **average** (mean) ki baat karte hain, to hum sirf ek central value dekh rahe hote hain. Lekin **variability** ya **dispersion** average ke aage ka concept hai. Yeh batata hai ke data values average se kitni door hain, yani data kitna spread hai.

Data Spread as the Heartbeat

Data ka spread ya dispersion wo hai jo ek dataset ke baare mein zyada information de sakta hai. Agar kisi dataset mein high variability ho, to iska matlab hai ke data kaafi diverse hai aur predictions ya generalizations karna mushkil ho sakta hai. Agar variability low ho, to data kaafi consistent hai, aur predictions zyada reliable ho sakte hain.

Decision Making

Variability ka analysis decision-making mein kaafi useful hota hai. Agar

aapko business, research, ya kisi analysis mein decision lena ho, to yeh samajhna zaroori hota hai ke aapke data ka spread kaisa hai. High variability ka matlab ho sakta hai ke aapko zyada careful analysis ki zaroorat hai, jab ke low variability ka matlab hai ke aap zyada confident ho sakte hain apne decisions mein.

Types of variability

Center --> Mean, Median, Mode

Range of spread --> Range=Maximum Value–Minimum Value

Interquartile Range (IQR 50%) -->

Data ko ascending order mein sort karna hota hai, phir **Q1 (Lower Quartile)**, **Q3 (Upper Quartile)**, aur IQR ko calculate karte hain. Yeh process outliers ko identify karne ke liye bhi use hota hai.

IQR=Q3−Q1

Q1 - 1.5 x IQR

Q3 + 1.5 x IQR

Variance-->Data ke values ki average squared distance from the mean.

I) **Population variance:**

Sigma square = Sum(x1 - mu) square divided by capital N

II) **sample variance:**

S square = Sum(x1 – x bar) square divided by n-1

Standard Deviation--> Variance ka square root. Yeh measure karta hai ke data kitna spread hai original unit's mein.

Gaussian | Normal Distribution

If SD ki value 1 ho gi to 68% data ho ga (mu ± 1σ)

If SD ki value 2 ho gi to 95% data ho ga (mu ± 2σ)

If SD ki value 3 ho gi to 99.7% data ho ga (mu ± 3σ)

Standard error-->

Sample mean ki **precision** ko measure karta hai.

Mean as an estimate of the population mean.

Standard error = standard deviation divided by under root of number of sample size

Data:

Data kisi bhi information ya facts ka collection hota hai jo analysis, understanding, ya decision-making ke liye use hota hai. Isko numbers, text, images, sounds, ya kisi aur form mein collect kiya ja sakta hai.

Types of Data:

1. **Qualitative Data (Categorical Data):**

Wo data jo categories ya qualities ko represent kare.

Example:

Gender: Male, Female

Colors: Red, Blue, Green

2. **Quantitative Data (Numerical Data):**

Wo data jo numbers ke form mein ho aur measure ya count kiya ja sake.

Example:

Age: 25, 30, 45

Height: 170 cm, 180 cm

Example of Data:

Class ke students ke marks:

85, 90, 78, 88, 76

Survey mein logon ka favorite food:

Biryani, Pizza, Pasta

Continuous Data:

Continuous data wo hota hai jo kisi bhi range mein infinite values le sakta hai. Iska measurement decimal ya fractional values mein hota hai, aur ye uninterrupted hota hai.

Examples:

Temperature: 36.5°C, 37.8°C

Height: 170.2 cm, 180.5 cm

Weight: 55.4 kg, 70.8 kg

Discrete Data:

Discrete data wo hota hai jo countable values par based hota hai. Ye data whole numbers mein hota hai aur ismein fractions ya decimals nahi hote.

Examples:

Number of students in a class: 25, 30

Number of cars in a parking lot: 10, 15

Tossing a coin (outcomes): Heads, Tails

Difference between Discrete and Continuous Data

Aspect	Discrete Data	Continuous Data
Values	Countable (finite/integer)	Infinite within a range (real numbers)
Measurement	Exact numbers (whole numbers)	Measured with precision (fractions/decimals)
Example	Students in a class (10, 20, etc.)	Height (170.5 cm, 172.8 cm, etc.)

Distributions for Continuous Data

1. **Uniform Distribution**

All outcomes have equal probability.

Example: Rolling a fair dice.

2. **Normal Distribution**

Bell-shaped curve, symmetric around the mean.

Example: Heights of people.

3. Exponential Distribution

Time between events in a Poisson process.

Example: Time to repair a machine.

4. T-Test Distribution

Used for hypothesis testing with small samples.

Example: Comparing means of two groups.

Distributions for Discrete Data

1. Binomial Distribution

Number of successes in a fixed number of trials.

Example: Tossing a coin 10 times.

2. Bernoulli Distribution

Single trial with two possible outcomes (success/failure).

Example: Flipping a coin once.

Symmetrical Data Distribution:

A data distribution is said to be **symmetrical** when the **mean**, **median**, and **mode** are the same. In other words, the left and right sides of the distribution are mirror images of each other.

Example:
A **normal distribution** (bell curve) is symmetrical because the mean, median, and mode all align at the center.

Positive Skew (Right Skew):

A **positive skew** (or right skew) refers to a distribution where the **right tail** (higher values) is longer or fatter than the left tail. In such distributions, the **mean** is greater than the **median**, and the **median** is greater than the **mode**.

Example:

Income distribution: A majority of people earn below the average, but a few high earners create a long right tail (positive skew).

Negative Skew (Left Skew):

A **negative skew** (or left skew) refers to a distribution where the **left tail** (lower values) is longer or fatter than the right tail. In such distributions, the **mean** is smaller than the **median**, and the **median** is smaller than the **mode**.

Example:

Age at retirement: Most people retire around 60-65 years, but a few people may retire earlier, creating a longer left tail.

Skewness:

Skewness measures the **degree of asymmetry** of a distribution around its mean. It indicates whether the data points are

skewed to the left (negative skew) or to the right (positive skew) of the mean.

Types of Skewness:

Positive Skew:

Distribution with a longer right tail (higher values).

Example: Income distribution, where most people earn less, but a few earn very high incomes.

Negative Skew:

Distribution with a longer left tail (lower values).

Example: Age at retirement, where most people retire around a certain age, but a few retire earlier.

Summary of Skewness:

Skew Type	Direction of Tail	Mean vs Median vs Mode	Example
Symmetrical	No tail (balanced)	Mean = Median = Mode	Normal distribution
Positive Skew	Right tail (longer)	Mean > Median > Mode	Income distribution
Negative Skew	Left tail (longer)	Mean < Median < Mode	Age at retirement

Skewness Formula:

Skewness (G) = n divided by [(n−1)(n−2)] ∑i=1n (xi–x¯divided by (s))^3

Where:

n = Number of data points

xi = Individual data point

x^{-} = Mean of the dataset

s = Standard deviation

Kurtosis: The Tailedness Measure

Kurtosis measures the "tailedness" of a data distribution, i.e., how the tails of the distribution behave and how sharp the peak is.

High Kurtosis (Leptokurtic): Distributions with heavy tails and a sharp peak.

Low Kurtosis (Platykurtic): Distributions with lighter tails and a flatter peak.

Example:

A dataset with **similar exam scores** might have high kurtosis (sharp peak and heavy tails), while a dataset with **more varied scores** will have low kurtosis (flatter peak).

Kurtosis Formula:

Kurtosis (K) = n(n+1) divided by [(n−1)(n−2)(n−3)] ∑i=1n((xi–x⁻) divided by

(s))^4 – [3(n−1)]^2 divided by [(n−2)(n−3)]

Where:

n = Number of data points

xi = Individual data point

$\bar{x}$ = Mean of the dataset

s = Standard deviation

This formula calculates the standardized fourth moment (the sum of the fourth power of deviations from the mean, divided by the standard deviation to the fourth power).

The term at the end adjusts for the kurtosis of a normal distribution, making the kurtosis of the normal distribution zero (this is sometimes called “excess kurtosis”).

Null Hypothesis (H_0): The data is normally distributed.

Alternative Hypothesis (H_1): The data is not normally distributed.

Shapiro-Wilk Test:

Purpose: Checks if the data follows a normal distribution.

Example: Suppose we have the ages of 10 people in a class: [18, 22, 21, 19, 20, 22, 21, 22, 20, 21].

shapiro wilk test to test the normally of the data in python code

stats.shapiro(data) --->jab p-value 0.5 sa zyada ho then data is normally distributed

D'Agostino's K^2 Test:

Purpose: A test for normality based on skewness and kurtosis (the shape of the distribution).

Example: Let's say you have the following scores of 10 students: [80, 85, 88, 90, 95, 85, 90, 92, 87, 88]. This test will check for asymmetry (skewness) and tail shape (kurtosis).

Data Collection:

Data Collection is the process of gathering and measuring information on variables of interest. It can be done using different methods based on whether the data is collected directly or from pre-existing sources.

Primary Data:

This is data that you **collect yourself** directly from the source. You gather, measure, and analyze this data personally. Primary data is typically used when you're performing a study or experiment.

Examples of Primary Data Collection Methods:

Research/Experiment: Collecting data through experiments or structured research activities. For example, conducting a lab experiment to observe reactions.

Interview Data Collection: Gathering information directly from individuals through interviews. This can be structured (set questions) or unstructured (open-ended).

Primary Reagent Data: This could refer to data collected from primary sources, like laboratory testing or scientific observations (specific to a domain like chemistry or biology).

Questionnaire (Google Form): Collecting data through structured surveys or questionnaires, often using online platforms like Google Forms or other survey tools.

Audio/Video Data (Podcast): Collecting data in the form of audio or video, for instance, through interviews, podcasts, or any other audio-visual means.

Advantages:

Data is fresh, tailored to your specific research needs.

Greater control over the data quality and methodology.

Disadvantages:

It is time-consuming and can be expensive.

Requires considerable effort to collect and analyze.

Secondary Data:

This is data that has already been **collected by someone else**. You are using existing data rather than collecting it yourself. In secondary data collection, the original data collector is not analyzing the data, you are.

Examples of Secondary Data:

Google: Searching online for already published data, reports, or articles.

GitHub: Accessing datasets or code repositories that have pre-collected data.

UCL (University College London) ML Datasets: Using datasets provided by institutions for machine learning research or training.

Advantages:

Data is readily available and can be accessed quickly.

It saves time and resources as the data is already collected.

Disadvantages:

May not be perfectly aligned with your specific research questions.

Data might be outdated or not tailored to your exact needs.

In summary, **Primary Data** is what you collect yourself, while **Secondary Data** is data that others have already collected and made available.

Sample Randomization:

Random sampling involves selecting a subset of individuals from a population, ensuring each individual has an equal chance of being chosen. This process reduces bias and error, making your data more representative and reliable. When randomization is done well, it helps eliminate selection bias and gives a clearer picture of the whole population.

Examples of variables to randomize in a study:

1. Age
2. Gender
3. Area
4. Qualification
5. Special interests or skills
6. Rich or poor (socio-economic status)

Best Practices to Collect Data:

1. Goal or Aim (write them down)-->Clearly defined
2. Choose the right data collation method
3. Ensure data accuracy and reliability
4. Ensure Privacy of data and ethics
5. Plan of data store and organize
6. Train your team
7. Pilot test
8. Document the process of data collation

Why is Sample Randomization Important?

1. Reduces Bias
2. Increases Validity
3. Prevents Manipulation

Type of Sampling:

1. Representative or Probability sampling:

1.1. Simple Random Sampling-->the fair game

1.2. Systematic Sampling -->Orderly approach

1.3. Stratified Sampling -->the mini-group

1.4. Cluster Sampling -->the miniature mirror

2. Non-Representative or Non-Probability Sampling

2.1. Convenience Sampling --->mary karebi janany waly

2.2. Haphazard Sampling --->ap gumy pery jaha ap ko koi mily use masly wala usy la ly

2.3. Purposive Sampling --->jis ma pahly sy purpos define ho

2.4. Judgmental Sampling --->the expert's choice

2.5. Snow ball Sampling --->the networking web

2.6. Quota Sampling --->the proportional mimic

Descriptive Statistics:

Descriptive statistics are techniques used to summarize, describe, and present data in an easily understandable format. They allow you to explore the basic characteristics of a data set, like its central tendency, variability, and distribution, without making inferences or predictions.

Components of Descriptive Statistics:

1. Population vs sample
2. central tendency-->mean, maiden, mode
3. Measure of variability
4. Outliers
5. Data visualization tools and plots

Data Analysis: (Composition, Distribution, Comparison, Relationship)

Andrew Abela's Plotting Guide:

Andrew Abela's guide to plotting emphasizes the importance of clear, effective visualization that tells a compelling story about the data. This often involves:

1. Choosing the right plot
2. Simplifying the visualization
3. Labeling axes

T-test:

Jab ap ko population ky SD pta ho or sample size above 30 ho tab T-test or Z-test use kia jata a jab ap ko population ka SD pta na ho tab T-test use karty ha

Use:

1. T-test tab use hota a jab less than 30 sample ho
2. 2 means ko jab apass ma compair karna hota a

Types of T-Tests:

1. one sample T-test
2. independent sample T-test
3. repeated measure sample T-test
4. unequal variance sample T-test

Z-test:

The Z-test is used when the sample size is greater than 30 or when the population standard deviation is known.

Z-test is ideal for large samples because the sampling distribution of the sample mean approximates a normal distribution due to the Central Limit Theorem.

Parametric Test :--> more reliable result: First we have meet the assumptions

(2, 5, 16, 18, 20) (25, 38, 52, 100, 120) Not equal

step1: Normality test:

1: Shapiro-wilk test -->specific (reliable)

2: Kolmogorov-Smirnov test -->general(less reliable)

step2: Homogeneity test: the variance of the variable in data are equal

1: Levene's test

step3: Purpose: know the purpose of your research question

1: comparison (difference)

2: relationship (connection)

step4: Data type::know the type of data you are working with

1: categorical-->qualitative and no numerical (e.g. character, factors)

2: numerical-->quantitative and numerical (e.g. numerical variable, int and float)

step5: statistical test::choose a statistical test from three main families

1: chi-squared-->purpose: comparison Data: categorical only

2: t-test/anova-->purpose: comparison Data: categorical and numerical

3: correlation-->purpose: relationship Data: numerical only

Non-Parametric Test :--> Less reliable result: calculates the rank of data: NO need to meet the assumptions

(1, 2, 3, 4, 5) (1, 2, 3, 4, 5) equal based on the rankings

1: Chi-squared

2: T-test/anova

1. one-sample Wilcoxon signed rank test
2. (unpaid)mann whitney's U-test [(paid) Wilcoxon]
3. (anova) kruskal-waallis test)

3: Correlation---> (spearman's correlation) and (Kendall's tau) regression

Variable:

Asy value jo change hoty jay

Independent:

1. Independent Variable:

A variable that is **controlled or changed** in an experiment.

Example: Type of food in a diet experiment.

2. Dependent Variable:

A variable that is **measured or observed** as it changes in response to the independent variable.

Example: Body weight measured in a diet experiment.

2. Inferential Statistics:

Inferential statistics use data from a sample to make inferences or predictions about a larger population. It involves analyzing the data and applying probability to draw conclusions.

Sample k base sy population py result dalna

1. Conclusion
2. Predict
3. Hypotheses test

Important Concepts in statistics:

1. **Depend vs independent variable**--> Relationship between controlled (IV) and observed (DV) variables.

2. **Hypothesis testing-->** Validating assumptions with statistical tests.
3. **Confidence interval-->** a range to estimate population parameters.
4. **Multiple comparison-->** Adjusts for testing multiple groups.
5. **Regression-->** Predicts relationships between variables.
6. **ANONA-->** Tests differences across three or more groups.

Hypothesis:

Ho: Null hypothesis

H1: Alternate hypothesis

Steps of Hypothesis:

1. Ho: Null hypothesis
2. H1: Alternate hypothesis
3. Data collection to test the hypothesis
4. Statistical Analysis
5. Results

Ho: Null hypothesis: male = female

Drink use karny sy koi farg nahi pra

When we reject H0 then use H1

H1: Alternate hypothesis:male != female

Drink use karny sy farg pra

H1 = fsd < lhr

H2 = fsd > lhr

After test jab P ki value less than 0.05 then H0 reject

After test jab P ki value greater than 0.05 then H0 failed to reject

Confidence Interval:

Example: maan lijiye aapne lahore ke ek college mein students ka average test Score calculate kiya hai. aapka sample mean 70 hai aur aapne 95% confidence interval calculate kiy hai jo 68 se 72 ke beech hai. iska matlab ya hai ke app 95% sure hain ke poore college ke students ka average score is range mein hoga.

Formula:

CI = mean -+ (Z x SE)

Z is the Z-score corresponding to the desired confidence level (for

Example. 1.96 for a 95% confidence interval)

SE is the standard error of the sample mean

SE = SD divided by square root of sample number