STATISTICS

Statistics should be like a story telling

Statistics is the discipline that concerns the collection, organization, analysis, interpretation, and presentation of data. It helps in transforming data into information to make informed decisions.

Key Processes in Statistics:

- **Data Collection**: Gathering raw data from various sources.
- **Organizing the Data**: Structuring and cleaning data for analysis.
- **Analysing the Data**: Using statistical methods to understand patterns and trends.
- **Interpreting the Data**: Drawing conclusions from the analysis.
- **Presenting the Data**: Communicating the findings effectively.

Statistics should be like storytelling, where data is used to convey meaningful insights.

Types of Data:

1. **Quantitative (Numerical) Data**:
    - **Discrete**: Countable values (e.g., 1, 2, 3).
    - **Continuous**: Any value within a range (e.g., 1.1, 1.2, 1.3).
    - *Examples*: Temperature, Height.
2. **Qualitative (Categorical) Data**:
    - Data that describes categories or groups.
    - *Examples*: Colors, Brands.

Levels of Measurement:

1. **Nominal Level**:
    - Categories without a specific order.
    - *Examples*: Gender, Marital Status.
2. **Ordinal Level**:
    - Categories with a meaningful order but not evenly spaced.
    - *Examples*: Movie ratings (e.g., flop, average, hit, superhit), Satisfaction levels (e.g., good, bad, ugly).
3. **Interval Level**:
    - Numeric scales with equal intervals but no true zero.
    - Can have negative values.
    - *Examples*: Temperature in Celsius or Fahrenheit.
4. **Ratio Level**:
    - Numeric scales with equal intervals and a true zero.
    - Cannot have negative values.
    - *Examples*: Weight, Height, Age.

Temperature is which level of data?

If you convert temperature into different units then its not equal

---

Population and Sample

Population:

- A large group of data which we are experimenting on.

Sample:

- A small (subset) group of the population we will work on to draw conclusions (inferences).

Inferential Statistics:

- Making conclusions based on evidence and reasoning from a sample to the broader population.

Descriptive Statistics

Descriptive Statistics:

- Works on the population to summarize and describe data.

Frequency Table:

- Used for categorical data to show the number of occurrences (frequency) for each category.

Example:

Class	Class Frequency
Male	4
Female	4

Box Plot:

- y-axis: Frequency (number of values)
- x-axis: Class (categories)

Tabular Representation: Frequency Table

Graphical Representation: Bar plot/bar graph

- x-axis: Class (categorical)
- y-axis: Class Frequency (numerical)

Relative Frequency:

- Often represented in a pie chart.

Example:

Class	Class Frequency	Relative Frequency
Boys	30	37.5%
Girls	50	62.5%
Total	80	100%

Calculating Percentages:

$$\frac{value\ of\ the\ category * 100}{Total\ no\ of\ values}$$

Example Calculation:

- Total = 80 (100%)
- Female = 50

$$\frac{50 * 100}{80} = \frac{5 * 25}{2} = \frac{125}{2} = 62.5\%$$

- Male = 30

$$\frac{30 * 100}{80} = \frac{75}{2} = 37.5\%$$

Related Questions

What is the name of the tabular representation?

- Frequency Table

What are the column names in the frequency table?

- Class, Class Frequency, Relative Frequency

Example: Boys & Girls Marks

Raw Data: 3, 7, 9, 11, 15, 18, 20, 23, 25, 30

Observations on data:

- We divide data into intervals and check how many observations fall in each interval.

Class Interval	Class Interval Frequency
0-5	1
5-10	2
10-15	1
15-20	3
20-25	2
25-30	1

Graphical Representation: Histogram

- x-axis: Numerical (Class Interval)
- y-axis: Numerical (Class Interval Frequency)

Categorical vs. Numerical Data

Categorical Data:

- **Table**: Frequency Table (Class vs. Class Frequency)
- **Graph**: Bar graph (x-axis: Class, y-axis: Class Frequency)
- **Pie Chart**: Relative Frequency

Numerical Data:

- **Table**: Frequency Distribution Table (Class Interval vs. Class Interval Frequency)
- **Graph**: Histogram (x-axis: Class Interval, y-axis: Class Interval Frequency)
- **Data Distribution Plot**

Variables, Columns, Features

- Age, Gender, Income → Variables, Columns, Features

Univariate Analysis:

- Analysing a single variable (e.g., Age).

Bivariate Analysis:

- Analysing two variables (e.g., Age & Income).

Multivariate Analysis:

- Analysing more than two variables (e.g., Age, Income, and Gender).

Feature Selection and Mutually Exclusive Events

Feature Selection:

- Choosing relevant features for analysis to build better ML models.

Mutually Exclusive Events:

- Events that cannot happen simultaneously (e.g., a coin toss resulting in either heads or tails, but not both).

Raw Data to Histogram

1. **Determine Min and Max Values**.
2. **Count Number of Data Observations**.
3. **Decide on the Number of Class Intervals**.

Use the formula 2^K determine the appropriate number of class intervals, where k is adjusted to ensure the total number of observations is appropriately categorized.

To see where the number of data observations land in 2^K formula

4. **what is Interval width**

$$\frac{Max - Min}{Class\ intervals}$$

1. **Define Population and Sample.**

- **Population**: A large set of data we experiment on.
- **Sample**: A small group (subset) of the population from which we draw conclusions.

2. **What is the difference between Descriptive and Inferential Statistics?**

- **Descriptive Statistics**: Involves measuring and analyzing data to summarize and describe its main features.
- **Inferential Statistics**: Involves drawing conclusions from a sample group of data based on reasoning and calculations, to make inferences about the larger population.

3. **Explain what a Frequency Table is and give an example.**

- **Frequency Table**: Represents categorical data in a table format, showing the number of occurrences (frequency) for each category.
- **Example**:

Class	Class Frequency
Boys	30
Girls	50

4. **How do you calculate Relative Frequency and what graphical representation is best suited for it?**

- **Relative Frequency**:

Graphical Representation: Best suited for a pie chart.

5. **Describe the key components of a Box Plot.**

- **Components**:
    - **X-axis**: Represents the categories or classes.
    - **Y-axis**: Represents the values (often used for frequency in certain contexts, but primarily for showing distribution and spread of numerical data).

6. **What is the difference between a Bar Graph and a Histogram?**

- **Bar Graph**: Used for categorical data, where each bar represents a category.
- **Histogram**: Used for numerical data, where each bar represents a class interval.

7. **Give an example of univariate, bivariate, and multivariate analysis.**

- **Univariate Analysis**: Analysing a single variable, e.g., age.
- **Bivariate Analysis**: Analysing the relationship between two variables, e.g., age and income.
- **Multivariate Analysis**: Analysing the relationship among more than two variables, e.g., age, income, and gender.

8. **Explain the concept of mutually exclusive events with an example.**

- **Mutually Exclusive Events**: Events that cannot happen simultaneously.
- **Example**: Tossing a coin results in either heads or tails, but not both.

9. **How do you determine the number of class intervals for a histogram?**

- Using the 2k2^k2k rule, where k is adjusted to ensure the total number of observations is appropriately categorized.

10. **What are the steps involved in creating a histogram from raw data?**

1. Determine the minimum and maximum values.
2. Calculate the number of class intervals.
3. Calculate the interval width.
4. Create intervals and count the frequency of observations in each interval.
5. Plot the histogram with intervals on the x-axis and frequencies on the y-axis.

Describing Data:

Data analysis can be approached through two main types:

1. **Measure of Central Tendency (Centre Point Analysis):**
    - **Mean**: The average of the data.
    - **Median**: The middle value when data is sorted.
    - **Mode**: The most frequently occurring value in the dataset.
2. **Data Flow Analysis (Data Dispersion):**
    - **Range**: The difference between the maximum and minimum values.
    - **Mean Deviation**: The average deviation of each observation from the mean.
    - **Absolute Mean Deviation**: The average of the absolute differences between each data point and the mean.
    - **Variance**: The average of the squared differences from the mean.
    - **Standard Deviation**: The square root of the variance.

Measure of Central Tendency (Centre Point Analysis):

1. **Mean (Average):**
    - Formula:

$$\bar{x} = \frac{x_1 + x_2 + x_3 + x_4 \ldots x_n}{n}$$

$$sample\ mean = \sum_{i=1}^{n} x_i$$

$$Poupulation\ \ mean = \sum_{i=1}^{n} x_i$$

2. **Median (Middle Value):**
    - Observations must be in sorted order.
    - Median = 50th percentile of the data.

Mean vs. Median:

- Mean is sensitive to outliers, while the median is robust against them.

3. **Mode:**

- The most frequent value in the dataset.

Key Insights:

- **Mean**: Represents the average value.
- **Median**: Represents the middle value (50th percentile).
- **Mode**: Represents the most repeated value.
- **Distribution Characteristics**:
  - **Right Skew (Positive Skew)**: Mean > Median > Mode.
  - **Left Skew (Negative Skew)**: Mode > Median > Mean.
  - **No Skew (Normal Distribution)**: Mean = Median = Mode.
- **Asymptotes in Distribution Plots**: The edges of a distribution plot that never touch the real axis.

Data Flow Analysis: Data Dispersion/Data Variance:

1. **Range**:
   - Formula: Range=Max−Min
   - It focuses on the extreme values, ignoring the middle values.
2. **Mean Deviation**:

   It considers how much each observation deviates from the mean

   Example:

   1 2 3 4 5

   Mean = 3

$$\bar{x} = 3$$

$$x_1 = 1, x_2 = 2, x_3 = 3, x_4 = 4, x_5 = 5$$

$$1st\ observtaion\ from\ mean\ point\ x_1 - \bar{x} = 1 - 3 = -2$$

$$2nd\ observtaion\ from\ mean\ point\ x_2 - \bar{x} = 2 - 3 = -1$$

$$3rd\ observtaion\ from\ mean\ point\ x_3 - \bar{x} = 3 - 3 = 0$$

$$4th\ observtaion\ from\ mean\ point\ x_4 - \bar{x} = 4 - 3 = 1$$

$$5th\ observtaion\ from\ mean\ point\ x_5 - \bar{x} = 5 - 3 = 2$$

What is the total Deviation

$$-2 - 1 + 0 + 1 + 2 = 0$$

$$(x_1 - \bar{x}) + (x_2 - \bar{x}) + (x_3 - \bar{x}) \dots\dots + (x_n - \bar{x})$$

$$Mean\ divation = \frac{1}{n}\sum_{i=1}^{n}(\boldsymbol{x_i} - \bar{\boldsymbol{x}})$$

$$\boldsymbol{\frac{1}{n}quantifying\ the\ result}$$

Drawback: When deviations are summed, the result might be zero, failing to provide meaningful information.

3. **Absolute Mean Deviation**:
    - Uses absolute values to avoid cancellation of deviations.

      Example:

      1 2 3 4 5

      Mean = 3

$$\bar{x} = 3$$

$$x_1 = 1, x_2 = 2, x_3 = 3, x_4 = 4, x_5 = 5$$

$$1st\ observtaion\ from\ mean\ point\ |x_1 - \bar{x}| = 1 - 3 = |-2| = 2$$

$$2nd\ observtaion\ from\ mean\ point\ |x_2 - \bar{x}| = 2 - 3 = |-1| = 1$$

$$3rd\ observtaion\ from\ mean\ point\ |x_3 - \bar{x}| = 3 - 3 = |0| = 0$$

$$4th\ observtaion\ from\ mean\ point\ |x_4 - \bar{x}| = |4 - 3| = |1| = 1$$

$$5th\ observtaion\ from\ mean\ point\ |x_5 - \bar{x}| = |5 - 3| = |2| = 2$$

What is the total Absolute Mean Deviation

$$2 + 1 + 0 + 1 + 2 = 4$$

$$|(x_1 - \bar{x})| + |(x_2 - \bar{x})| + |(x_3 - \bar{x})| \dots \dots + |(x_n - \bar{x})|$$

$$Absolute\ Mean\ divation = \frac{1}{n}\sum_{i=1}^{n} |\boldsymbol{x_i} - \bar{\boldsymbol{x}}|$$

Drawback: Absolute values lead to discontinuities at zero, making some mathematical operations (like differentiation) problematic.

4. **Variance**:
    - Measures the spread of data points around the mean by squaring deviations.

      Example:

      1 2 3 4 5

      Mean = 3

$$\bar{x} = 3$$

$$x_1 = 1, x_2 = 2, x_3 = 3, x_4 = 4, x_5 = 5$$

$$1st\ observtaion\ from\ mean\ point\ (x_1 - \bar{x})^2 = (1 - 3)^2 = 4$$

$$2nd\ observtaion\ from\ mean\ point\ (x_2 - \bar{x})^2 = (2 - 3)^2 = 1$$

$$3rd\ observtaion\ from\ mean\ point (x_3 - \bar{x})^2 = (3 - 3)^2 = 0$$

$$4th\ observtaion\ from\ mean\ point\ (x_4 - \bar{x})^2 = (4 - 3)^2 = 1$$

$$5th\ observtaion\ from\ mean\ point\ (x_5 - \bar{x})^2 = (5 - 3)^2 = 4$$

$$4 + 1 + 0 + 1 + 4 = 10$$

$$(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + (x_3 - \bar{x})^2 \dots \dots + (x_n - \bar{x})^2$$

$$variance\ \sigma = \frac{1}{n}\sum_{i=1}^{n}(\boldsymbol{x_i} - \overline{\boldsymbol{x}})^2$$

Drawback: Units of variance are squared, which might complicate interpretation.

5. **Standard Deviation**:

$$Standard\ Deviation\ \sigma^2 = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(\boldsymbol{x_i} - \overline{\boldsymbol{x}})^2}$$

   - Provides a measure of dispersion in the same units as the original data.

Percentiles and Quartiles:

- **Percentile**: Represents the relative standing of a value within a dataset.

Formula: Percentile Position

$$(N+1) * \frac{p}{100}$$

   - Example: 50th percentile corresponds to the median.

- **Quartile**: Divides data into four equal parts.
   - 1st Quartile (25th percentile)
   - 2nd Quartile (50th percentile, or Median)
   - 3rd Quartile (75th percentile)

Questions on Central Tendency:

1. **What is the difference between mean, median, and mode?**

   Mean: Average of all the observations of a data set

   Median: Centre point of a data set

   Mode: most repeated value in a data set

2. **Explain why the median is more robust than the mean in the presence of outliers.**

   In Median all the observations are get sorted and the middle values are going to consider, so the outliers are never getting in the dataset

   Where mean gets affect if any large or small observations are in a data set

3. **If a dataset is positively skewed, how will the mean, median, and mode be related?**

Mode>Median>Mean

4. **How do you calculate the mean of a dataset? Provide the formula and explain each component.**

   Sum of observations / Total no of observations

5. **Describe a situation where the mode might be more informative than the mean or median.**

Questions on Data Dispersion:

6. **What is the range, and what does it tell you about a dataset?**

   Range is data dispersion measure

   It tells about the range of observations in the data set

7. **Why might the range be insufficient as the only measure of data dispersion?**
8. **How is mean deviation calculated, and what is its drawback?**
9. **Explain the concept of absolute mean deviation and how it addresses the drawbacks of mean deviation.**
10. **What is variance, and why do we square the deviations when calculating it?**
11. **How is standard deviation related to variance?**
12. **Why is standard deviation preferred over variance when interpreting the spread of data?**

Questions on Percentiles and Quartiles:

13. **What is a percentile, and how would you calculate the position of the 75th percentile in a dataset?**
14. **How do quartiles divide a dataset, and what does each quartile represent?**
15. **If you have a dataset with 10 observations, how would you determine the observation number corresponding to the 25th percentile?**

Application Questions:

16. **Given a dataset: 2, 4, 6, 8, 10, calculate the mean, median, and mode.**

    Mean = 6

    Median = 6

    Mode = No mode

17. **For the same dataset above, calculate the range, mean deviation, and standard deviation.**

    Range = 2-10 = 8

18. **If a dataset has the following values: 5, 12, 18, 23, 23, 29, 34, which measure of central tendency would be most affected by an outlier of 100 added to the dataset? Explain why.**

Mean will be most affected by the outlier because the mean will be greater than the max value in the data set

19. **Explain how you would identify skewness in a dataset based on the relationship between the mean, median, and mode.**
20. **Given a small dataset, how would you calculate the interquartile range (IQR) and what does it represent?**

Outlier Analysis:

Observations that impact the overall data is called outliers

Outliers might exit in

Q3 – max

Min – Q1

Outliers are more than Q3 and less than Q1

With the help of IQR Interquartile range

a measure of statistical dispersion, which is the spread of the data a difference between 25 and 75 percentiles

$$outlier = Q_3 + K * IQR$$

$$Q_3 + K * (Q_3 - Q_1)$$

K= 1.5 = Mild outlier

$Q_3 + 1.5 * IQR \ \& \ Q_1 - 1.5 * IQR$

K= 3 = Huge outlier

$Q_3 + 3 * IQR \ \& \ Q_1 - 3 * IQR$

How much far you are travelling

How to deal with outliers?

No specific rule

Remove the outlier, but when to remove (intuition process)

When outliers are less assuming 2% then remove the outliers then you have 98% of data left

Data Augmentation:

Create multiple observations from single observations

Variance – Co-Variance:

Age	Income
30	50k
31	55k
32	60k

How data is varying

In age column separately → univariant

In income column separately → univariant

How the data is varying in Income with respect to Age or data varying in Age with respect to Income

Bi-variant

Data is Varying in univariant → variance

Data is varying in one variance with respect to another variant → co-variance

Co-Variance:

We are trying to understand how a data is varying in one variant with respect to another variant

Co-Variance Matrix:

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

$$Co - Variance\ (x, y) = \frac{1}{n}\sum_{i=1}^{n}(\bar{\boldsymbol{x}} - \boldsymbol{x_i})\,(\bar{\boldsymbol{y}} - \boldsymbol{y_i})$$

What is Statistics

Data Types of Statistics

Types of Statistics

Levels of Data

Nominal-ordinal-interval-ratio

Population vs Sample

Frequency Table

Bar chart

Histogram Distribution plot

Central Tendency

Mean-median-mode

Mean vs Median

Mode meaning with respect to distribution plot

Positive skew – No skew – Negative skew

Outlier Meaning

Data dispersion

Range - Mean Deviation – Absolute Mean Deviation – Variance – Standard Deviation