### **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.

key statistical definitions

**Population**: The entire set of individuals or items you're interested in studying. Entire data that has ever existed w.r.t the problem statement.

**Sample**: A subset of the population data selected for analysis.

**Sampling**: The process of extracting sample data from Population Data.

**Parameter**: A measurable characteristic of a population (e.g., population mean μ).

**Statistic**: A measurable characteristic from a sample (e.g., sample mean x̄).

data types

Nominal: Categories that **cannot be ordered** or ranked. Only **labels or names**.

**No numerical meaning**. Cannot perform mathematical operations

Eg only Labels Coke or Pepsi, Gender( Male or Female) , Blood type (A, B, AB, O), Eye color (Brown, Blue, Green), Yes/No responses

Ordinary : Categories that have a meaningful order, but the differences between values are not measurable. **Ranking is possible**, but intervals are not equal. Can say one value is greater than another, but not **how much greater**

.

Conveys only preference information. Eg I prefer Coke to Pepsi,

Satisfaction level (Poor < Fair < Good < Excellent)

Education level (High school < Bachelor's < Master's)

Survey ratings (1 to 5 stars)

Interval : Numeric data with **equal intervals**, but **no true zero** point. Can measure the **difference** between values. Cannot calculate **ratios** (e.g., 40°C is not “twice as hot” as 20°C)

Conveys relative magnitude. Can be ranked /ordered but not measurable. Degree, school ranking but not ratio. It is not Zero but in between. Eg I rate Coke a 7 and Pepsi a 4 on a scale of 10.

Temperature in Celsius or Fahrenheit. Temperature will not start from zero but in Between.

Dates (e.g., 1990, 2020)

Time of day (e.g., 10:00 AM)

Ratio : Conveys information in a absolute scale. ( I paid Rs 11 for a coke and Rs12 for a Pepsi)

Numeric data with equal intervals **and a true zero** point.

All mathematical operations are valid: addition, subtraction, multiplication, division.

Can calculate **ratios** (e.g., 4 kg is twice as heavy as 2 kg).

Eg Height, weight, age, distance

Temperature in **Kelvin**

Number of items sold

**Quantitative (Numerical)**:

Discrete: Countable (e.g., number of students).

Continuous: Measurable (e.g., height, weight)  
**Qualitative (Categorical)**:  
*Nominal*: No intrinsic order (e.g., gender, color).

Ordinal: With intrinsic order (e.g., rating scale: poor, average, good).

#### Population vs Sample:

| **Concept** | **Population** | **Sample** |
| --- | --- | --- |
| Size | Large / entire group | Small / subset |
| Use | True values  (rarely accessible) | Practical and cost-effective |
| Notation | μ, σ, P | x̄, s, p |

* + Learn how to calculate measures of central tendency and dispersion.

measures of central tendency and dispersion   
Central Tendency:  
**Mean (Average)**: Sum of all values / number of values.

**Median**: Middle value when data is ordered.

**Mode**: Most frequent value in the dataset. Most occurring value in the data.

Dispersion:

**Range**: Max – Min.   
**Variance (σ² or s²)**: Average of squared deviations from the mean.

**Standard Deviation (σ or s)**: Square root of variance.  
**Interquartile Range (IQR)**: Q3 − Q1; measures the spread of the middle 50%.The middle half of a data set falls within the inter-quartile range.

**Deviation** is the distance from the mean. **Deviation score = observation - true mean**

**Variance** = mean or average of squared deviation scores.

**How far the Data Values present around the mean.**

**Average Squared distance of every data point from mean.**  
**Standard Deviation** = square root of variance.

.

* + Understand how confidence intervals and hypothesis testing are used to infer characteristics about a population from sample data.

Confidence Intervals & Hypothesis Testing

Confidence Interval (CI): A **range of values** used to estimate the **true population parameter**.

Example: "We are 95% confident that the population mean lies between [a, b]".

Hypothesis Testing: Used to make decisions or inferences about population parameters.

**Null Hypothesis (H₀)**: No effect or status quo.

**Alternative Hypothesis (H₁)**: There is an effect or difference.

1. **Hands-on Learning:**
   * Manually calculate measures such as mean, median, mode, variance, and standard deviation.

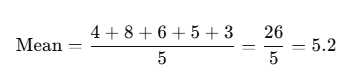
dataset:

Data: 4, 8, 6, 5, 3

Mean (Average)

Mean = (4+8+6+5+3​)/ 5=26​/5=5.2





Median (Middle value)

Sort the data: 3, 4, 5, 6, 8

Middle value (odd number of values) = \*\*5\*\*

Median = 5

Mode (Most frequent value) All values occur once , **No mode** (or all are modes).

Variance (Measure of spread)

Step 1: Find the Mean=5.2

Step 2: Subtract Mean from each value and square the result:

(3−5.2)2=4.84

(4−5.2)2=1.44,

(5−5.2)2=0.04,

(6−5.2)2=0.64,

(8−5.2)2=7.84,

Step 3: Average of squared differences: (4.84+1.44+0.04+0.64+7.84) / 5 = 2.96

Variance = 2.96

Standard Deviation

Standard Deviation= = 1.72

* Interpret statistical results, including the concept of confidence intervals and hypothesis testing.

Interpret Statistical Results

Confidence Interval (CI)

Suppose your sample mean is 100, standard deviation = 15, and n = 25. You want a 95% CI:   
CI=ˉ ± Z × (n​​)

For 95% confidence, Z ≈ 1.96:

CI=100 ± 1.96 × (25​15​) = 100 ± 1.96 × 3 = 100 ± 5.88

We are 95% confident that the population mean lies between **94.12 and 105.88**

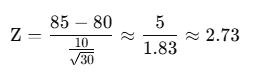
Hypothesis Testing

Suppose you want to test whether a new teaching method affects scores (mean = 80), and your sample mean is 85

**H₀**: μ = 80 (no effect)

**H₁**: μ ≠ 80 (some effect)

sample size = 30, std dev = 10



Z = (​85−80) / (10/ ) ​≈ (5/1.835 )​≈2.73

If Z > 1.96 or < -1.96 → **Reject H₀**

Here, Z = 2.73 > 1.96 → **Statistically significant**

1. **Model Evaluation:**
   * Practice interpreting the results of statistical tests and applying them to make data-driven decisions.

Interpreting the Results of Statistical Tests  
Statistical tests help assess whether patterns in your data are **statistically significant** or could have occurred by **random chance**. The interpretation centers around **p-values**, **test statistics**, and **confidence levels**.

#### Common Statistical Tests & Their Use:

| **Test** | **Use Case** | **Example** |
| --- | --- | --- |
| **Z-test / t-test** | Compare means | Is average test score higher with new teaching method? |
| **Chi-square test** | Categorical associations | Is gender associated with purchase behavior? |
| **ANOVA** | Compare multiple means | Do 3 types of fertilizers produce different crop yields? |
| **Correlation test (Pearson)** | Relationship between variables | Are hours studied and scores related? |

### 

### 

### 

### **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 entire group of individuals or items you want to study or draw conclusions about.  
  **Sample**: A smaller group selected from the population, used to analyze and make inferences.

* **b) Descriptive Statistics and Inferential Statistics  
    
  Descriptive Statistics**: Methods used to summarize, organize, and present data using numbers, tables, or graphs.  
  **Inferential Statistics**: Techniques that use sample data to make predictions or generalizations about a larger population.

* **c) Parameter and Statistic  
    
  Parameter**: A fixed numerical value that describes a characteristic of a population (e.g., population mean).  
  **Statistic**: A value calculated from a sample that estimates a population parameter (e.g., sample mean).
* **d) Qualitative and Quantitative Data  
  Qualitative Data**: Descriptive data that categorizes without numerical values (e.g., colors, names, labels).  
  **Quantitative Data**: Numerical data that can be measured or counted (e.g., height, age, salary).

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

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

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

* **a) Mean :** The **mean** is the sum of all data values divided by the total number of values. It represents the **average** of the dataset.

Mean= Sum of all values / Number of values​

Eg Data Set = 4,6,8,10,12

Mean = (4+6+8+10+12) / 5 = 40/5 = 8

The average value is 8

* **b) Median:** The **median** is the **middle value** when the dataset is arranged in ascending or descending order.  
  \*If the number of observations is **odd**, the median is the **middle value**.  
  \*If the number of observations is **even**, the median is the **average of the two middle values**.

Dataset = **5, 7, 9, 11, 13** (odd count)

The odd count is 5 , the mid value is 3rd ;;;;

Median = 9

Dataset = **2, 4, 6, 8** (even count)

The even count is 4, the mid value is 2nd and 3rd

Median = (4 + 6) / 2 = 10 / 2 = 5

The median divides the dataset into two equal halves

* **c) Mode :** The **mode** is the value that **appears most frequently** in the dataset.

A dataset can have:  
**No mode** (if all values occur once)  
**One mode** (unimodal)  
**Two modes** (bimodal)  
**More than two modes** (multimodal)

Dataset = **3, 5, 7, 5, 8, 5, 9**

**Here mode is 5 as it is appearing 3 times.**

#### **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**  
  Add all the numbers Sum=12+18+14+16+18+20+18+15+12+18+14+16+18+20+18+15=260Count how many numbers are there  
  Count=16  
    
  Mean = Sum / Count = 260 / 16 = 16.25
* **b) Median**Sort the data : 12, 12, 14, 14, 15, 15, 16, 16, 18, 18, 18, 18, 18, 18, 20, 20  
  Median is the middle value(s): There are **16 numbers** (even), so the median is the average of the 8th and 9th numbers.

8th number = 16  
9th number = 18

Median=(16+18) / 2 ​=34​ / 2 = 17

* **c) Mode  
  Find the number that appears most frequently**:  
  **18** appears **6 times**, more than any other number.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**Nominal data consists of **categories or labels**.  
There is **no order** or ranking between the categories.  
Only **names**, **labels**, or **classifications** are used.  
**Mathematical calculations are not meaningful** here.

**Example**   
Gender: **Male, Female, Others**Blood group: **A, B, AB, O**  
Brands: **Coke, Pepsi, Fanta**

**b) Ordinal**Ordinal data shows **categories with a meaningful order** or **ranking**.  
The order matters, but the **differences between values are not measurable** or equal.

Examples:--

Satisfaction levels: **Poor < Fair < Good < Excellent**  
Education levels: **High School < Bachelor’s < Master’s < Ph.D.**  
Competition results: **1st place, 2nd place, 3rd place**

We know the **order**, but **not the exact difference** between levels

**c) Interval**Interval data has **numeric values** with **equal intervals** between them.  
However, **zero is not absolute** (zero does **not** mean absence).  
We can perform **addition** and **subtraction**, but **ratios** are **not meaningful**.

Example

Temperature in **Celsius** or **Fahrenheit**:  
20°C is **not twice as hot** as 10°C.  
Dates: **1990, 2000, 2010**  
IQ scores

Equal spacing exists, but **no true zero**.

**d) Ratio**Ratio data is like interval data but has an **absolute zero point**.  
**All mathematical operations** (addition, subtraction, multiplication, division) are valid.  
Ratios **make sense** here.

**Example:**

Height: **160 cm, 170 cm, 180 cm**  
Weight: **50 kg, 60 kg**  
Age: **10 years, 20 years**  
Income: **₹20,000, ₹40,000**

Zero means **complete absence**, and we can compare **how many times bigger or smaller** one value is.

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

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

**Problem:**

* **a) Define Variance and Standard Deviation.**

Variance (σ² or s²)   
Variance measures **how far the data values are spread out from the mean**.  
It is the **average of the squared differences** between each data point and the mean.  
σ² = (∑ (xi​−xˉ)2​ ) / N  
where:  
xi= each data point  
 xˉ = mean of the data  
 N = number of data points

Larger variance → **more spread out** data;   
smaller variance → **data is close to the mean**.

Example: Data: 2, 4, 6  
Mean (xˉ) = 4  
Variance = ((2−4)2+(4−4)2+(6−4)2) / 3

= (4+0+4) / 3=2.67

#### Standard Deviation (σ or s) Standard deviation is the **square root of variance**. It tells us, **on average, how far the data points are from the mean**. σ = squareroot σ2

Standard deviation is expressed in the **same units** as the original data, making it easier to interpret.

From the above variance:  
σ= squareroot 2.67≈1.63

* **b) Explain why Standard Deviation is more interpretable than Variance.** Standard Deviation is Easier to understand directly as it uses same value as data eg m, Kg etc. It is widely used in reports and summary.

**Standard Deviation**-- Square root of variance, shows **average deviation from the mean** in **original units** , **easier to understand**.

**A standard deviation** of **5 cm** clearly tells us that, **on average, students’ heights differ from the mean by about 5 cm**.

**Variance** --- Measures the **spread** but in **squared units** which is harder to interpret.

#### **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**

**N =9**

Mean = x

Mean = (8 +10+12+14+16+10+12+14+16) / 9 = 112 / 9 =12.4444

Compute each squared deviation

|  |  |  |
| --- | --- | --- |
| **x** | **x-x-** | **(x-x-)2** |
| 8 | 8 - 12.44 = -4.44 | 19.71 |
| 10 | 10 - 12.44 = -2.44 | 5.95 |
| 12 | 12 - 12.44 = -0.44 | 0.19 |
| 14 | 14 - 12.44 = 1.56 | 2.43 |
| 16 | 16 - 12.44 = 3.56 | 12.67 |
| 10 | 10 - 12.44 = -2.44 | 5.95 |
| 12 | 12 - 12.44 = -0.44 | 0.19 |
| 14 | 14 - 12.44 = 1.56 | 2.43 |
| 16 | 16 - 12.44 = 3.56 | 12.6 |

Squared deviations (sum) ∑(xi​−xˉ)2 = 62.2222

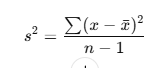
1 9.71+5.95+0.19+2.43+12.67+5.95+0.19+2.43+

Now add carefully:  
19.71 + 5.95 = 25.66  
25.66 + 0.19 = 25.85  
25.85 + 2.43 = 28.28  
28.28 + 12.67 = 40.95  
40.95 + 5.95 = 46.90  
46.90 + 0.19 = 47.09  
47.09 + 2.43 = 49.52  
49.52 + 12.67 = 62.19

Sum of (x−mean)² = 62.19

Sample Variance (s²)

s²= ∑(x−xˉ)2​) / (n−1)



s²=62.19 / 8 = 7.77

Sample Variance = 7.77

* **b) Sample Standard Deviation**



s= Sq Root of (s2​) =Sq Root of (7.77) ​=2.79

Sample Standard Deviation = 2.79 (approx)

### **Final Answers**

| **Measure** | **Symbol** | **Value** |
| --- | --- | --- |
| Mean |  | **12.44** |
| Sample Variance |  | **7.77** |
| Sample Standard Deviation |  | **2.79** |

#### **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) Calculate the Range**Range=Maximum value−Minimum value

Range=50−22=28

✅ **Range = 28**

**c) Find Q1, Q3, and IQR**Q2 (Median) n=9 (odd),

Q2=Middle value=35 (5th Value)

Q1 = Median of **lower half** (values before overall median).   
Lower half = **22, 25, 29, 31**  
Q1 = (25 + 29) / 2 = 54 / 2 = 27

Q3 (Third Quartile)  
Q3 = Median of **upper half** (values after overall median).  
Upper half = **40, 45, 48, 50**

Q3 = (45 + 48) / 2 = 93/ 2= 46.5

IQR (Interquartile Range)

IQR= Q3 - Q1 = 46.5 – 27= 19.5

✅ **Q1 = 27, Q3 = 46.5, IQR = 19.5**

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

**Objective:** Summarize distribution of data.

**Problem:**

* **Define the Five-number Summary and explain each component:**

To **summarize the distribution** of a dataset and **identify outliers** using the **five-number summary** and **boxplots**.The **five-number summary** provides a quick overview of a dataset by identifying its key points. It is a quick way to describe a dataset by showing **five key points**:

* + Minimum - The smallest value in the dataset. Shows the lower boundary of data.

Eg: In Dataset = {10, 12, 14, 16, 20}, Minimum = **10**

Q1 (First Quartile): The value below which **25% of the data** lies. Represents the lower quartile boundary. Q1 = 25th percentile.  
Eg: In Dataset = {10, 12, 14, 16, 20}, Q1= **12**

* + Median Q2: The **middle value** when data is sorted. Divides the dataset into **two equal halves**. Median = 50th percentile

Eg: In Dataset = {10, 12, 14, 16, 20}, Median Q2= **14**

* + Q3 (Third Quartile) : The value below which **75% of the data** lies. Represents the upper quartile boundary. Q3 = 75th percentile

Eg: In Dataset = {10, 12, 14, 16, 20}, Median Q3= **16**

* + Maximum : The largest value in the dataset. Shows the upper boundary of data. For the same dataset, Maximum = **20.**

Eg: In Dataset = {10, 12, 14, 16, 20}, Maximum= **20**

|  |
| --- |
|  |

* **Describe how boxplots help in detecting outliers.**

Arrange the dataset in ascending order.   
**Median (Q2):**If **n** is odd → middle value.  
If **n** is even → average of the two middle values.  
**Q1:** Median of the **lower half** (below the overall median).  
**Q3:** Median of the **upper half** (above the overall median).

A **boxplot** (or **whisker plot**) is a graphical representation of the five-number summary.  
**Box:** Represents the **Interquartile Range (IQR)**, from **Q1 to Q3**.  
**Median Line:** Inside the box, shows the dataset’s center.  
**Whiskers:** Extend from the box to **Minimum** and **Maximum** (within limits).  
**Outliers:** Points that lie **beyond 1.5 × IQR** from Q1 or Q3 are plotted separately.

**Lower Bound:** Q1−1.5×IQRQ1 - 1.5 \times IQRQ1−1.5×IQR  
**Upper Bound:** Q3+1.5×IQRQ3 + 1.5 \times IQRQ3+1.5×IQR  
Any value **outside** this range is an **outlier**.

**Example**

Dataset = {10, 12, 14, 16, 20, 25, 40}

Q1 = 12, Q3 = 25  
IQR = Q3 − Q1 = 25 − 12 = **13**  
Lower Bound = 12 − 1.5(13) = **−7.5**  
Upper Bound = 25 + 1.5(13) = **44.5**  
Any value above **44.5** or below **−7.5** is an **outlier**.  
Here, **40** is **within range**, so **no outliers**.

**Minimum** Lowest data point  
Q1 25% of data below this point  
Median Middle of the dataset  
Q3 75% of data below this point  
Maximum Highest data point  
IQR Spread of middle 50% (Q3 − Q1)  
Outliers Beyond 1.5 × IQR

#### 

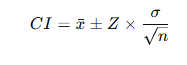
#### **Q9. Confidence Interval for the Mean**

**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)

To calculate the 95% confidence interval for the population mean given a sample mean of 162 cm, sample size of 36 and sample standard deviation of 6 cm, using a Z value of 1.96:   
Sample size (**n**) = 36  
Sample mean (**x̄**) x Bar= 162 cm  
Standard deviation (**σ**)sigma = 6 cm  
Z-value for 95% confidence = **1.96**

To calculate the confidence interval for the mean (CI):



CI = 162 ± 1.96 \* 6 / Sq Root of 36 = 162 ± 1.96 (6/6) = 162 ± 1.96(1)= 162 ± 1.96

Calculate Standard Error (SE) SE= σ(sigma) / Sq Root n  
SE = 6 / Sq Root 36 = 6/6 = 1

Calculate Margin of Error (ME)  
ME=Z×SE=1.96×1=1.96

Confidence Interval CI:   
CI= 162±1.96

Lower Limit = 162 – 1.96 = 160.04 Cm  
Upper Limit = 162 + 1.96 = 163.96 Cm  
  
**95% Confidence Interval** = **(160.04 cm, 163.96 cm)**This means we are **95% confident** that the true population mean height lies **between 160.04 cm and 163.96 cm**.

​

Population mean (**μ**) = ₹30,000

Sample mean (**x̄**) = ₹31,000

Standard deviation (**σ**) = ₹4,900

Sample size (**n**) = 49

Significance level (**α**) = 5%

**Z-critical** = ±1.96

a)State the Hypotheses  
We need to check if the average salary has **increased**.

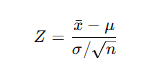
**Null Hypothesis (H₀):** μ = ₹30,000  
The average salary **has not increased**.  
**Alternative Hypothesis (H₁):** μ > ₹30,000.   
The average salary **has increased**. This is right-tailed test

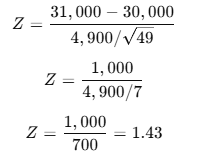
b) Calculate the Z-score  


Standard Error (SE): E= SE= σ(sigma) / Sq Root n



SE = 4900 / Sq Root 49 = 4900 / 7 = 700

Calculate Z  
  
Z = (31000 – 30000 ) / 700 = 1000 / 700 = 1.43



Compare Z-score with Critical Value  
**Calculated Z** = **1.43  
Critical Z** = **+1.96** (for 5% level, right-tailed)

## c) Conclusion Using Critical Value

* **Critical Z (Right-tailed, α=0.05):** 1.645  
  (We use 1.645 since this is a **one-tailed test**, not ±1.96.)

Now compare:

| **Z (calculated)** | **Z (critical)** | **Decision** |
| --- | --- | --- |
| 1.43 | 1.645 | Fail to reject H₀ |

**Conclusion:**  
Since **Z = 1.43 < 1.645**,  
we **fail to reject the null hypothesis**.

### **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.