Part A: Theory & Definitions

1. Definitions with Examples from the Dataset

a) Types of Data: Numerical & Categorical

Numerical Data
Numerical data represents quantities that can be measured or counted and on which mathematical operations can be performed.

Examples from the dataset:

- **Age_of_Household_Head** – Measured in years.
- **Household_Income** – Monthly income in local currency.
- **Family_Size** – Number of family members.

Categorical Data
Categorical data represents qualitative characteristics or categories and cannot be meaningfully used in arithmetic calculations.

Examples from the dataset:

- **Household_ID** – Unique identifier.
- **Education_Level** – Primary, Secondary, Graduate, Post-Graduate.
- **Owns_House** – Yes or No.
- **Urban_Rural** – Urban or Rural.

---

b) Types of Statistics: Descriptive vs. Inferential

Descriptive Statistics
Descriptive statistics summarize and organize data to understand its main characteristics.

Examples:

- Mean income of households.
- Median age of household heads.
- Income distribution using histograms.

Inferential Statistics
Inferential statistics use a sample to make predictions or generalizations about a larger population.

Examples:

- Estimating average income of all households based on the dataset.
- Testing whether urban households earn more than rural households.

---

c) What is Descriptive Statistics?

Descriptive statistics is the branch of statistics that focuses on collecting, summarizing, and presenting data in a meaningful way. It includes measures such as mean, median, mode, range, variance, standard deviation, skewness, and kurtosis to describe the dataset without making predictions beyond the data.

---

2. Differences Between Statistical Measures

a) Mean, Median, and Mode

Measure	Definition	Key Property
Mean	Average of all values	Sensitive to extreme values
Median	Middle value after sorting	Robust to outliers
Mode	Most frequent value	Useful for categorical data

Example:
If household incomes have a few very high values, the mean will be higher than the median, indicating income inequality.

---

b) Range, Variance, and Standard Deviation

Measure	Definition	Interpretation
Range	Maximum − Minimum	Overall spread
Variance	Average squared deviation from mean	Dispersion magnitude
Standard Deviation	Square root of variance	Spread in original units

Standard deviation is preferred over variance because it is in the same unit as the data.

Question 3: Explain the following terms with neat and clean diagram along with its formula

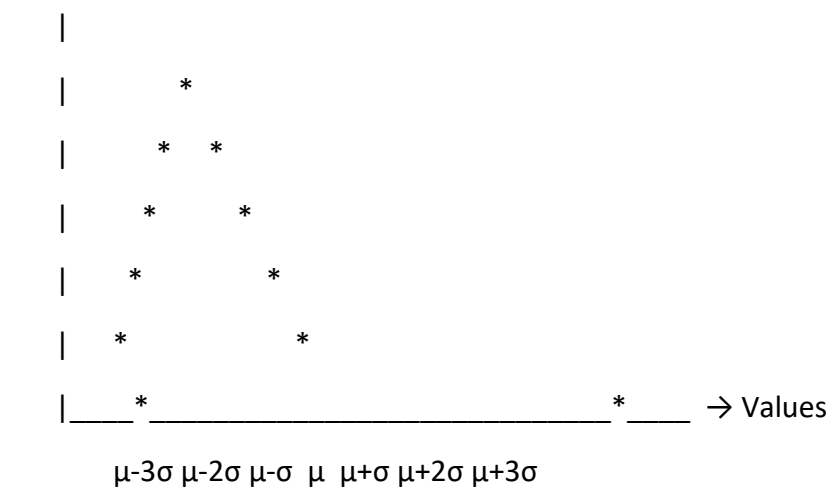
---

1. Gaussian (Normal) Distribution

Definition

A Gaussian or Normal Distribution is a continuous probability distribution that is symmetric around its mean. In this distribution, the mean, median, and mode are equal, and most data points cluster around the center.

Diagram (Bell Curve)



Formula

$$
f(x) = \frac{1}{\sigma \sqrt{2\pi}} ; e^{-\frac{(x-\mu)^2}{2\sigma^2}}
$$

Where:

- μ = Mean
- σ = Standard deviation

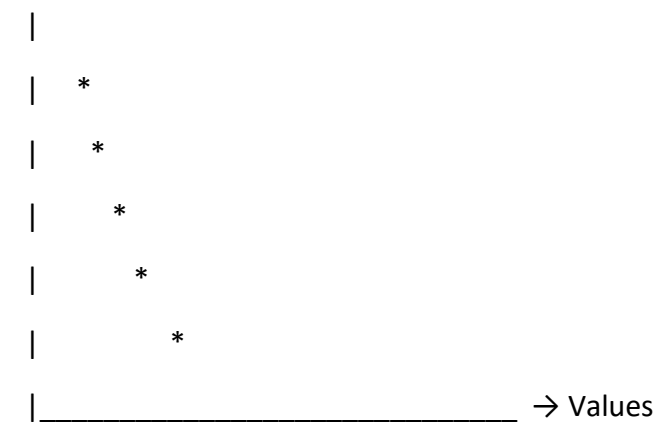
---

2. Log-Normal Distribution

Definition

A Log-Normal Distribution is a distribution where the logarithm of the variable follows a normal distribution. It is positively skewed and commonly used for income data.

Diagram (Right-Skewed Curve)



Formula

$$
f(x) = \frac{1}{x\sigma\sqrt{2\pi}} ; e^{-\frac{(\ln x - \mu)^2}{2\sigma^2}}, \quad x>0
$$

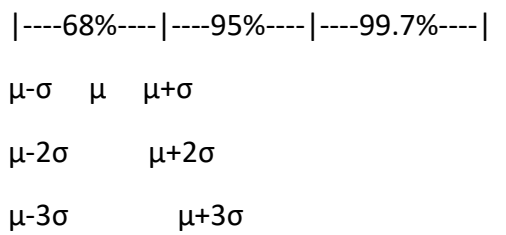
---

3. 3-Sigma Rule (Empirical Rule)

Definition

The 3-Sigma Rule states that for a normal distribution, nearly all data lies within three standard deviations from the mean.

Diagram



Rule

- 68% data lies within ±1σ
- 95% data lies within ±2σ
- 99.7% data lies within ±3σ

---

4. Percentiles

Definition

Percentiles divide the dataset into 100 equal parts. The p-th percentile indicates the value below which p% of the data falls.

Diagram

0% 25% 50% 75% 100%

|------|-------|-------|--------|

Min Q1 Median Q3 Max

Formula

$$
P_k = \frac{k}{100} \times (n+1)
$$

Where:

- (k) = percentile number
- (n) = total observations

---

5. Quartiles

Definition

Quartiles divide the dataset into four equal parts.

- Q1 = 25th percentile
- Q2 = 50th percentile (Median)
- Q3 = 75th percentile

Diagram

Min ---- Q1 ---- Median ---- Q3 ---- Max

Formula

$$
Q_k = \frac{k(n+1)}{4}, \quad k = 1,2,3
$$

---

6. Five Number Summary

Definition

The five-number summary gives a complete overview of data distribution using five key values.

Components

- Minimum
- First Quartile (Q1)
- Median
- Third Quartile (Q3)
- Maximum

Diagram (Boxplot)

Min —┬——[Q1 | Median | Q3]——┬— Max

---

7. Skewness

Definition

Skewness measures the asymmetry of a distribution.

Types & Diagrams

Positive Skew (Right Skewed)

```
 *
  *
   *
     *
       *
```

Negative Skew (Left Skewed)

```
       *
     *
   *
 *
```

*

Formula

$$
\text{Skewness} = \frac{1}{n} \sum \left(\frac{x - \mu}{\sigma}\right)^3
$$

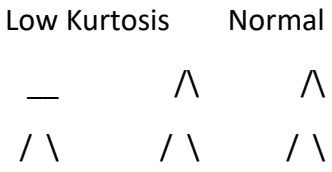
---

8. Kurtosis

Definition

Kurtosis measures the sharpness of the peak and presence of outliers in a distribution.

Diagram



Formula

$$
\text{Kurtosis} = \frac{1}{n} \sum \left(\frac{x - \mu}{\sigma}\right)^4
$$