Experiment No 5

Title: Create a data set and do statistical analysis on the data using R.

Objective:

The objective of this practical is to learn how to:

1. **Create a dataset** in R.
2. **Perform basic statistical analysis** like calculating the mean, median, and standard deviation.
3. **Visualize the data** using basic plots.
4. **Draw basic conclusions** from the analysis, such as identifying trends or patterns.

Theory:

1. Dataset Creation

A dataset in R is typically structured as a **data frame**, where:

- **Variables** are columns (e.g., age, height, weight).
- **Observations** are rows (each representing a single unit of data like a person or object).

R's `data.frame()` function is a fundamental tool for creating datasets, allowing us to input or import structured data for analysis. Each variable in the dataset can represent different data types (numeric, categorical, etc.), and these variables are what we analyze statistically.

2. Statistical Analysis

This step focuses on extracting meaningful summaries from the dataset. Some essential statistical concepts are:

- **Mean**: This is the **arithmetic average** and represents the central tendency of a variable.

  $$\text{Mean} = \sum x_i / n$$

  Where x_i are the individual data points, and n is the number of data points.

- **Median**: The **median** is the middle value that separates the dataset into two equal halves. It's less sensitive to outliers than the mean, making it useful when dealing with skewed data.

- **Standard Deviation (SD)**: SD measures the **spread** or **dispersion** of the data points around the mean. A low SD indicates that the data points are close to the mean, while a high SD shows that the data points are spread out over a wider range.

$$\mathrm{SD} = \sqrt{\frac{1}{n}\sum (x_i - \mu)^2}$$

  Where xi is each individual data point, and μ is the mean.

- **Variance**: The square of the standard deviation, used to measure the degree of dispersion in the dataset.

These statistical measures give us an idea of how the data is distributed and provide insight into its central tendencies and variability.

Data Visualization

Data visualization is crucial for identifying patterns, trends, or anomalies in the dataset. The following visual tools help us understand the data better:

- **Histogram**: A histogram displays the **frequency distribution** of a single variable. It shows how often each range of values occurs, providing a sense of the data's distribution (normal, skewed, etc.).
- **Scatter Plot**: This graph plots two variables against each other, providing insights into their relationship or correlation. For example, plotting height vs. weight might show a positive correlation (as height increases, weight tends to increase).
- **Box Plot**: A box plot (or box-and-whisker plot) is useful for identifying **outliers** and the **spread** of the data. It displays the minimum, first quartile (Q1), median, third quartile (Q3), and maximum. The interquartile range (IQR = Q3 - Q1) helps understand the spread of the middle 50% of the data.

4. Drawing Conclusions

Once we have the statistical results and visual representations, we can derive **insights** from the data. For example:

- **Outliers**: Boxplots help identify extreme values that might skew the results.
- **Distribution**: Histograms and summary statistics (mean, median) tell us whether the data is normally distributed or skewed.

- **Relationships**: Scatter plots help identify correlations between variables (e.g., a positive or negative trend between height and weight).

In real-world applications, these techniques are essential in fields like economics, healthcare, engineering, and any domain where decision-making is based on data analysis.

Problem Statement:

Health Data Analysis:

- You have a dataset of individuals with information on their age, height, weight, and BMI (Body Mass Index). Analyze the data to:
    - Calculate the mean, median, and standard deviation of each variable.
    - Determine if there is a correlation between height and weight.
    - Visualize the distribution of age using a histogram and check for any outliers using a boxplot.

Sales Performance of Products:

- A company provides sales data for its products, including variables like product price, units sold, and profit. Your task is to:
    - Calculate the average sales, median price, and total profit.
    - Analyze the variability in sales across products using standard deviation.
    - Visualize the sales and profit data using scatter plots and draw conclusions about the relationship between price and profit.

Student Scores Analysis:

- You have a dataset containing the exam scores of students in three subjects: Mathematics, Science, and English. Analyze the scores to:
    - Compute the mean, median, and mode for each subject.
    - Identify the subject with the highest variability in scores (using standard deviation).
    - Create a boxplot for each subject and analyze the spread and outliers.
    - Visualize the relationship between Mathematics and Science scores using a scatter plot and determine if there is a trend.

Outcome:

The students will understand the use of statistical analysis for drawing meaningful conclusions from data.

Online Reference Websites:

- https://www.geeksforgeeks.org/r-statistics/
- https://www.programiz.com/r/dataset
- https://www.geeksforgeeks.org/how-to-calculate-correlation-between-multiple-variables-in-r/

Viva Questions:

1.	How you can produce co-relations and covariances?
2.	What is difference between matrix and dataframes?
3.	Which function in R language is used to find out whether the means of 2 groups are equal to each other or not?
4.	How can you add datasets in R?
5.	How do you calculate descriptive statistics (mean, median, variance) in R?
6.	How do you create and interpret a correlation matrix in R?

Conclusion:

The students learn power of statistical analysis for drawing meaningful conclusions from data.