Descriptive Biostatistics, Data Types and Python for Data Analysis

1. Why Does Data Matter?

Healthcare today is driven not only by medicines but also by data. Every patient consultation, laboratory investigation, prescription, imaging report, and adverse drug reaction generates valuable information that can improve patient care and healthcare decisions.

Examples of data generated in healthcare include:

- Patient demographics
- Laboratory test results
- Medical images
- Electronic Health Records (EHRs)
- Prescriptions
- Pharmacovigilance reports

In the pharmaceutical industry, data are also generated during:

- Drug manufacturing
- Quality control
- Batch production
- Clinical trials
- Drug distribution
- Research and development

Good data enable evidence-based decisions and improve patient safety.

2. The process of using data involves:

Data Collection → Data Organization → Statistical Analysis →

→ Interpretation → Decision Making

Statistics convert raw numbers into meaningful information that supports clinical and pharmaceutical decisions.

3. Types of Data

Data are broadly classified into:

A. Categorical (Qualitative): which represent labels or categories.

A1. Nominal Data (Categories with no natural order)

Examples: Blood group, Gender etc.

A2. Ordinal Data (Categories that have a meaningful order).

Examples: Pain score, Disease stage etc.

B. Numerical (Quantitative): which represent measurable quantities.

B1. Discrete Data

- Countable values
- Whole numbers only

Examples: Number of patients, Number of prescriptions, Number of ADRs etc.

B2. Continuous Data:

- Measured quantities
- and can take any value within a range

Examples: Height, Weight, Blood pressure etc.

C. Interval Data Equal intervals between values and No true zero

Examples: Temperature (°C), IQ score, Pain scale etc.

D. Ratio Data Equal intervals and has a true zero

Examples: Age, Body weight, Plasma drug concentration etc.

4. Descriptive Statistics

Descriptive statistics summarize data.

A. **Measures of Central Tendency**

- Mean (Arithmetic average.)
- Median (Middle value after arranging observations.)
- Mode (Most frequently occurring value.)

B. **Measures of Dispersion**

Describe how spread out the data are.

- Range
- Variance
- Standard Deviation (SD)

5. Data Distribution

Positive Skew: Tail extends to the right.

Mean > Median > Mode.

Negative Skew: Tail extends to the left.

Mean < Median < Mode.

Symmetric Distribution

Mean = Median = Mode.

Normal Distribution

Characteristics:

- Bell-shaped
- Symmetrical
- Single peak
- Mean = Median = Mode

Empirical Rule: 68% of observations lie within ±1 SD

95% within ±2 SD

99.7% within ±3 SD

Many biological variables approximately follow a normal distribution, including:

Height, Weight, Blood pressure, Body temperature, Hemoglobin, RBC count, WBC count, etc.

6. **Selecting Statistics According to Data Type**

Data Type	Common Measure of Central Tendency	Common Measure of Dispersion
Nominal	Mode	Frequency, Percentage
Ordinal	Median, Mode	Range, IQR
Numerical (Discrete)	Mean, Median, Mode	Range, SD, IQR
Numerical (Continuous)	Mean, Median, Mode	Range, SD, CV
Interval	Mean, Median, Mode	Range, SD
Ratio	Mean, Median, Mode	Range, SD, CV

7. **Introduction to Python Libraries**

A. **Statistics Module**

Used for basic descriptive statistics.

Common functions: statistics.mean()

statistics.median()

statistics.mode()

statistics.stdev()

B. **NumPy**

Used for: Numerical computing, Arrays, Mathematical operations, Matrix calculations

C. **Pandas**

Used for: Reading CSV/Excel files, Data cleaning, Filtering, Organizing data, Data analysis

D. **SciPy**

Used for: Scientific computing, Statistical analysis, Advanced mathematical functions