The Definitive Guide to Descriptive Statistics

From classifying data to understanding its shape, center, and spread. This is your complete guide to telling the story hidden within the numbers.

Part 1: The Building Blocks of Data

Before we can analyze data, we must understand its nature. Every statistical method you'll ever use depends on the type of data you have.

Qualitative Data (গুণবাচক ডেটা)

*Describes non-numerical qualities or categories. Think of it as **descriptive information**.*

Nominal Data
Categories with no intrinsic order.
(e.g., Eye Color: Blue, Brown; Country: Bangladesh, India)

Ordinal Data
Categories that have a meaningful order.
(e.g., T-Shirt Size: Small, Medium, Large; Rating: Poor, Good, Excellent)

Quantitative Data (সংখ্যাবাচক ডেটা)

*Represents numerical values that can be measured or counted. Think of it as **numerical information**.*

Discrete Data
Countable, distinct values (usually integers).
(e.g., Number of students: 35; Pages in a book: 250)

Continuous Data
Measurable values within a range.
(e.g., Height: 175.5 cm; Temperature in Dhaka: 31.2°C)

Part 2: Finding the "Center" of Your Data

Measures of central tendency provide a single value that represents the typical or central point of a dataset.

Mean (গড়)

The sum of all values divided by the number of values; the arithmetic average.

Why it Matters:
The mean is the most common measure of center and uses every value in the dataset, providing a comprehensive summary. However, it's highly sensitive to outliers (extreme values). A billionaire walking into a cafe drastically increases the mean income of the customers, even though nobody's actual income changed.

Median (মধ্যক)

The middle value when the dataset is sorted in ascending order.

Example: Find the median of {10, 5, 25, 15, 30}

1. **Sort the data:** {5, 10, 15, 25, 30}
2. **Find the middle:** The middle value is 15.

Why it Matters:
The median is resistant to outliers. It represents the true midpoint of the data, making it a better measure of center for skewed datasets, like income or house prices.

Mode (প্রচুরক)

The value that appears most frequently in the dataset.

Example: Find the mode of {Apple, Banana, Orange, Banana, Orange, Apple, Orange}

The mode is **Orange** as it appears 3 times.

Why it Matters:
The mode is the only measure of central tendency that can be used for qualitative (nominal) data. It's useful for identifying the most popular choice or common category.

Part 3: Measuring the Spread of Your Data

Measures of dispersion (or variability) describe how spread out or clustered together the data points are.

Range

The difference between the maximum and minimum values. Simple but sensitive to outliers.

Interquartile Range (IQR)

The range of the middle 50% of the data (IQR = Q3 - Q1). It is resistant to outliers.

Standard Deviation (סטיית תקן)

The average distance of each data point from the mean. It is the most common and powerful measure of spread.

Why it Matters:

- A **small** standard deviation indicates that data points are tightly clustered around the mean (high consistency).
- A **large** standard deviation indicates that data points are spread out over a wide range (low consistency).
- It provides a standardized way to understand variability and is the foundation for more advanced statistics.

Part 4: Visualizing the Story

Charts and graphs transform numbers into an intuitive visual story, revealing patterns, trends, and outliers at a glance.

Choosing the Right Chart

Bar Chart
Compares categories of qualitative data. Each bar represents a category, and its height represents a count or percentage.

Histogram
Shows the distribution of quantitative data. Bars are grouped into "bins" (ranges) and the height shows the frequency of data in that bin.

The Box Plot: A 5-in-1 Visual

A box plot (or box-and-whisker plot) is a powerful tool for visualizing the distribution of quantitative data, summarizing five key numbers in one chart.

The box represents the **Interquartile Range (IQR)**, containing the middle 50% of the data. The line inside the box is the **Median**. The "whiskers" extend to the minimum and maximum values (excluding outliers, which are sometimes plotted as individual points).

Part 5: Case Study - Analyzing Daily Cafe Sales

Let's apply everything to analyze the daily sales (in BDT) for a small cafe in Dhaka over 10 days.

```
Sales Data (BDT): { 5500, 6200, 5800, 7500, 6000,
5800, 15000, 6500, 6800, 5900 }
```

Step 1: Sort and Identify Key Values

```
Sorted: { 5500, 5800, 5800, 5900, 6000, 6200, 6500,
6800, 7500, 15000 }
```

Note the outlier: 15000 BDT. This might be from a special event.

Step 2: Central Tendency

Mean: 71000 / 10 = **7100 BDT**

Median: (6000 + 6200) / 2 = **6100 BDT**

Mode: **5800 BDT**

Interpretation: The mean is pulled higher by the 15000 outlier. The median of 6100 BDT is a more accurate representation of a typical day's sales.

Step 3: Dispersion (Spread)

Range: 15000 - 5500 = **9500 BDT**

IQR: Q1=5800, Q3=6800. IQR = 6800 - 5800 = **1000 BDT**

Std. Dev.:
≈
2891 BDT

Interpretation: The Range is huge due to the outlier. The IQR shows the middle 50% of sales are very consistent, varying by only 1000 BDT. The large standard deviation also reflects the outlier's impact.

Step 4: Conclusion & Visualization

A typical day brings in about 6100 BDT in sales. Most days are very consistent (IQR = 1000). However, there is significant variability overall (Std. Dev. = 2891) due to an exceptional sales day of 15000 BDT. A **Box Plot** would be the perfect visualization to show the consistent core sales (the "box") and the outlier (a single point far away from the "whisker").