ASSIGNMENT CASE STUDY

3. Interquartile Range (IQR):

● Find the IQR for the given dataset and explain its significance.

The interquartile range (IQR) is a measure of statistical dispersion that describes the spread of the middle 50% of a dataset. It is calculated as the difference between the third quartile (Q3) and the first quartile (Q1).

Significance of the IQR**:**

1. Measures Spread: The IQR describes the spread of the middle 50% of the data, making it a robust measure of variability.
2. Resistant to Outliers: Unlike the range, the IQR is not affected by extreme values (outliers), making it a reliable measure for skewed datasets.
3. Identifies Outliers: The IQR is used to detect outliers using the **1**.5 × IQR rule:
    - Any data point below Q1−1.5×IQR$Q1-1.5\times IQR$ or above Q3+1.5×IQR$Q3+1.5\times IQR$ is considered an outlier.
4. Box Plots: The IQR is a key component of box plots, which visually summarize the distribution of data

5. Finding Outliers Using Quartiles:

● Compute the Lower Bound and Upper Bound.

● Identify any outliers in the dataset.

Answer:

- The smallest value is 10, which is greater than the lower bound (5).
- The largest value is 24, which is less than the upper bound (29).

There are no outliers in this dataset. All values lie within the range [5,29].

7. Z-score Standardization:

● Compute the Z-scores for each value in the dataset and explain its significance in

data standardization.

Significance of z-scores in Data Standardization:

1. Standardizes Data:
    - Z-scores transform data into a standard scale with a mean of 0 and a standard deviation of 1.
    - This allows for easy comparison of values from different datasets or distributions.
2. Identifies Outliers:
    - Z-scores help identify outliers. A data point with a z-score greater than 3 or less than -3 is often considered an outlier.
3. Normalizes Data:
    - Z-scores are used to normalize data, which is essential for many machine learning algorithms (e.g., linear regression, PCA) that perform better when features are on the same scale.
4. Compares Different Distributions:
    - Z-scores allow you to compare values from different datasets or distributions, even if they have different units or scales.

12. Why Inferential Statistics?

Inferential statistical analysis involves drawing conclusions about a population by testing hypothesis and generating estimates. It operates under the assumption that the observed dataset is a sample taken from a larger population. Inferential statistics is distinct from descriptive statistics, which focuses on summarizing and describing the characteristics of the dataset itself without making broader inferences.

● Explain the difference between Correlation and Causation with an example.

Correlation is the analysis of association between two or more variables. Two variables are said to be correlated if the change in one variable results in a corresponding change in the other. For example; when the price of a commodity rises the supply for the commodity also raises. On the other hand, when price falls the supply also falls. Hence price and supply are correlated.

When two variables are correlated there need not be cause and effect relationship. It is possible that a high degree of correlation between the variables may be due to same cause affecting each variable. For example; a situation where ice cream sales increases when drowning cases increases. These two variables may seem correlated but in reality they are influenced by a third variable, hot weather. Hot weather leads to more people swimming (increasing drowning risk) and more people buying ice cream. Here, there is no direct causation between ice cream sales and drowning incidents.

13. Population vs. Sample:

A group of objects under study is called population. A population containing finite number of objects are called finite population. For example; the students in a school. Population having infinite number of objects is called infinite population. for example; population of stars in the sky.

A finite subset of population selected from it with objective of investigating its properties is called a sample. A sample is a representative part of the population. for example; we want to study the life span of motors produced by a company, we select some motors and study their life span. This selected number is called the sample.

● Why do we need sampling? Provide a real-world example.

The basic objective of sampling is to draw inference about the population. It is a tool which helps to know the characteristics of the population. for example; to examine the quality of rice in a ricebag a handful of rice is taken as sample to examine.

The process of sampling involves 3 elements:

1.selection of the sample

2. collecting information

3.making an inference about the population.

14. Hypothesis Testing Concepts:

● Define Null Hypothesis, Alternate Hypothesis, Significance Level (α), and P-value.

Null Hypothesis and Alternative Hypothesis:

A null hypothesis can be defined as a statistical hypothesis which is stated for the purpose of possible acceptance. Null Hypothesis is the original hypothesis, any other hypothesis than null is considered as the Alternate Hypothesis. So when Null is rejected we accept the other hypothesis known as the alterative hypothesis.

Significance Level(α):

Confidence with which a null hypothesis is rejected or accepted depends on significance level denoted by α. The level of significance is usually determined before conducting the test of hypothesis.

P-value:

The p-value is the probability of observing the results (or more extreme results) of your study, assuming the null hypothesis is true.It helps you decide whether to reject or fail to reject the null hypothesis.

A low p-value (typically ≤ 0.05) suggests that the observed data is unlikely under the null hypothesis, indicating that your results are statistically significant.

20. Summary and Insights:

● Summarize the key takeaways from the analysis performed above and describe

Descriptive analysis performed for section 1:

Summarizes and describes the main features of a dataset and provides a clear and concise understanding of the data. Key Tools used are :

- Measures of central tendency: Mean, Median, Mode.
- Measures of variability: Range, Variance, Standard Deviation, Interquartile Range (IQR).
- Data visualization: Histograms, Box plots, Scatter plots.

Inferential statistics performed for section 2:

Draws conclusions or makes predictions about a larger population based on a sample of data. Tests hypotheses and estimates population parameters. Key Tools used are:

- Hypothesis testing: t-tests, z-tests

how descriptive and inferential statistics can be used in real-world data analysis.

Descriptive Analysis Applications:

Business:

- Summarizing sales data to identify trends (e.g., monthly revenue, best-selling products).
- Creating dashboards to visualize key performance indicators (KPIs).

Healthcare:

- Describing patient demographics (e.g., average age, gender distribution).
- Summarizing clinical data (e.g., average blood pressure, cholesterol levels).

Education:

- Reporting average test scores and grade distributions.

- Visualizing student performance across different subjects.

Finance:

- Summarizing stock price movements over time.
- Calculating average returns or volatility for investment portfolios.

Marketing:

- Analysing customer demographics and purchase behaviour.
- Visualizing website traffic and engagement metrics.

Inferential Analysis Applications:

Business:

- Testing whether a new marketing strategy increases sales.
- Estimating customer satisfaction levels across a large population based on survey data.

Healthcare:

- Testing the effectiveness of a new drug compared to a placebo.
- Estimating the prevalence of a disease in a population based on sample data.

Education:

- Determining whether a new teaching method improves student performance.
- Estimating the average literacy rate in a country based on sample data.

Finance:

- Predicting future stock prices using regression analysis.
- Testing whether a new investment strategy yields higher returns.

Social Sciences:

- Analysing survey data to understand public opinion on social issues.
- Testing hypotheses about the impact of policy changes (e.g., minimum wage increases).

Manufacturing:

- Testing whether a new production process reduces defect rates.
- Estimating the average lifespan of a product based on sample testing.