**1. Definitions**

**Types of Data**

* **Numerical (Quantitative)**: Data that represents numbers and can be measured.
  + *Example*: Household\_Income = 50000, Age = 30.
  + Subtypes:
    - **Discrete** (countable, like number of children = 2).
    - **Continuous** (measurable, like income, height, temperature).
* **Categorical (Qualitative)**: Data that represents categories or labels.
  + *Example*: Gender = Male/Female, Department = HR/Finance/IT.
  + Subtypes:
    - **Nominal** (no order, like colors: Red, Blue, Green).
    - **Ordinal** (ordered categories, like education level: High School < Bachelor < Master).

**Types of Statistics**

* **Descriptive Statistics**: Summarizing and describing data using numbers, tables, or graphs.
  + *Example*: “Average household income in the dataset is ₹50,000, the maximum is ₹1,20,000.”
* **Inferential Statistics**: Drawing conclusions or making predictions about a population using a sample.
  + *Example*: Using data from 200 households to predict the average income of all households in the city.

**What is Descriptive Statistics?**

It is a branch of statistics that **organizes, summarizes, and presents data**. It does **not make predictions** beyond the dataset.

* Tools: Mean, Median, Mode, Range, Graphs, Tables.
* Example: "In our dataset, the mean household income is ₹50,000, and the income distribution is positively skewed."

**2. Differences**

**Measures of Central Tendency**

* **Mean**: Arithmetic average. Sensitive to outliers.
  + Example: (10 + 20 + 100) / 3 = 43.3
* **Median**: Middle value when data is sorted. Not affected by outliers.
  + Example: (10, 20, 100) → Median = 20
* **Mode**: Most frequent value.
  + Example: (10, 20, 20, 30) → Mode = 20

**Measures of Spread**

* **Range**: Difference between max and min.
  + Example: (100 – 10) = 90
* **Variance**: Average squared deviation from the mean.
* **Standard Deviation (σ)**: Square root of variance, shows spread in same unit as data.

**3. Key Terms with Diagrams & Formulas**

(I’ll describe diagrams clearly, but you can later plot them in Python with matplotlib or seaborn.)

**Gaussian (Normal) Distribution**

* Bell-shaped, symmetric around mean.
* Formula:

f(x)=1σ2πe−(x−μ)22σ2f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{ -\frac{(x-\mu)^2}{2\sigma^2} }f(x)=σ2π​1​e−2σ2(x−μ)2​

* Example: Heights, exam scores.
* **Diagram**: Bell curve centered at mean (µ).

**Log-Normal Distribution**

* If a variable’s **logarithm is normally distributed**, the variable itself follows log-normal.
* Skewed to the right (positively skewed).
* Example: Income distribution.
* Formula:

f(x)=1xσ2πe−(ln⁡x−μ)22σ2,x>0f(x) = \frac{1}{x\sigma\sqrt{2\pi}} e^{ -\frac{(\ln x-\mu)^2}{2\sigma^2} }, \quad x>0f(x)=xσ2π​1​e−2σ2(lnx−μ)2​,x>0

**3-Sigma Rule (Empirical Rule)**

For a normal distribution:

* 68% of data lies within **µ ± 1σ**
* 95% of data lies within **µ ± 2σ**
* 99.7% of data lies within **µ ± 3σ**

**Diagram**: Bell curve with shaded regions at ±1σ, ±2σ, ±3σ.

**Percentiles**

* A percentile is a value below which a given % of observations fall.
  + Example: 90th percentile = score above which top 10% students lie.

**Quartiles**

* Special percentiles dividing data into 4 equal parts:
  + Q1 = 25th percentile
  + Q2 = Median = 50th percentile
  + Q3 = 75th percentile

**Five Number Summary**

1. Minimum
2. Q1 (25th percentile)
3. Median (Q2)
4. Q3 (75th percentile)
5. Maximum

Often visualized using a **boxplot**.

**Skewness**

* Measure of asymmetry of data.
  + Positive Skew (right tail longer, e.g., incomes).
  + Negative Skew (left tail longer).
* Formula:

Skewness=∑(xi−xˉ)3/n(σ3)\text{Skewness} = \frac{ \sum (x\_i - \bar{x})^3 / n }{ (\sigma^3) }Skewness=(σ3)∑(xi​−xˉ)3/n​

**Kurtosis**

* Measures **tailedness** of distribution.
  + High kurtosis = more outliers (heavy tails).
  + Low kurtosis = light tails.
* Formula:

Kurtosis=∑(xi−xˉ)4/n(σ4)\text{Kurtosis} = \frac{ \sum (x\_i - \bar{x})^4 / n }{ (\sigma^4) }Kurtosis=(σ4)∑(xi​−xˉ)4/n​