Class #17: Introduction to Statistics - Part 1

1. Understanding quantitative analysis.
2. Frequency distribution basics.
3. Comparing data presentation methods: Bar Graph vs Histogram.
4. Methods of central tendency (Mean, Median, Mode).
5. Methods of dispersion measurement.
6. Exploring range, variance, and standard deviation.
7. Quartiles, deciles, percentiles, and coefficient of variation.
8. Five-number summary and creating a box plot. **Assignment #17:**

1. Understanding quantitative analysis.

বৈশিষ্ট্য	Qualitative (গুণগত)	Quantitative (পরিমাণগত)
সংজ্ঞা	গুণগত তথ্য যা অনুভূতি, অভিজ্ঞতা বা বৈশিষ্ট্য প্রকাশ করে।	পরিমাণগত তথ্য যা সংখ্যা বা মাপের মাধ্যমে প্রকাশিত হয়।
উদাহরণ	যেমন: মানুষের আচরণ, রঙ, অনুভূতি	যেমন: বয়স, উচ্চতা, তাপমাত্রা
দৃষ্টিভঙ্গি	বর্ণনামূলক, গুণগত	গাণিতিক, পরিমাণগত
তথ্য সংগ্রহের পদ্ধতি	সাক্ষাৎকার, ফোকাস গ্রুপ, পর্যবেক্ষণ	জরিপ, গণনা, পরিসংখ্যান
প্রশ্নের ধরন	কেন? কিভাবে? (উদাহরণ: কেন মানুষ সুখী হয়?)	কতটা? কখন? কোথায়? (উদাহরণ: কতজন মানুষ সুখী?)
ডেটা বিশ্লেষণ	থিম বা বিষয় ভিত্তিক বিশ্লেষণ (যেমন: অভ্যন্তরীণ অনুভূতি)	সংখ্যা বা পরিসংখ্যান ভিত্তিক বিশ্লেষণ (যেমন: কতজন সুখী?)
নমুনা পদ্ধতি	সাধারণত নির্দিষ্ট সংখ্যক মানুষের উপর গবেষণা করা হয় (Non-probability sample)	সাধারণত বৃহৎ পরিসরে নমুনা নেওয়া হয় (Probability-based sample)
সাধারণীকরণযোগ্যতা	সাধারণত অন্যান্য জনগণের ওপর ফলাফল প্রযোজ্য নয়। (Non-generalizable)	অন্যান্য জনগণের ওপর ফলাফল প্রযোজ্য হতে পারে। (Generalizable)

গবেষক কি	গবেষক নিজে সরাসরি তথ্য সংগ্রহ করেন (গবেষক যন্ত্র হিসেবে)	গবেষক বিভিন্ন টুল, যন্ত্র বা সফটওয়্যার ব্যবহার করেন

Basics of Statistics

Definition: Science of collection, presentation, analysis, and reasonable interpretation of data. Statistics presents a rigorous scientific method for gaining insight into data.

2. Frequency distribution basics.

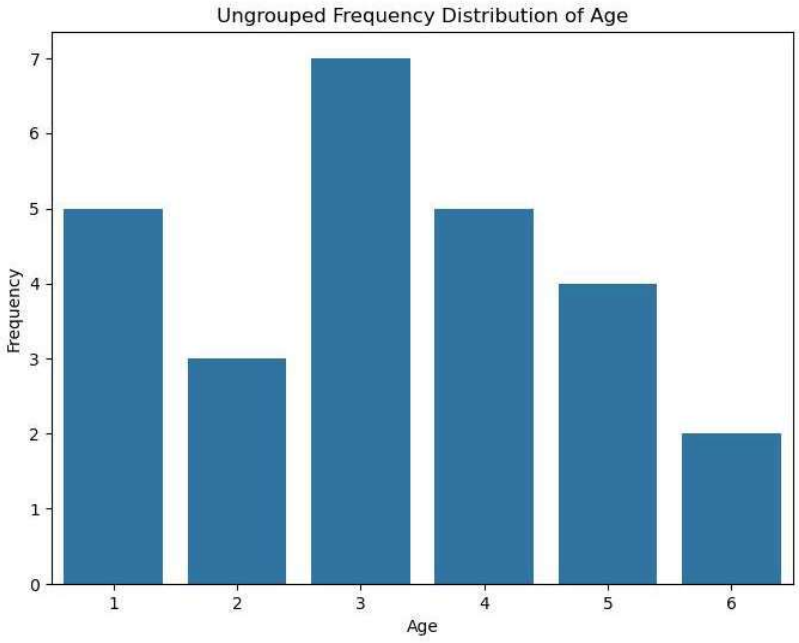
Frequency Distribution (ফ্রিকোয়েন্সি ডিস্ট্রিবিউশন) হলো একটি পদ্ধতি যেখানে তথ্যের প্রতিটি মান বা গোষ্ঠী কতবার ঘটে তা দেখানো হয়। সহজভাবে বললে, এটি আমাদের জানায় যে কোনো নির্দিষ্ট পরিসরের মধ্যে কতবার কিছু ঘটে।

এখানে দুটি ধরনের ফ্রিকোয়েন্সি ডিস্ট্রিবিউশন দেওয়া হয়েছে: **Ungrouped Frequency Distribution** (অগ্রুপিত ফ্রিকোয়েন্সি ডিস্ট্রিবিউশন) এবং **Grouped Frequency Distribution** (গ্রুপিত ফ্রিকোয়েন্সি ডিস্ট্রিবিউশন)।

1. Ungrouped Frequency Distribution (অগ্রুপিত ফ্রিকোয়েন্সি ডিস্ট্রিবিউশন):

এখানে প্রতিটি বয়সের জন্য ফ্রিকোয়েন্সি (কতবার প্রতিটি বয়স পাওয়া গেছে) দেখানো হয়েছে।

Age	Frequency
1	5
2	3
3	7
4	5
5	4
6	2

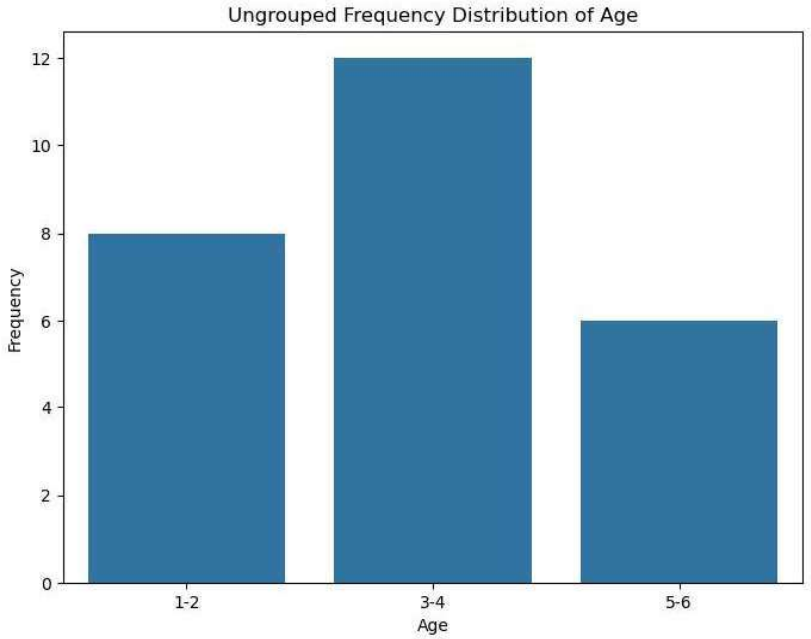


এখন, এই তথ্যের মাধ্যমে, আপনি দেখতে পাচ্ছেন যে বয়স 1-এর জন্য 5 বার, বয়স 2-এর জন্য 3 বার, বয়স 3-এর জন্য 7 বার, এবং এইভাবে অন্যান্য বয়সগুলির জন্য কতবার ঘটেছে তা জানাচ্ছে।

2. Grouped Frequency Distribution (গ্রুপিত ফ্রিকোয়েন্সি ডিস্ট্রিবিউশন):

এখানে, বয়সের সংখ্যা ছোট গ্রুপে ভাগ করা হয়েছে এবং প্রতিটি গ্রুপের জন্য মোট ফ্রিকোয়েন্সি দেখানো হয়েছে। এতে ডেটা আরও সহজে বিশ্লেষণ করা যায়।

Age Group	Frequency
1-2	8
3-4	12
5-6	6



5+3=8

7+5=12

4+2=6

এখানে, বয়স 1 এবং 2 কে একত্রে গ্রুপ করা হয়েছে এবং এই গ্রুপের মোট ফ্রিকোয়েন্সি হলো 8। একইভাবে, বয়স 3 এবং 4 একটি গ্রুপে এবং বয়স 5 ও 6 আরেকটি গ্রুপে রাখা হয়েছে।

এভাবে, **Grouped Frequency Distribution** আপনাকে ডেটাকে বড় গোষ্ঠীতে ভাগ করে বিশ্লেষণ করার সুযোগ দেয়, যা তুলনামূলকভাবে সহজ এবং পরিষ্কার।

3. Comparing data presentation methods: Bar Graph vs Histogram.

Data Presentation: Two types of statistical presentation of data -**graphical and numerical.**

Graphical Presentation: We look for the overall pattern and for striking deviations from that pattern. Over all pattern usually described by shape, center, and spread of the data. An individual value that falls outside the overall pattern is called an *outlier.*

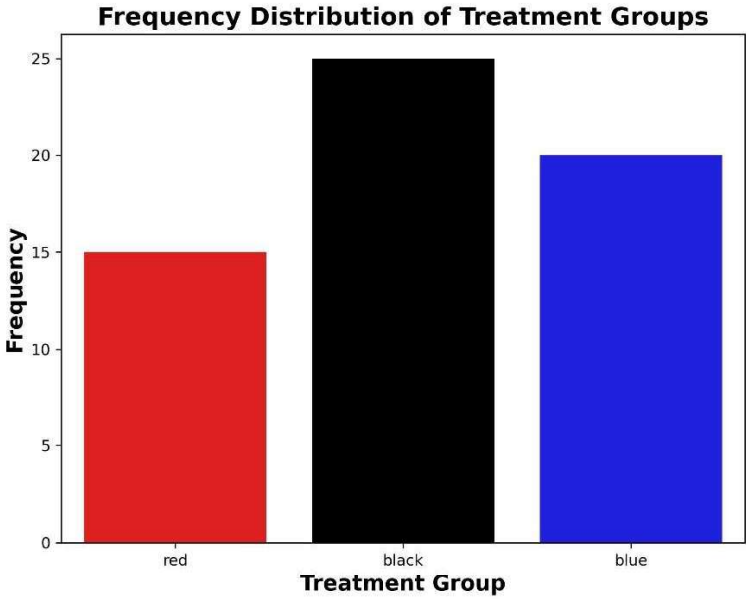
1. Bar diagram and Pie charts are used for categorical variables.
2. Histogram, Box-plot are used for numerical variable.

Data Presentation –Categorical Variable

Bar Diagram: A **bar diagram**(or a bar graph) is a rectangular bar shaped statistical graphic which is divided into several bar to illustrate numerical proportion.

red, black, blue এইগুলা কোনো সংখ্যা না । ক্যাটাগরি ধরা যায় ।

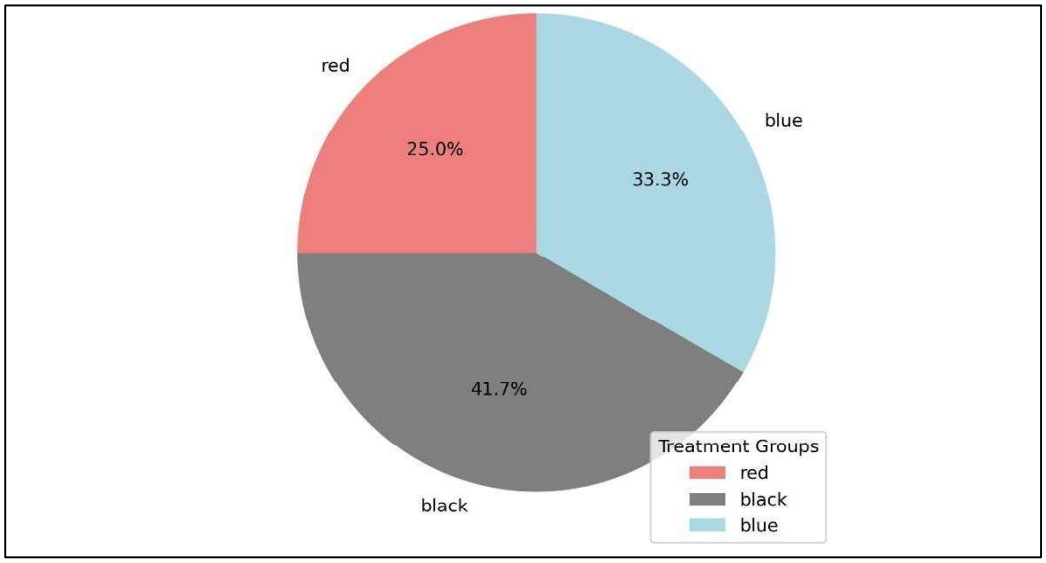
Treatment Group	Frequency
red	15
black	25
blue	20
Total	**60**



Pie Chart: A **pie chart** (or a circle **chart**) is a circular statistical graphic which is divided into slices to illustrate numerical proportion. Pie chart এর সাথে সম্পর্ক percentage

Treatment Group	Frequency	Proportion	Percent (%)
red	15	(15/60)=0.25	25.0
black	25	(25/60)=0.417	41.7
blue	20	(20/60)=0.333	33.3
Total	**60**	**1.00**	**100**

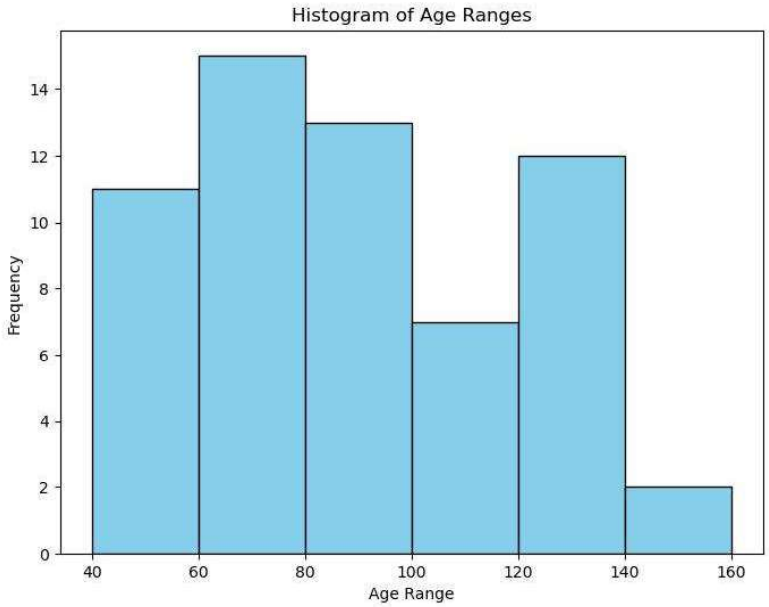
Figure 2: Pie Chart of Subjects in Treatment Groups



Graphical Presentation –Numerical Variable

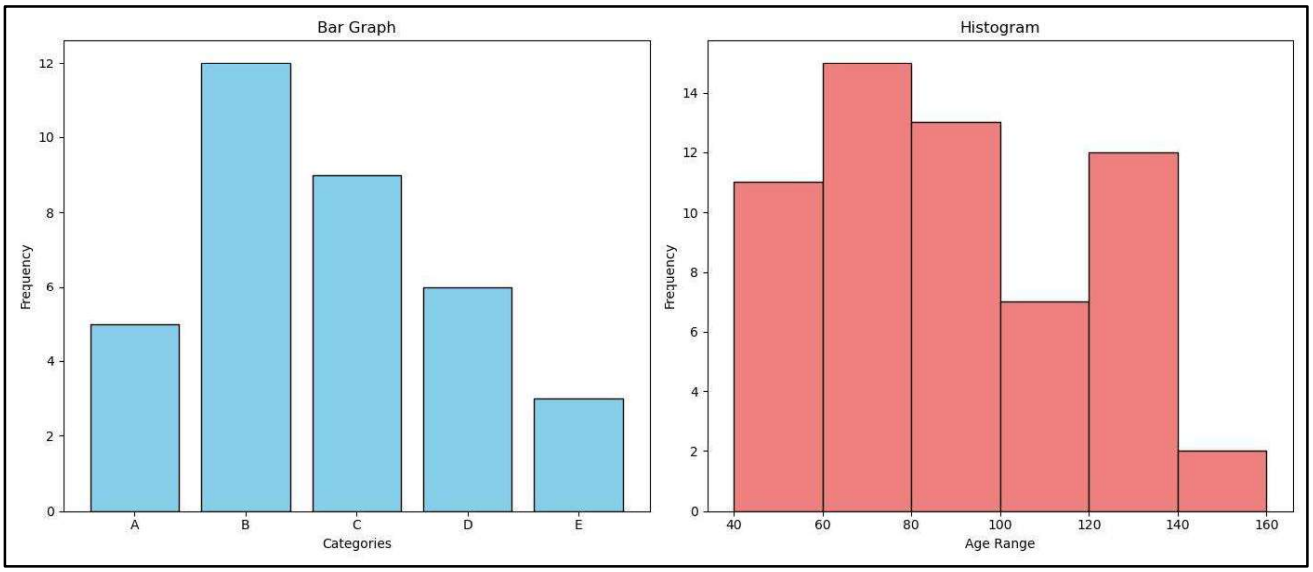
Histogram is a graphical representation of the distribution of numerical data. Overall pattern can be described by its **shape**, **center**, and **spread**. The following age distribution is **right skewed**. The **center** lies between **80 to 100**. **No outliers.**

Age Range	Frequency
40-59	11
60-79	15
80-99	13
100-119	7
120-139	12
140-159	2



comparison between a **Bar Graph** and a **Histogram** in table format:

Feature	Bar Graph	Histogram
Data Type	Categorical (discrete data) A,B,C,D	Continuous (numeric data) 40,60,80,100,
Bars	Separate bars with gaps between them	Adjacent bars without gaps
X-Axis	Represents categories (e.g., colors, names)	Represents ranges or intervals of data
Purpose	Used to compare different categories	Used to show distribution of continuous data
Example	Number of people in different cities	Distribution of ages in a population
Bar Width	All bars have the same width	The width of bars depends on the interval size
Y-Axis	Frequency or count of categories	Frequency or count of data within each interval



4. Methods of central tendency (Mean, Median, Mode).

Central Tendency

Central tendency is defined as "the statistical measure that identifies a single value as representative of an entire distribution or sample.

Central Tendency (কেন্দ্রীয় প্রবণতা) ডেটার মধ্যে একটি সাধারণ মান বা গড় স্থাপন করার প্রক্রিয়া। এটি আমাদের ডেটা সেটের "কেন্দ্র" বা মধ্যবর্তী অবস্থান বুঝতে সাহায্য করে। তিনটি প্রধান পদ্ধতি রয়েছে যার মাধ্যমে আমরা কেন্দ্রীয় প্রবণতা বের করি:

1. Mean (গড়):

গড় বা **Mean** হলো ডেটার সব মান যোগ করে, তারপর সংখ্যা দিয়ে ভাগ করা।

কিভাবে হিসাব করবেন:

- ডেটার সব মান যোগ করুন।
- তারপর সেই যোগফলকে ডেটার মানের মোট সংখ্যা দিয়ে ভাগ করুন।

উদাহরণ: ধরা যাক, ডেটা সেটটি হলো: 2, 4, 6, 8, 10

গড় (Mean) বের করতে:

Mean= \frac{2 + 4 + 6 + 8 + 10}{5} Mean= \frac{30}{5} = 6

Meanএর ব্যবহার: mean ব্যবহার হয় সাধারণত যখন ডেটা সমানভাবে বিতরণ থাকে এবং outlayer (অস্বাভাবিক মান) খুব বেশি প্রভাব ফেলে না।

2. Median (মিডিয়ান):

Median হলো ডেটার মধ্যে **মাঝের মান**। যদি ডেটা সাজানো থাকে, তাহলে মিডিয়ান হবে ডেটার **মাঝের মান**। যদি ডেটার সংখ্যা বিজোড় (odd) হয়, তাহলে মাঝের মান সরাসরি পাওয়া যাবে। আর যদি সংখ্যা জোড়া (even) হয়, তাহলে দুটি মাঝের মানের গড় নিতে হবে।

কিভাবে হিসাব করবেন:

- প্রথমে ডেটা সাজিয়ে নিন।
- তারপর, যদি ডেটার সংখ্যা বিজোড় হয়, মাঝের মানটি হলো \left(\frac{n + 1}{2}\right) তম মান।
- যদি ডেটার সংখ্যা জোড়া হয়, তখন দুইটি মাঝের মানের গড় নিন।

উদাহরণ 1 (বিজোড় সংখ্যা): ধরা যাক, ডেটা সেট: 1, 3, 5, 7, 9

মিডিয়ান হবে 5, কারণ এটি ডেটার মধ্যে মাঝের মান।

উদাহরণ 2 (জোড়া সংখ্যা): ধরা যাক, ডেটা সেট: 1, 3, 5, 7

মিডিয়ান হবে: Median= \left(\frac{3 + 5}{2}\right) = 4

Medianএর ব্যবহার: Median ব্যবহার করা হয় যখন ডেটাতে কিছু অস্বাভাবিক মান (outlayer) থাকে, কারণ Median outlayer দ্বারা প্রভাবিত হয় না।

3. Mode (মোড):

Mode হলো সেই মান যা ডেটার মধ্যে সবচেয়ে বেশি বার পুনরাবৃত্তি হয়। অর্থাৎ, কোন মানটি সবচেয়ে বেশি আসে সেটিই হচ্ছে মোড। Mode: যদি একই সংখ্যা বেশিবার না থাকে, তাহলে mode বা প্রচুরক নাই ।

কিভাবে হিসাব করবেন:

- ডেটার মধ্যে যেই মানটি সবচেয়ে বেশি বার আসবে, সেটি হবে মোড।

উদাহরণ: ধরা যাক, ডেটা সেট: 2, 4, 4, 6, 8, 8, 8

এখানে, 8 সবচেয়ে বেশি (৩ বার) এসেছে, তাই **Mode** হবে 8।

মোডের ব্যবহার: মোড ব্যবহার করা হয় যখন আমরা দেখতে চাই কোন মানটি সবচেয়ে সাধারণ বা সবচেয়ে বেশি ঘটে।

Missing value থাকলে data train করা যায় না । সেজন্য, missing value কে mean, median, mode দিতে হয় ।

- **Mean:** ডেটার সব মান যোগ করে গড় বের করা।
- **Median:** ডেটা সাজানোর পর মাঝের মান বের করা।
- **Mode:** যেই মানটি সবচেয়ে বেশি আসে, সেটি হলো মোড।

5. Methods of dispersion measurement.

Measures of Dispersion (বিচ্ছুরণের পরিমাপ): dispersion পরিমাপ করে যে ডেটার মানগুলি গড় মান থেকে কতটুকু ছড়িয়ে (বা বিচ্যুত) আছে।

Dispersion (বিচ্ছুরণ) ডেটার বিস্তার বা পরিবর্তনশীলতা পরিমাপ করে। এটি ডেটার মধ্যে কতটুকু পার্থক্য বা বিচ্যুতি রয়েছে তা বোঝাতে সাহায্য করে। পাঁচটি প্রধান **dispersion** পরিমাপ পদ্ধতি হলো:

1. **Range:** সর্বোচ্চ এবং সর্বনিম্ন মানের পার্থক্য।
2. **Variance:** গড় থেকে মানগুলির বিচ্যুতি এর বর্গের গড়।
3. **Standard Deviation:** Variance এর বর্গমূল।
4. **Interquartile Range (IQR):** প্রথম এবং তৃতীয় কুয়ারটাইলের পার্থক্য।
5. **Coefficient of Variation (CV):** Standard Deviation এবং Mean এর অনুপাত।

এগুলো ডেটার প্রকৃত বিস্তার বা বিচ্ছুরণ বোঝাতে ব্যবহৃত হয়।

6. Exploring range, variance, and standard deviation.

Variance: mean থেকে প্রতিটি সংখ্যা কত দুরত্বে আছে। তাই হচ্ছে variance.

Standard Deviation: variance কে বর্গমুল করলে Standard Deviation.

Measures of Dispersion (বিচ্ছুরণের পরিমাপের প্রধান পদ্ধতিগুলি):

1. Range (রেঞ্জ):

- **Range** হলো ডেটা সেটের সর্বোচ্চ মান এবং সর্বনিম্ন মানের মধ্যে পার্থক্য।
- এটি খুবই সাধারণ এবং সহজ একটি পদ্ধতি, কিন্তু ডেটার outlayer (অস্বাভাবিক মান) দ্বারা প্রভাবিত হতে পারে।

ফর্মুলা: Range = Maximum Value - Minimum Value

উদাহরণ: ডেটা সেট: 3, 5, 7, 8, 10 Range হবে: 10 – 3 = 7

2. Variance (ভ্যারিয়েন্স):

- **Variance** হলো ডেটার মানগুলির গড় বিচ্যুতি (deviation) এর বর্গের গড়। এটি পরিমাপ করে ডেটার মানগুলো গড় থেকে কতটুকু বিচ্যুত হচ্ছে।
- এটি একটি গুরুত্বপূর্ণ পরিমাপ, তবে এর ইউনিট থাকে ডেটার আসল ইউনিটের বর্গের সমান, যা কিছুটা অসুবিধাজনক হতে পারে।

ফর্মুলা: Variance $= \frac{\Sigma(X_i - \mu)^2}{N}$, যেখানে X_i হলো ডেটার প্রতিটি মান, μ হলো গড় এবং N হলো ডেটার সংখ্যা।

উদাহরণ: ডেটা সেট: 3, 5, 7, 8, 10 গড় $\mu=6.6$

variance হবে:

$$\frac{(3-6.6)^2+(5-6.6)^2+(7-6.6)^2+(8-6.6)^2+(10-6.6)^2}{5}$$

$$=\frac{(12.96+2.56+0.16+1.96+11.56)}{5}=\frac{29.2}{5}=5.84$$

3. Standard Deviation (স্ট্যান্ডার্ড ডেভিয়েশন):

- **Standard Deviation (SD)** হলো **Variance** এর বর্গমূল, যা ডেটার বিচ্ছুরণের পরিমাণ পরিমাপ করে এবং ইউনিট ডেটার আসল ইউনিটে প্রকাশিত হয়, যা বোঝার জন্য সহজ।
- **Standard Deviation** সবচেয়ে বেশি ব্যবহৃত পদ্ধতি, কারণ এটি ডেটার প্রকৃত বিস্তার বা বিচ্ছুরণ বুঝতে সাহায্য করে।

ফর্মুলা: $Standard\ Deviation = \sqrt{\text{Variance}}$

উদাহরণ: আগের উদাহরণের **Variance** ছিল 5.84, তাই **Standard Deviation** হবে: $\sqrt{5.84} \approx 2.42$

4. Interquartile Range (IQR):

- **Interquartile Range (IQR)** হলো ডেটার প্রথম কুয়ারটাইল (Q1) এবং তৃতীয় কুয়ারটাইল (Q3) এর মধ্যে পার্থক্য। এটি ডেটার মধ্যবর্তী 50% এর বিস্তার পরিমাপ করে এবং আউটলাইয়ারগুলিকে এড়িয়ে যায়।
- এটি আউটলাইয়ার প্রভাবিত হওয়ার পরিবর্তে ডেটার প্রকৃত বিস্তার বুঝতে সাহায্য করে।

ফর্মুলা: IQR=Q3-Q1

উদাহরণ: ডেটা সেট: 1, 3, 5, 7, 9 Q1 = 3, Q3 = 7

IQR হবে: 7-3=47 - 3 = 4

5. Coefficient of Variation (CV):

- **Coefficient of Variation (CV)** হলো **Standard Deviation** এবং **Mean** এর অনুপাত, যা ডেটার পরিবর্তনশীলতা অন্য ডেটার তুলনায় ব্যাখ্যা করতে সাহায্য করে।
- এটি পরিমাণগত পরিমাপ, যা বিভিন্ন সেটের মধ্যে বিচ্ছুরণের তুলনা করতে ব্যবহৃত হয়।

ফর্মুলা:

$$CV=\frac{\text{Standard Deviation}}{\text{Mean}}\times 100$$

উদাহরণ: ধরা যাক, ডেটা সেটের গড় (Mean) হলো 50 এবং **Standard Deviation** হলো 5।

তাহলে: $CV=\frac{5}{50}\times 100=10\%$

7. Quartiles, deciles, percentiles, and coefficient of variation.

1. Quartiles (কুয়ারটাইল)

Quartiles হলো ডেটাকে চারটি সমান ভাগে ভাগ করার জন্য ব্যবহৃত পরিমাপ। এটি ডেটার ভ্যালুগুলির বিশ্লেষণ করতে সহায়ক, বিশেষ করে ডেটার অবস্থান বা বিতরণ বোঝার জন্য।

- **Q1 (First Quartile):** ডেটার প্রথম ২৫% মানকে প্রস্থানে ভাগ করার পরে যে মানটি আসে, সেটি প্রথম কুয়ারটাইল (Q1)।
- **Q2 (Second Quartile):** এটি গড় মান বা **Median** (মিডিয়ান) হিসেবে পরিচিত। এটি ডেটার মাঝের মান।
- **Q3 (Third Quartile):** ডেটার ৭৫% মান পর্যন্ত পৌঁছানোর পরে যে মান আসে, সেটি তৃতীয় কুয়ারটাইল (Q3)।
- **Interquartile Range (IQR):** Q3 এবং Q1 এর মধ্যে পার্থক্য, অর্থাৎ: IQR=Q3-Q1

উদাহরণ: ডেটা: 1, 3, 5, 7, 9

- Q1=3, Q2=5 (মিডিয়ান), Q3=7
- IQR = Q3-Q1=7-3=4

2. Deciles (ডেসাইল)

Deciles হলো ডেটাকে ১০টি সমান ভাগে ভাগ করার জন্য ব্যবহৃত পরিমাপ। এর মধ্যে ৯টি পয়েন্ট রয়েছে, যেগুলি ডেটার ১০%, 20%, ... 90% মানের কাছাকাছি অবস্থান নির্দেশ করে।

- **D1 (First Decile):** ডেটার ১০% জায়গা বা প্রথম ১০% মান।
- **D2 (Second Decile):** ২০% জায়গা বা প্রথম ২০% মান।
- ...
- **D9 (Ninth Decile):** ৯০% জায়গা বা প্রথম ৯০% মান।

উদাহরণ: ডেটা: 1, 3, 5, 7, 9

- **D1 = 3, D5 = 5, D9 = 9** হতে পারে, যদি ডেটা সঠিকভাবে ভাগ করা যায়।

3. Percentiles (পার্সেন্টাইল)

Percentiles ডেটাকে ১০০টি সমান ভাগে ভাগ করে, যাতে ১টি পার্সেন্টাইল প্রতিটি ভাগের সীমানা নির্দেশ করে। এটি **অন্তর্নিহিত** ডেটা বৈশিষ্ট্য এবং পারফরম্যান্স বিশ্লেষণ করতে সাহায্য করে, বিশেষ করে শিক্ষা, স্বাস্থ্য এবং সামাজিক বিজ্ঞান শাখায়।

- **P1 (1st Percentile):** ১% জায়গা বা প্রথম ১% মান।
- **P50 (50th Percentile):** এটি **Median** (মিডিয়ান) এর সমান।

- **P99 (99th Percentile):** ৯৯% জায়গা বা প্রথম ৯৯% মান।

উদাহরণ: ডেটা: 1,3,5,7,91, 3, 5, 7, 9

- **P1 = 1.5, P50 = 5, P99 = 9** হতে পারে, যদি সঠিকভাবে হিসাব করা হয়।

4. Coefficient of Variation (CV) - ভেরিয়েন্সের সহগ

Coefficient of Variation (CV) হলো **Standard Deviation** (স্ট্যান্ডার্ড ডেভিয়েশন) এবং **Mean** (গড়) এর অনুপাত, যা ডেটার পরিবর্তনশীলতা বা বিচ্যুতি পরিমাপ করতে ব্যবহৃত হয়। এটি ডেটার বিস্তার বা পরিবর্তনশীলতার তুলনা করতে খুবই কার্যকর, বিশেষত যখন ডেটা সেটের একাধিক সমান প্রকৃতি থাকে।

ফর্মুলা:

$$Coefficient\ of\ Variation(CV) = \frac{\text{Standard Deviation (SD)}}{\text{Mean } (\mu)} \times 100$$

উদাহরণ: ধরা যাক, ডেটা সেটের গড় (Mean) হলো 50 এবং **Standard Deviation** হলো 5। তাহলে:

$$CV = \frac{5}{50} \times 100 = 10\%$$

8. Five-number summary and creating a box plot. **Assignment #17:**

Five-Number Summary (পাঁচটি সংখ্যার সারাংশ)

Five-Number Summary হলো ডেটার ৫টি মূল সংখ্যার একটি সারাংশ, যা ডেটার বর্ণনা বা বিস্তার বোঝাতে সাহায্য করে। এটি বিশেষভাবে **Box Plot** তৈরি করতে ব্যবহার করা হয়।

চলুন, **Five-Number Summary** (পাঁচটি সংখ্যার সারাংশ) ১৮টি সংখ্যা দিয়ে দেখি।

ধরা যাক, আমাদের ডেটাসেটটি হল:

ডেটা: 2,4,6,7,9,10,12,14,16,18,20,22,24,25,26,28,30,32

1. **Minimum (সর্বনিম্ন):** এটি ডেটার সবচেয়ে ছোট মান। আমাদের ডেটাতে সবচেয়ে ছোট মান হল 2। তাই, Minimum = 2।
2. **First Quartile (Q1):**
   - ডেটার ২৫% মানের কাছাকাছি প্রথম কুয়ারটাইল। এটি ডেটার প্রথম ২৫% ভাগের শেষ মান। Q1 হলো প্রথম কুয়ারটাইল। আমাদের ১৮টি সংখ্যা রয়েছে, সুতরাং, প্রথম ৯টি সংখ্যা থেকে মধ্য মান বের করা হবে। $18/4=4.5 \approx 5$
   - প্রথম ৯টি সংখ্যাঃ 2,4,6,7,9,10,12,14,16
   - এর মধ্যে Median বা মধ্য মান হলো 9। তাই, Q1 = 9।
3. **Median (Q2):**

- এটি ডেটার মাঝের মান বা **Second Quartile (Q2)**। এটি ডেটার গড় মান। ডেটা সেটে মোট ১৮টি (জোড়) সংখ্যা রয়েছে, তাই Median এর জন্য ৯তম এবং ১০ম সংখ্যার গড় নেওয়া হবে।
- ৯ম এবং ১০ম সংখ্যা: 16, 18
- Median (Q2) = $\frac{16+1}{2}$ = 17

4. **Third Quartile (Q3):**
   - ডেটার ৭৫% মানের কাছাকাছি তৃতীয় কুয়ারটাইল। এটি ডেটার শেষ ২৫% ভাগের প্রথম মান। Q3 হলো তৃতীয় কুয়ারটাইল। ডেটার পরের ৯টি সংখ্যা থেকে মধ্য মান বের করা হবে। 18/4 * 3 = 4.5 * 3 = 13.5 ≈ 13.4
   - পরের ৯টি সংখ্যাঃ 18,20,22,24,25,26,28,30,32
   - এর মধ্যে Median বা মধ্য মান হলো 25। তাই, Q3 = 25।
5. **Maximum (সর্বাধিক):** এটি ডেটার সবচেয়ে বড় মান। আমাদের ডেটাতে সবচেয়ে বড় মান হল **32**। তাই, **Maximum = 32**।

উদাহরণ: ডেটা: 2,4,6,7,9,10,12,14,16,18,20,22,24,25,26,28,30,32

- **Minimum** = 2
- **Q1 (First Quartile)** = $\frac{18}{4} \times 1 = 4.5 \approx 5^{th}$ মান হচ্ছে 9
- **Median (Q2)** = (16+18)/2 = 17
- **Q3 (Third Quartile)** = $\frac{18}{4} \times 3 = 13.5 \approx 14^{th}$ মান হচ্ছে 25
- **Maximum** = 32

এই 18টি সংখ্যার মাধ্যমে ডেটার বিস্তার এবং পার্থক্য বুঝতে সহজ হয়।

Box Plot (বক্স প্লট)

একটি **Box Plot** তৈরি করতে, নিচের ধাপগুলো অনুসরণ করতে হবে:

1. ডেটা সাজানো:

- প্রথমে ডেটাকে **ছোট থেকে বড়** অনুযায়ী সাজান।
- যেমন, আমাদের উদাহরণ ডেটাসেট ছিল: 2, 4, 6, 7, 9, 10, 12, 14, 16, 18, 20, 22, 24, 25, 26, 28, 30, 32

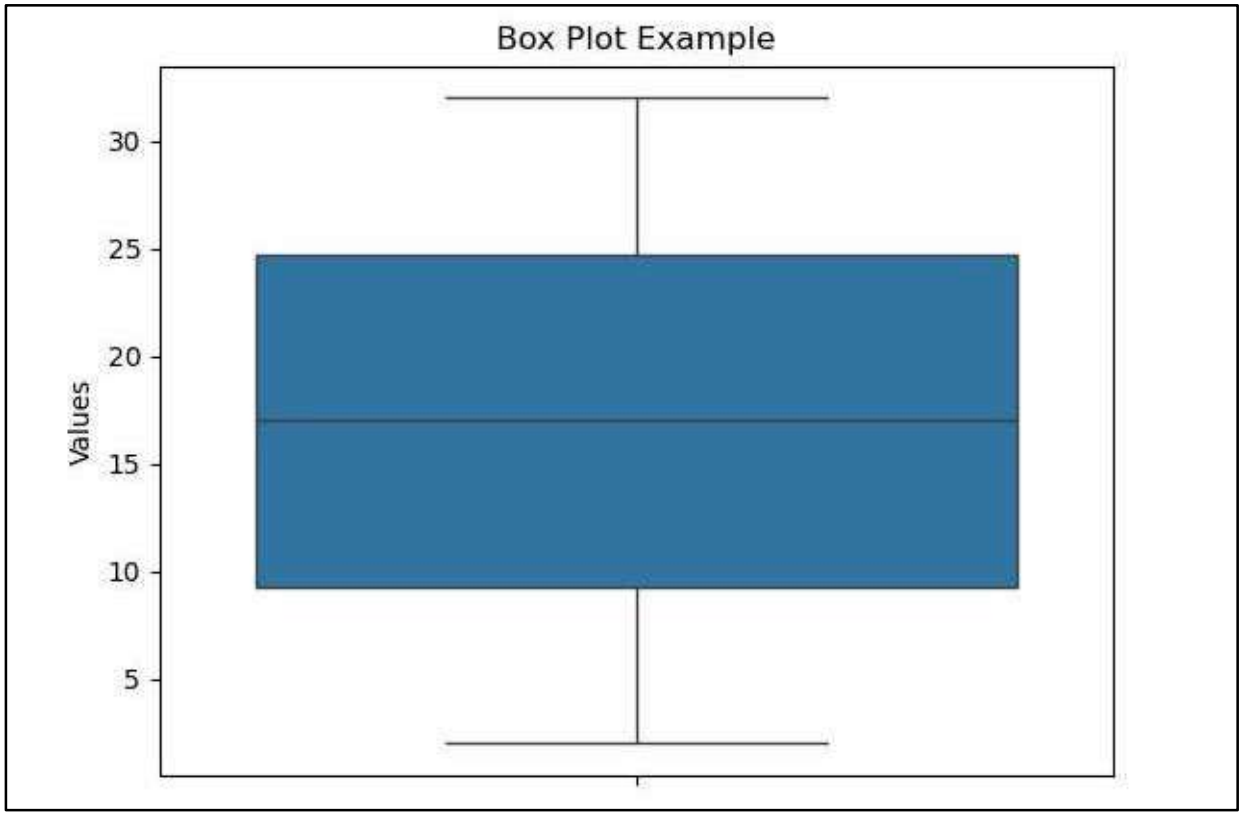
2. Five-Number Summary বের করা:

- **Minimum** = 2
- **Q1 (First Quartile)** = $\frac{18}{4} \times 1 = 4.5 \approx 5^{th}$ মান হচ্ছে 9
- **Median (Q2)** = (16+18)/2 = 17

- **Q3 (Third Quartile)** = $\frac{18}{4} \times 3 = 13.5 \approx 14^{th}$ মান হচ্ছে 25
- **Maximum** = 32

3. Box Plot এর বক্স তৈরি করা:

- **বক্স**: Q1 এবং Q3 এর মধ্যে একটি বক্স আঁকুন। এর মানে হচ্ছে বক্সের নিচের প্রান্ত হবে **Q1** (9) এবং উপরের প্রান্ত হবে **Q3** (25)।
- বক্সের মধ্যে একটি লাইন আঁকুন যা **Median (Q2)** কে চিহ্নিত করবে। এখানে **Median** (Q2) হবে **17**। বক্সের মধ্যে এটি একটি সোজা রেখা হিসাবে প্রদর্শিত হবে।



এখন, **Box Plot** এ:

- **বক্স** হবে 9 থেকে 25 পর্যন্ত ।
- **Median (17)** বক্সের মাঝখানে থাকবে ।
- **Whiskers** হবে 2 (Minimum) এবং 32 (Maximum) এর মধ্যে ।
- আউটলাইয়ার নেই, কারণ সব মান **Whiskers** এর মধ্যে রয়েছে ।

4. Whiskers আঁকানো:

- **Whiskers** হলো বক্সের বাইরে দুটি রেখা যা ডেটার **Minimum** এবং **Maximum** মানকে চিহ্নিত করে।
- **Minimum** (2) থেকে বক্সের নিচ পর্যন্ত একটি রেখা আঁকুন।

- **Maximum** (32) থেকে বক্সের উপরের প্রান্ত পর্যন্ত একটি রেখা আঁকুন।
- Whiskers এর সাহায্যে আমরা ডেটার বিস্তার বা রেঞ্জ বুঝতে পারি।

5. Outliers চিহ্নিত করা (যদি থাকে):

- যদি কোনো মান **Whiskers** এর বাইরে চলে যায়, তাহলে সেটি **Outlier** (আউটলাইয়ার) হিসাবে চিহ্নিত হবে। আউটলাইয়ার সাধারণত বক্স প্লটে একটি আলাদা চিহ্ন বা পয়েন্ট দিয়ে দেখানো হয়।
- আউটলাইয়ার চিনে রাখার জন্য সাধারণত **1.5 IQR (Interquartile Range)** নিয়ম ব্যবহার করা হয়। যেখানে:
    - **IQR = Q3 - Q1 = 25-9 = 16**
    - আউটলাইয়ার হচ্ছে এমন মান, যা Q1−1.5×IQR অথবা Q3+1.5×IQR এর বাইরে চলে যায়।

      (9-1.5×16) অথবা (25+1.5×16)

      -15 অথবা 49

Box Plot এর অংশগুলো:

1. **Box:** Q1 এবং Q3 এর মধ্যে একটিভ বক্স, যা ৫০% ডেটার বিস্তার দেখায়।
2. **Whiskers:** বক্সের বাইরে দুটি রেখা, যা **Minimum** এবং **Maximum** মান দেখায়।
3. **Median (Q2):** বক্সের মধ্যে একটি লাইন, যা ডেটার মাঝের মান চিহ্নিত করে।
4. **Outliers:** ডেটার এমন মান যা বক্স এবং whiskers এর বাইরে থাকে, এবং বিশেষ চিহ্ন (যেমন পয়েন্ট) দিয়ে চিহ্নিত হয়।

Box Plot এবং Five-Number Summary এর সম্পর্ক:

- **Box Plot** দ্বারা আমরা **Five-Number Summary** দেখাতে পারি এবং ডেটার বিস্তার, স্কিউনেস (skewness), এবং আউটলাইয়ারের উপস্থিতি সহজেই বিশ্লেষণ করতে পারি।
- **Box Plot** হল একটি খুবই কার্যকরী টুল, বিশেষ করে ডেটার বিভাজন এবং বৈশিষ্ট্যসমূহ বিশ্লেষণ করতে।

সারাংশ:

- **Five-Number Summary:** ডেটার পাঁচটি মূল পরিমাপ (Minimum, Q1, Median, Q3, Maximum)।
- **Box Plot:** একটি গ্রাফিক্যাল উপস্থাপনা যা **Five-Number Summary** কে বক্স এবং রেখা আকারে দেখায় এবং আউটলাইয়ার চিহ্নিত করতে সাহায্য করে।