### **Mini Project: Descriptive and Inferential Statistics**

**Problem Statement:**This project is designed to develop foundational understanding in **Descriptive and Inferential Statistics**. The assignment covers key concepts such as **Measures of Central Tendency** (Mean, Median, Mode), **Measures of Dispersion** (Range, Variance, Standard Deviation), and **key inferential methods** like **Confidence Intervals** and **Hypothesis Testing**. By completing this project, students will gain both theoretical knowledge and practical skills in interpreting and analyzing data.

**Guidelines for Students:**

1. **Foundational Knowledge:**
   * Understand the key statistical definitions, data types, and the concept of population vs sample.
   * Learn how to calculate measures of central tendency and dispersion.
   * Understand how confidence intervals and hypothesis testing are used to infer characteristics about a population from sample data.
2. **Hands-on Learning:**
   * Manually calculate measures such as mean, median, mode, variance, and standard deviation.
   * Interpret statistical results, including the concept of confidence intervals and hypothesis testing.
3. **Model Evaluation:**
   * Practice interpreting the results of statistical tests and applying them to make data-driven decisions.

### 

### 

### 

### **Step-by-Step Project Outline:**

#### **Q1. Key Statistical Definitions**

**Objective:** Understand foundational statistical terms.

**Problem:** Write short definitions (2-3 lines each) for the following:

#### **a) Population and Sample**

* Population: The complete set of individuals or items that are of interest in a statistical study.
* Sample: A subset of the population that is selected for analysis, used to make inferences about the population.

#### **b) Descriptive Statistics and Inferential Statistics**

* Descriptive Statistics: Involves summarizing and describing the features of a dataset (e.g., mean, median, mode).
* Inferential Statistics: Uses sample data to make generalizations or predictions about a population, typically involving hypothesis testing and confidence intervals.

#### **c) Parameter and Statistic**

* Parameter: A numerical characteristic or measure of a population (e.g., population mean).
* Statistic: A numerical characteristic or measure derived from a sample (e.g., sample mean).

#### **d) Qualitative and Quantitative Data**

* Qualitative Data: Data that describes categories or qualities (e.g., colors, types of animals).
* Quantitative Data: Data that represents amounts or counts, which can be measured numerically (e.g., height, weight, age).

#### **Q2. Measures of Central Tendency - Definitions**

**Objective:** Learn basic concepts of data centering.

**Problem:** Define the following terms with one example each:

#### **a) Mean**

* Mean: The sum of all values in a dataset divided by the number of values.  
   *Example*: For the dataset [2, 3, 5], the mean is (2+3+5)/3 = 3.33.

#### **b) Median**

* Median: The middle value of a dataset when it is arranged in ascending order. If the dataset has an even number of values, it is the average of the two middle values.  
   *Example*: For the dataset [1, 3, 5], the median is 3.

#### **c) Mode**

* Mode: The value that occurs most frequently in a dataset.  
   *Example*: For the dataset [1, 2, 2, 3, 4], the mode is 2.

#### **Q3. Manual Calculation of Mean, Median, and Mode**

**Objective:** Apply manual formulas to real data.

**Problem:** Given the dataset:  
 12, 18, 14, 16, 18, 20, 18, 15, 12, 18, 14, 16, 18, 20, 18, 15

Calculate:

#### **a) Mean**

* Sum of values = 12 + 18 + 14 + 16 + 18 + 20 + 18 + 15 + 12 + 18 + 14 + 16 + 18 + 20 + 18 + 15 = 240
* Number of values = 16
* Mean = 240 / 16 = 15

#### **b) Median**

* Arrange in ascending order: 12, 12, 14, 14, 15, 15, 16, 16, 18, 18, 18, 18, 18, 20, 20
* Median = Middle value, which is the average of the 8th and 9th values:  
   Median = (16 + 18) / 2 = 17

#### **c) Mode**

* The most frequent value is 18, so Mode = 18

#### **Q4. Levels of Measurement**

**Objective:** Understand classification of data types.

**Problem:** Define and give one example for each level of measurement:

#### **a) Nominal**

* Definition: Data categorized into distinct groups without a natural order or ranking.  
   *Example*: Gender (Male, Female, Other).

#### **b) Ordinal**

* Definition: Data with a defined order, but the intervals between values are not meaningful.  
   *Example*: Likert scale responses (Strongly Agree, Agree, Neutral, Disagree, Strongly Disagree).

#### **c) Interval**

* Definition: Data with equal intervals between values, but no true zero.  
   *Example*: Temperature in Celsius (0°C does not mean 'no temperature').

#### **d) Ratio**

* Definition: Data with equal intervals and a true zero point.  
   *Example*: Height (0 cm means no height).

#### **Q5. Variance and Standard Deviation - Theory**

**Objective:** Understand spread/variability in data.

**Problem:**

#### **a) Variance and Standard Deviation**

* Variance: A measure of the spread of a dataset, calculated as the average of the squared differences from the mean.
* Standard Deviation: The square root of the variance, indicating how much individual data points deviate from the mean.

#### **b) Why Standard Deviation is More Interpretable than Variance**

* Standard deviation is in the same units as the original data, making it easier to interpret, whereas variance is in squared units, which can be harder to relate to the data.

#### **Q6. Manual Calculation - Variance and Standard Deviation**

**Objective:** Practice computing data spread.

**Problem:** Given the data:  
 8, 10, 12, 14, 16, 10, 12, 14, 16

Calculate:

#### **a) Sample Variance**

* Mean = (8 + 10 + 12 + 14 + 16 + 10 + 12 + 14 + 16) / 9 = 12
* Variance = Sum of squared differences from the mean, divided by (n-1)  
   Variance = [(8-12)² + (10-12)² + (12-12)² + (14-12)² + (16-12)² + (10-12)² + (12-12)² + (14-12)² + (16-12)²] / 8  
   Variance = 16.5

#### **b) Sample Standard Deviation**

* Standard Deviation = √16.5 = 4.06

#### **Q7. Range and Interquartile Range (IQR)**

**Objective:** Use position-based dispersion metrics.

**Problem:** Given the dataset:  
 22, 29, 25, 31, 35, 40, 45, 48, 50

#### **a) Arrange data in ascending order**

22, 25, 29, 31, 35, 40, 45, 48, 50

#### **b) Range**

* Range = Max - Min = 50 - 22 = 28

#### **c) Q1, Q3, and IQR**

* Q1 (First Quartile) = 25
* Q3 (Third Quartile) = 45
* IQR = Q3 - Q1 = 45 - 25 = 20

#### **Q8. Five-number Summary and Boxplot Concept**

**Objective:** Summarize distribution of data.

**Problem:**

* **Define the Five-number Summary and explain each component:**
  + **Minimum**: The smallest value in the dataset.
  + **Q1 (First Quartile)**: The median of the lower half of the dataset.
  + **Median**: The middle value.
  + **Q3 (Third Quartile)**: The median of the upper half of the dataset.
  + **Maximum**: The largest value.
* **Describe how boxplots help in detecting outliers.**Boxplots visualize the five-number summary and highlight potential outliers (values outside 1.5 times the IQR from Q1 and Q3).

#### **Q9. Confidence Interval for the Mean**

**Objective:** Estimate population means using sample data.

**Problem:** A sample of 36 students has an average height of 162 cm with a standard deviation of 6 cm.  
 Calculate the 95% Confidence Interval for the population mean.  
 (Hint: Use Z = 1.96 for 95% confidence)

Given:

* Sample mean = 162 cm
* Standard deviation = 6 cm
* Sample size = 36
* Z = 1.96 for 95% confidence

#### **Confidence Interval Calculation**

Standard Error (SE) = 6 / sqrt(36) = 1

Margin of Error = 1.96 × 1 = 1.96

{95% CI} = 162 \pm 1.96 = (160.04, 163.96)

Confidence Interval: 160.04cm to 163.96cm

#### **Q10. Hypothesis Testing - One Sample Z-Test**

**Objective:** Make decisions using statistical testing.

**Problem:** The average salary in a city is ₹30,000. A random sample of 49 employees has an average salary of ₹31,000 with a standard deviation of ₹4,900.  
 Test the hypothesis at the 5% level of significance to determine if the average salary has increased.

* **a) State the null and alternative hypothesis:**
* **b) Calculate the Z-score:**
* **c) Conclude the result using critical value (±1.96)**

**Given:**

* Population mean = ₹30,000
* Sample mean = ₹31,000
* Sample size = 49
* Standard deviation = ₹4,900
* Significance level = 0.05

#### **a) Null and Alternative Hypothesis**

* **Null Hypothesis (H₀)**: The average salary has not increased.  
   μ=30,000\mu = 30,000μ=30,000
* **Alternative Hypothesis (H₁)**: The average salary has increased.  
   μ>30,000\mu > 30,000μ>30,000

#### **b) Z-score Calculation**

SE = 4,900 / sqrt (49) = 4,900/7 = 700

Z = (31,000 - 30,000) / 700 = 1,000 / 700 ≈ 1.43

#### **c) Conclusion**

* Since the Z-score (1.4286) is less than the critical value (±1.96), we **fail to reject the null hypothesis**, meaning there is insufficient evidence to conclude that the average salary has increased.

### **Dataset for the Project:**

You can use synthetic datasets or refer to the following sources for real-world data:

* **Kaggle Datasets:**
  + [Kaggle Datasets](https://www.kaggle.com/datasets)

### **Expected Outcomes:**

* Students will gain proficiency in computing and interpreting key statistical measures.
* They will develop skills to perform confidence intervals and hypothesis testing.
* This project will allow students to apply descriptive and inferential statistical methods to real-world datasets.