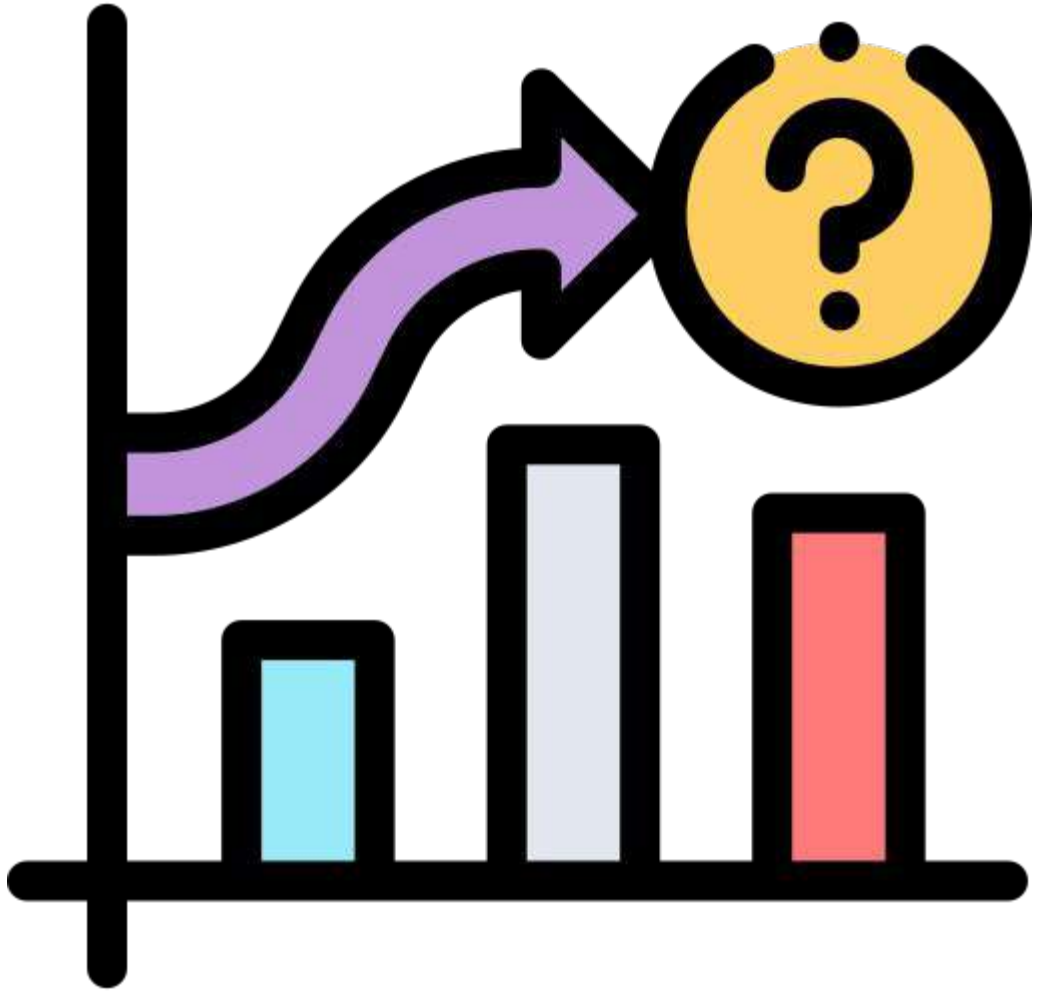
Measure of Central Tendency & Dispersion with various Chart Types



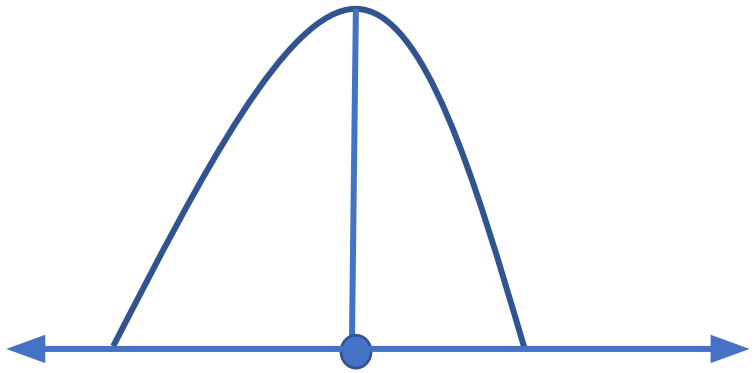
Dr. Aanchal Anant Awasthi, Ph.D.
https://www.youtube.com/c/sscrindia

Content

Central Tendency	Dispersion	Charts/Graphs
Mean Median Mode	Range Inter Quartile Range Quartile Deviation Variance Standard Deviation	Pie Chart Bar Diagram Clustered Bar Diagram Histogram Line Chart Box and Whisker Plot Scatter Plot

Central Tendency

- It’s a quality of a data set to cluster around some value
- This value is known as “Central Value”



Central Value

Central Tendency

Example

Sample: 300 women to create awareness about cancer screening

- Average age of our sample=55 years

This number is nothing but central value.

Measure of Central Tendency

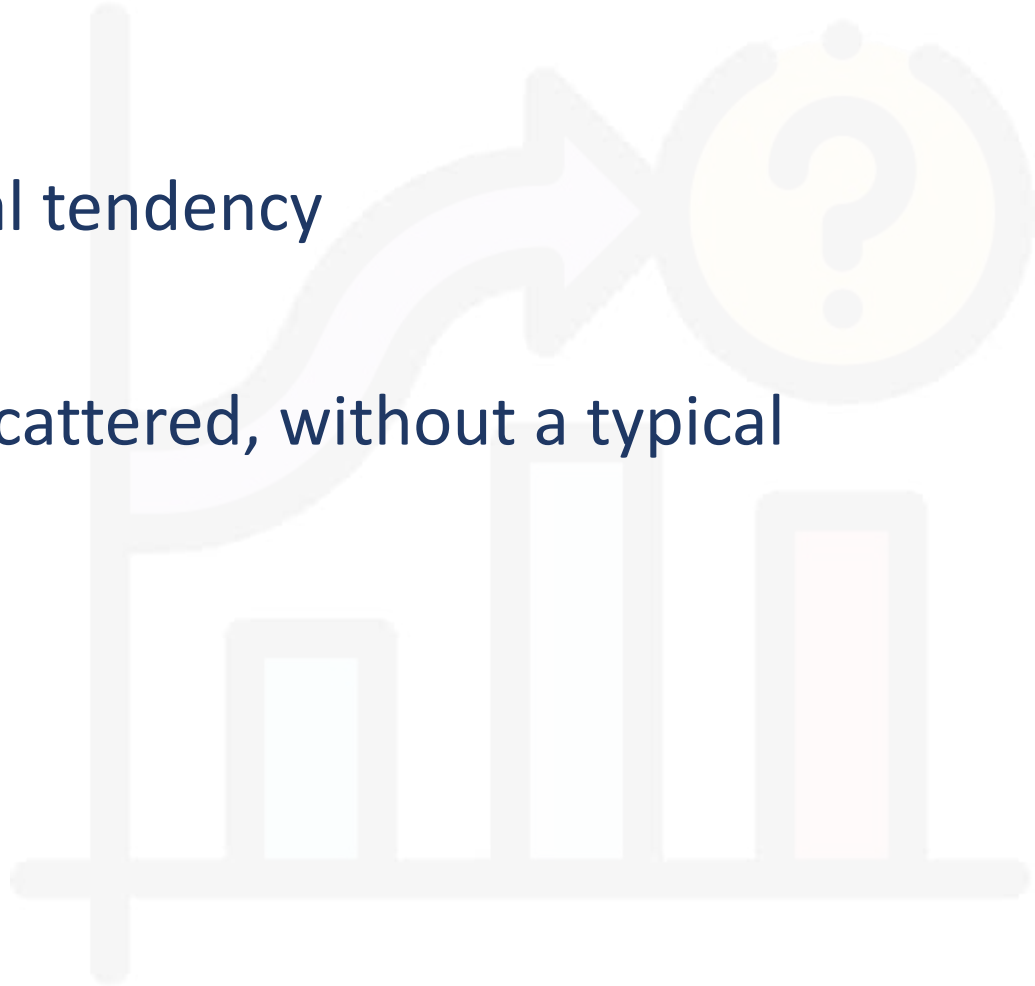
Mean

Median

Mode

Arithmetic Mean

- Most commonly used measure of central tendency
- It’s most important value when data is scattered, without a typical pattern.



Mean = Sum of All Observations / Total Number of Observation

50 60 70 80 90

$$\bar{x} = \frac{1}{N}\Sigma f x$$

$$\bar{x} = {}^{(50+60+70+80+90)}/_{5}$$

$$\bar{x} = 70$$

Interpretation: On the basis of given data, mean value is computed as 70.

Characteristics of Mean

Merit	Demerit
It is unique	When we have extreme or scarcely representative values (too big or too small), the mean may not be representative
It's easy to calculate & understand	It can't be calculated in case of open ended intervals
If we multiply all the values by a constant, the mean is also multiplied by the same constant	It can't be use if we are dealing with qualitative data
If a constant is summed to each value, the mean is summed in that constant also	
It's least affected by sampling fluctuations	

Median

The median is the **middle observation** in a set of observations that have been ranked in numerical order

Calculation of Median

Size of the sample=Odd Number

- Median=((n+1)/2)th term

30	19	15	22	21	24	28

Calculation of Median

In cases where there are an **even number** of observations, the median lies between the **two middle observations**,

Median is the value of the midpoint between those observations

Calculation of Median

Sample Size=Even Number

- Median = Average of $(n/2)^{th}$ term and $((n/2)+1)^{th}$ term

30	19	15	23	24	28

Characteristics of Median

Merit	Demerit
It's easy to understand & calculate	In case of even number of cases median can not be calculated exactly
It's not affected by extreme values	It's not based upon all observation
It can be calculated for distribution having open ended classes (<35 years, 65 years & above)	
It can be calculated in case of qualitative data arranged in ascending or descending order	

Mode

- The mode is usually defined as the most frequent value

Application of Mode

- Which t-shirt size sale most?

- Which Method of payment is preferred by youths?

(credit card/debit card/ net banking/ UPI Apps/ Cash)

Steps for computation of mode

- Determine all distinct values of characteristic under study
- Count frequency for each distinct value of that characteristics
- The most frequent value is mode

Characteristics of Mode

Merit	Demerit
Its easy to understand and calculate	It can be more than one
It can be calculated in case of categorical data	Its not based upon all the observations
It can be calculated in case of open ended intervals	Sometimes mode can't exist

Dispersion

It measures the extent of variation in a dataset

Data Set 1(n=12): 21, 21, 24, 20, 20,21,20, 24, 20, 25, 23, 24

Data Set 2(n=12): 11, 5, 15, 10, 20, 40, 60, 50, 80, 100, 55, 10

Location + Dispersion = Better Insights

- Consider a data set 1:

1, 3, 5, 7, 9

- Consider another data set 2:

1, 5, 9

- **Mean=5**
- **Dispersion?**

Location + Dispersion = Better Vision

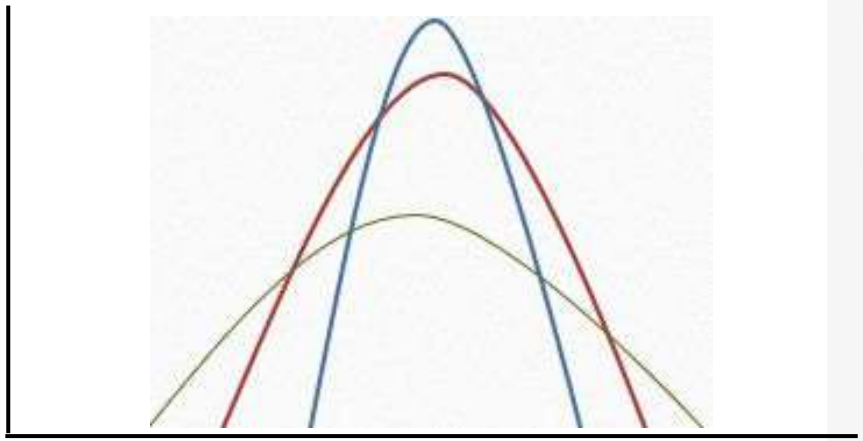


Figure 1: Distribution of Data: Dispersion

Importance of Dispersion

- Dispersion helps us to understand distribution of data
- Gives better insights about data
- It forms basis for statistical theory

Absolute Measures of Dispersion

- Range
- Quartile Deviation
- Inter Quartile Range
- Variance
- Standard Deviation

Range

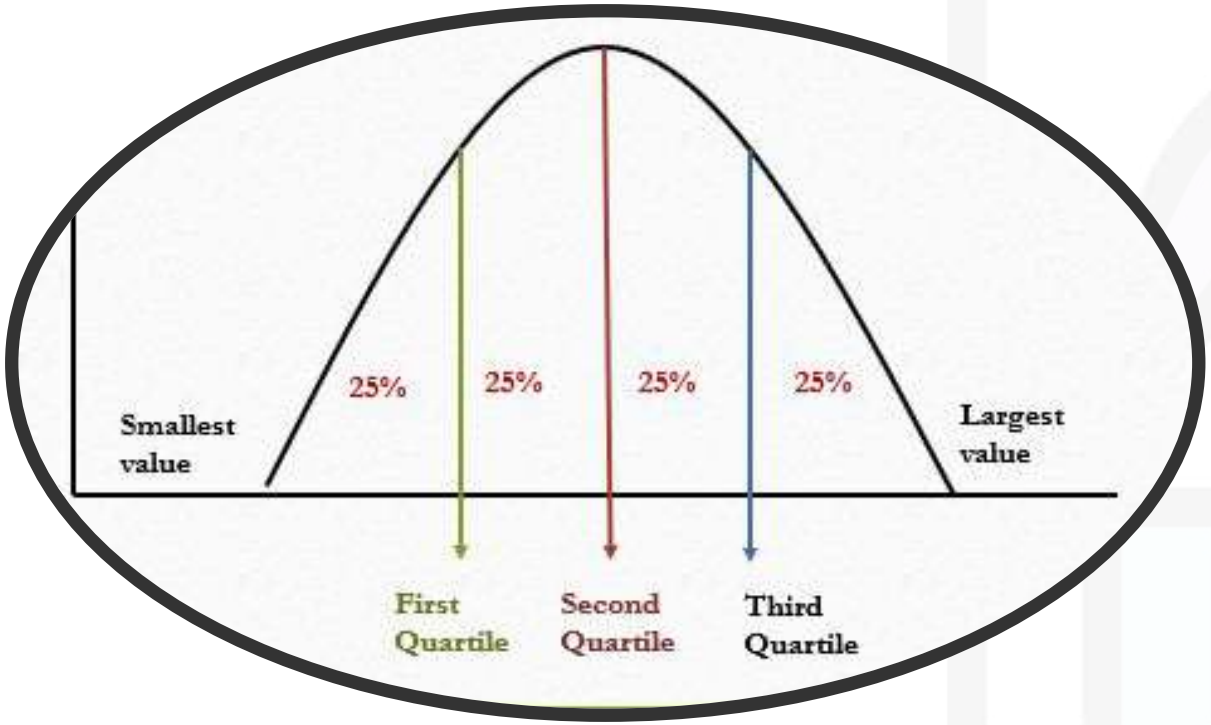
- It’s the difference between largest & smallest value in a data set
- Consider a dataset of time to get ready in the morning for 5 days
- Range = 40-25Minutes=15 Minutes

Day	1	2	3	4	5
Time (minutes)	25	39	35	40	31

Limitation of Range

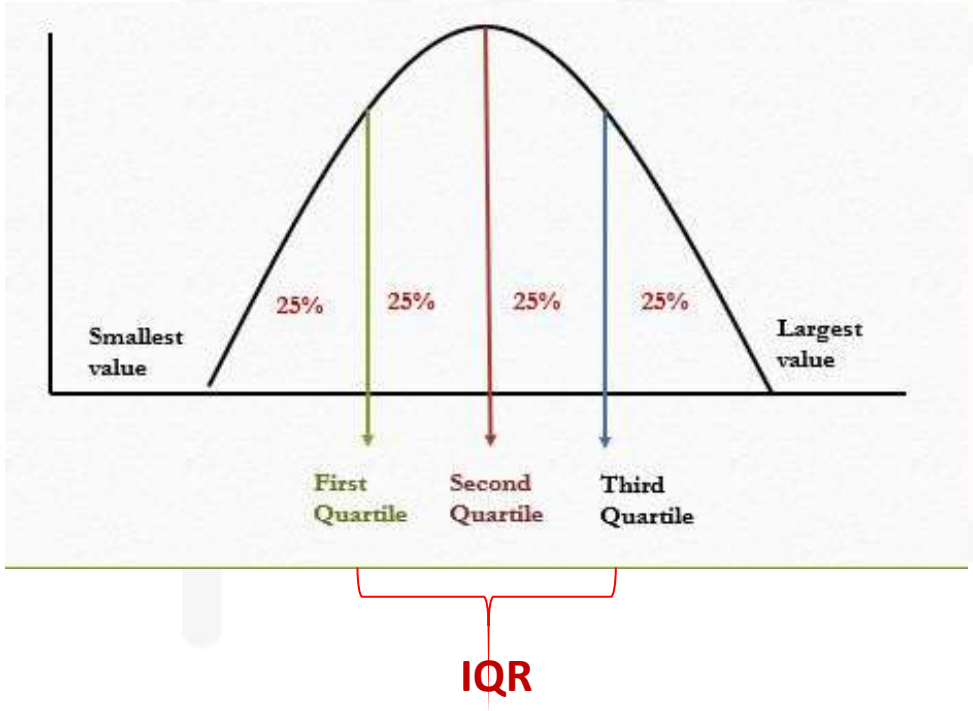
It does not take into account how the values are distributed between the smallest & largest values.

Quartiles



Inter Quartile Range (Concept)

- IQR=Third Quartile – First Quartile



Variance & Standard Deviation

- It conveys how widely or tightly the observations are distributed from the center
- Average variation of datapoints around central value
- Variance / Standard deviation is widely reported along with mean.

Variance & Standard Deviation

- Average variation of datapoints around central value

Weight (Kg)	Datapoint-Mean Weight (Kg)	(Datapoint-Mean Weight)^2
1		
3		
5		
7		
9		
Total		

Formula for Variance and standard deviation

$$\sigma^2 = \frac{\sum (x_i - \bar{x})^2}{n}$$

$$\sigma = \sqrt{\frac{\sum (x_i - \bar{x})^2}{n}}$$

Charts/Graphs

- Charts allows us to represent complex data in a way that is easier to understand and interpret.

Objectives

- Distribution
- Composition
- Comparison
- Relationship between variables

Most Common Graphs

- Pie-Chart
- Bar chart
- Line Graph
- Histogram
- Box-Whisker Plot
- Scatter Plot

Pie Charts

- Pie graphs shows parts or percentages of a whole

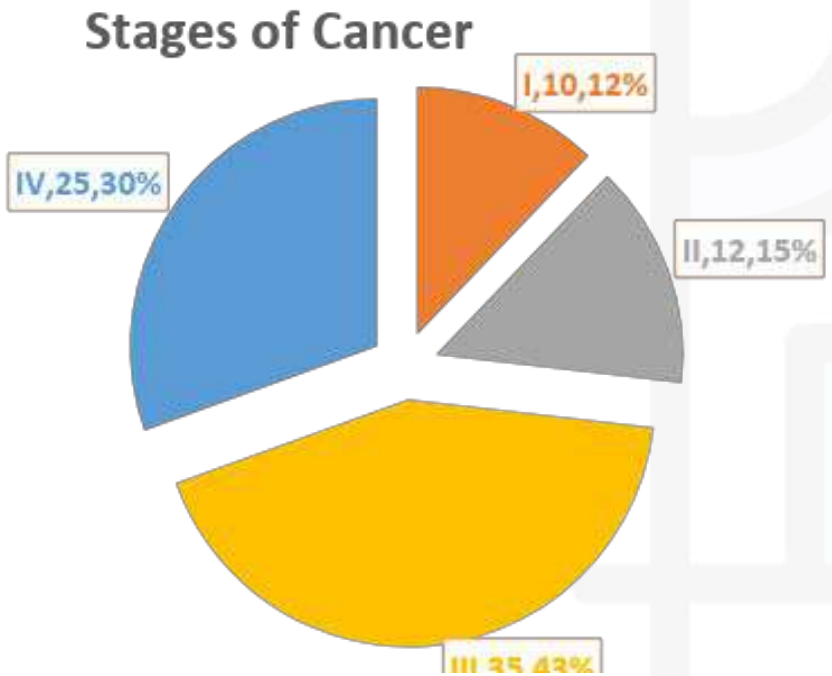


Figure 1: Distribution of Cancer staging in a cancer hospital of North India

Limitations

- Difficult to visualize the differences between estimates of almost similar size.

- Pie graphs simply don’t work when comparing data.

Bar Charts

Vertical Bar Graphs

- Using vertical bars going up from bottom
- Length are proportional to quantities they represent
- Vertical bar graphs are best when we have only few groups/categories to trepresent on chart

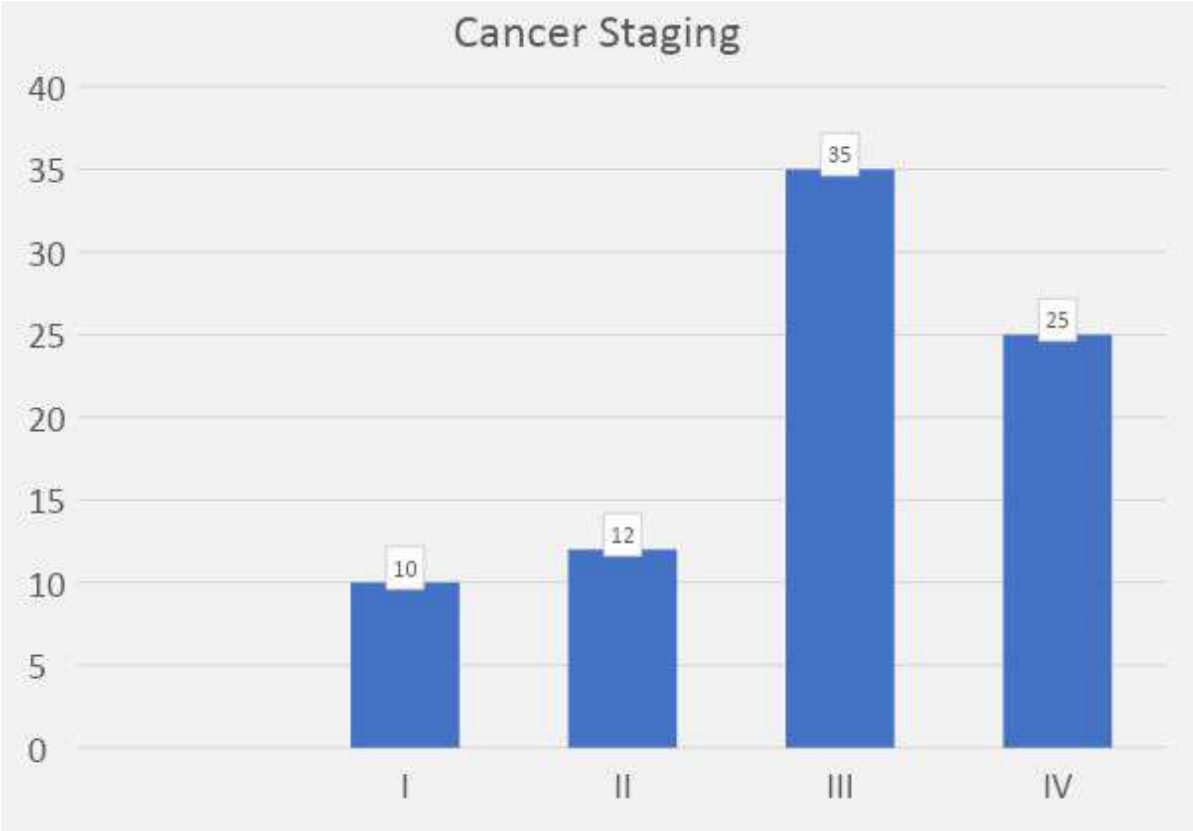


Figure 2: Distribution of oral cancer staging in a cancer hospital of North India

Bar Charts

Horizontal Bar Graphs

- These are the same as vertical bar graphs, but turned on their side
- Horizontal bar graphs are best to use when we have several groups to represent
- Horizontal bar graphs are also appropriate to use when the category labels are too long to appear neatly on the x-axis

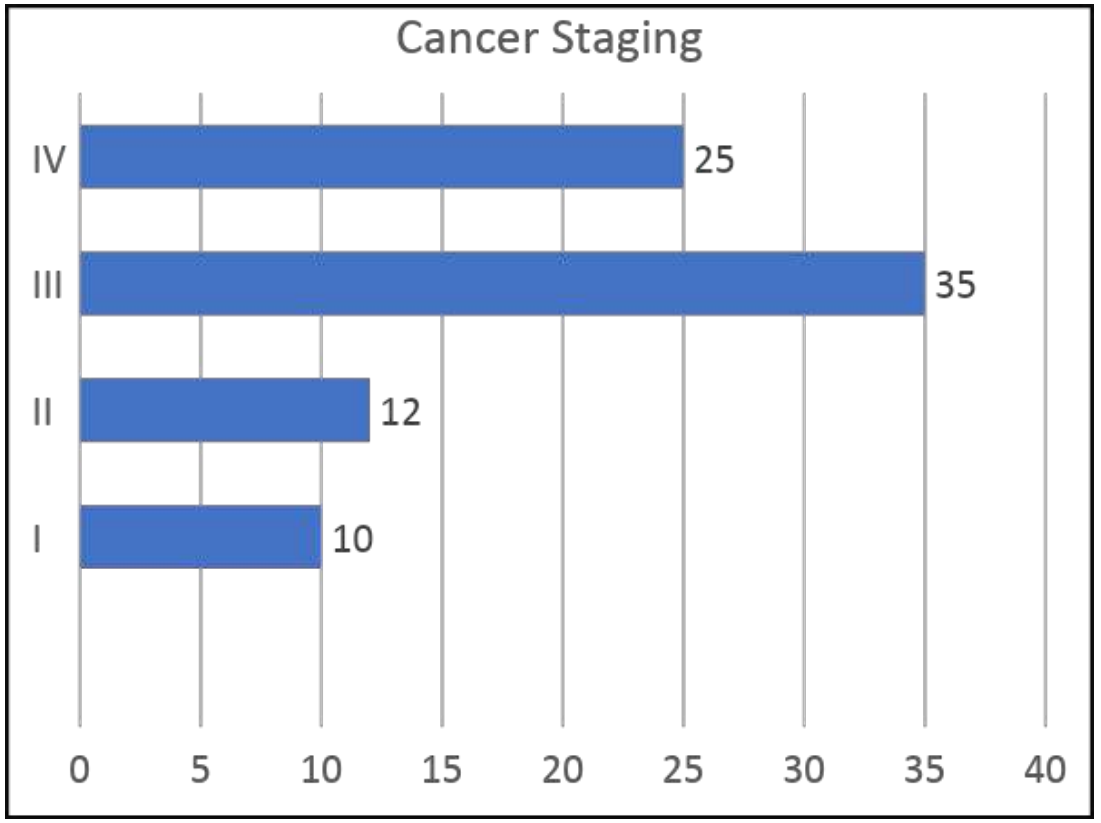


Figure 3: Distribution of oral cancer staging in a cancer hospital of North India

Bar Charts

Clustered Bar Graphs

- Clustered or grouped bar graphs are bar graphs that show two or more categories on one graph

- Plotting multiple categories on one graph increases the amount of information

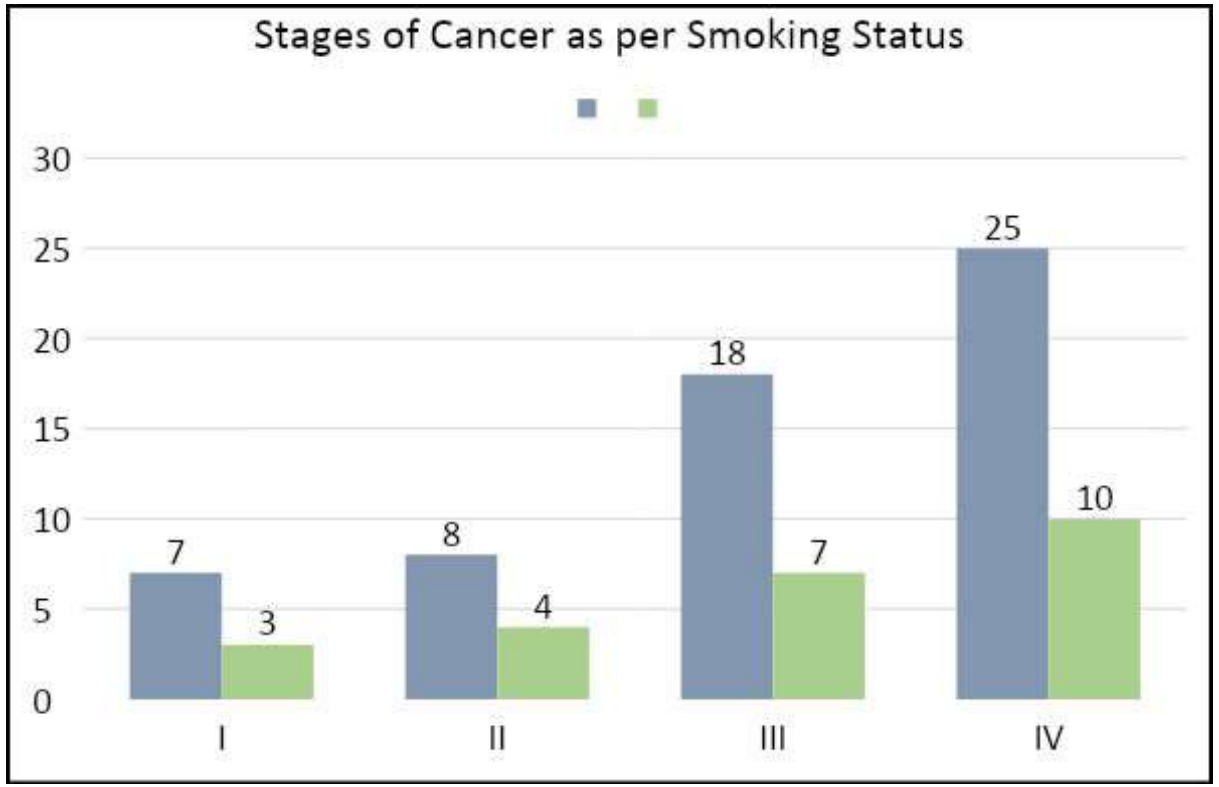


Figure 4: Distribution of Cancer staging along with smoking status in a Cancer Hospital of South India

Line Graphs

- Used to illustrate trends over time for continuous data
- They can also be used to compare two variables over time

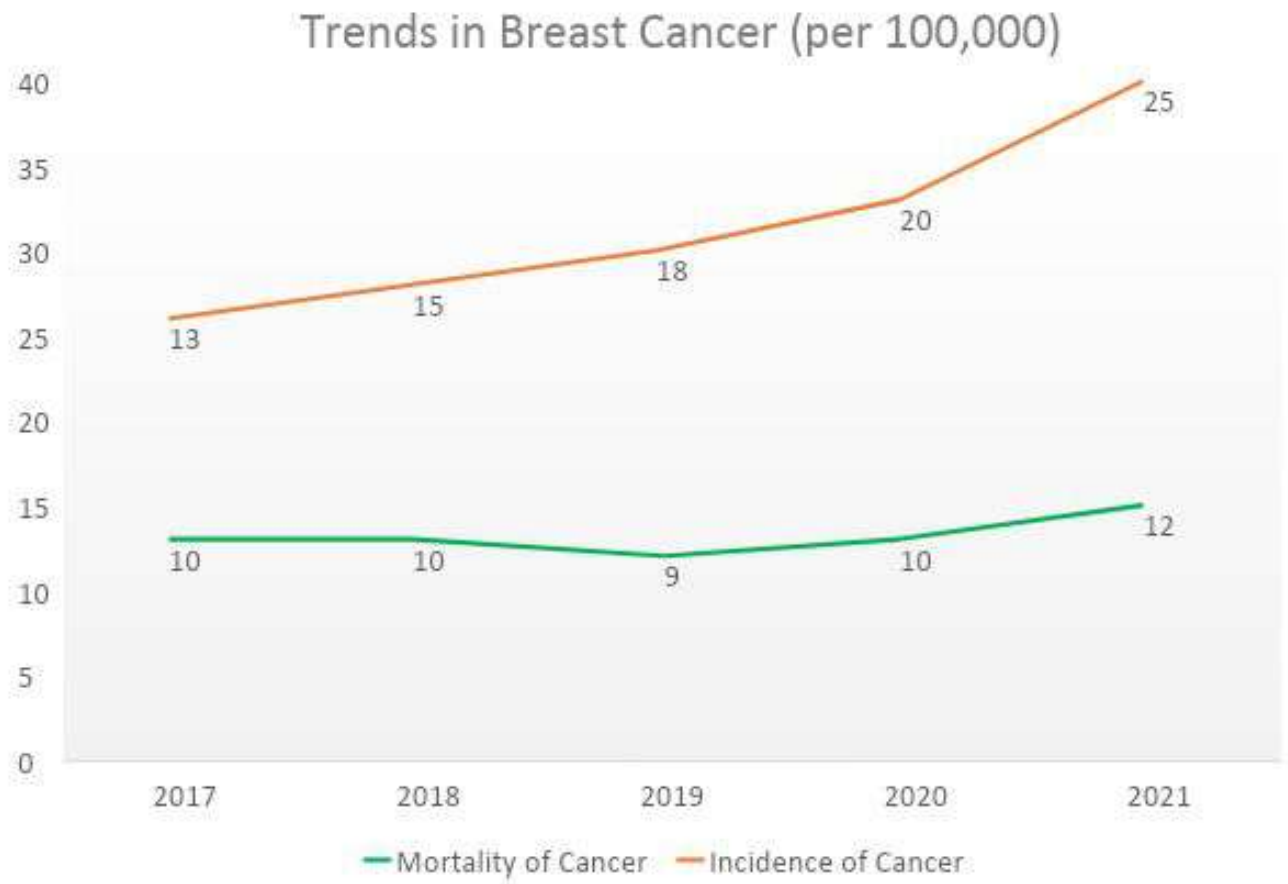


Figure 10:Trends in Breast Cancer (per 100,000)

Histogram

- A histogram shows the underlying frequency distribution (shape) of a set of continuous data
- Data should be grouped into exclusive ranges
- They are connected bars
- The width of each bar is proportional to the width of each category, and the height is proportional to the frequency of that category.

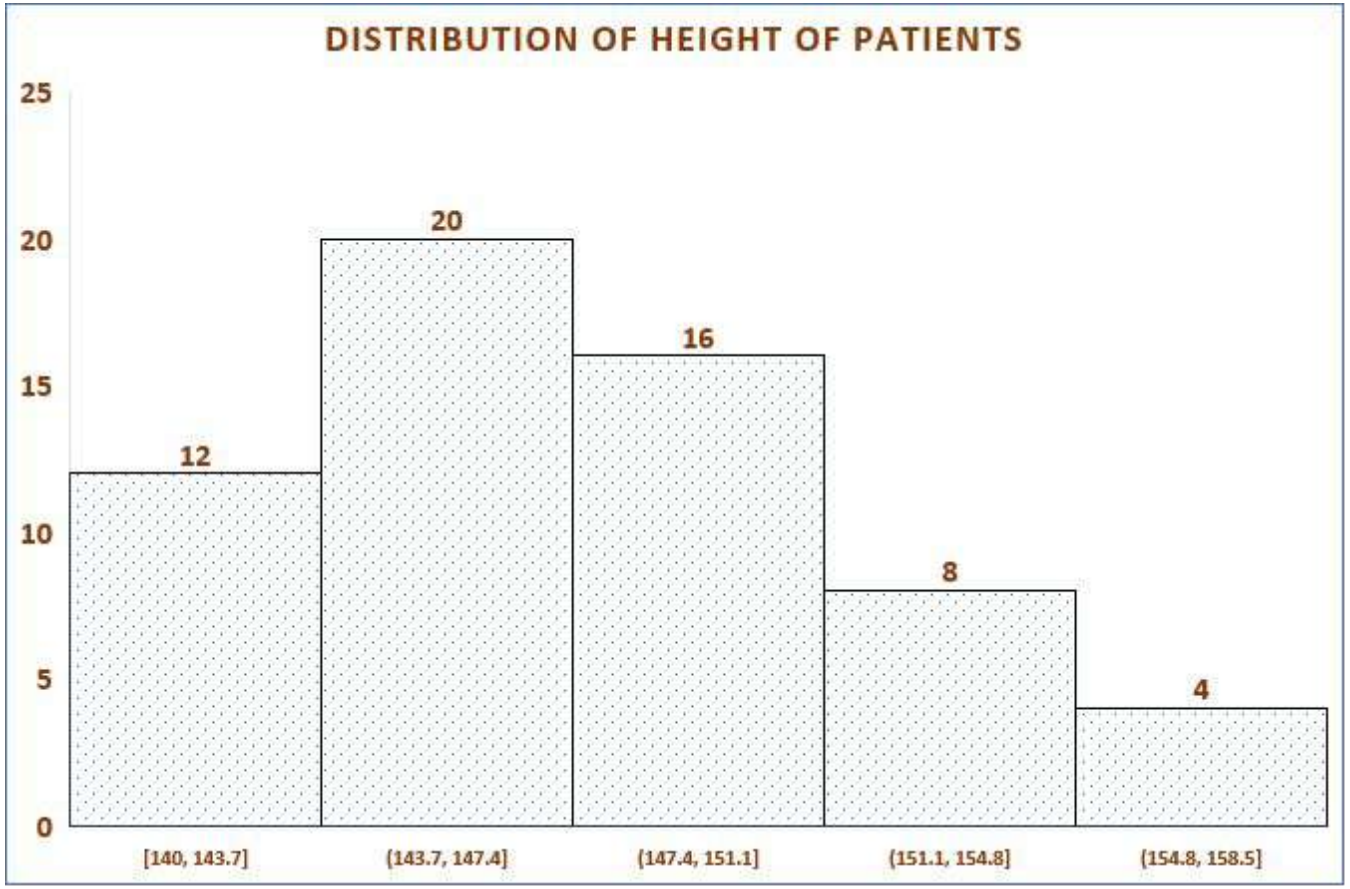


Figure 11: Distribution of heights of Patients

Difference between Bar diagram & Histogram

Bar Diagram

- Gaps between bars are immaterial
- Plots categorical data
- Bars can be reordered
- Height of the bar represents frequency
- Width of the bar is immaterial

Histogram

- Bars are adjacent to each other
- Shows frequency distribution of numerical data
- Bars can not be reordered
- Height of the bar is proportion to frequency
- Width of the bar is equal to interval range

Box Whisker Plot

- Often used in exploratory data analysis

- Five number summary:

    - the minimum value
    - the lower quartile
    - the median value
    - the upper quartile
    - The maximum value

Box Whisker Plot

Boxplots show robust measures of location and spread as well as providing information about symmetry and outliers

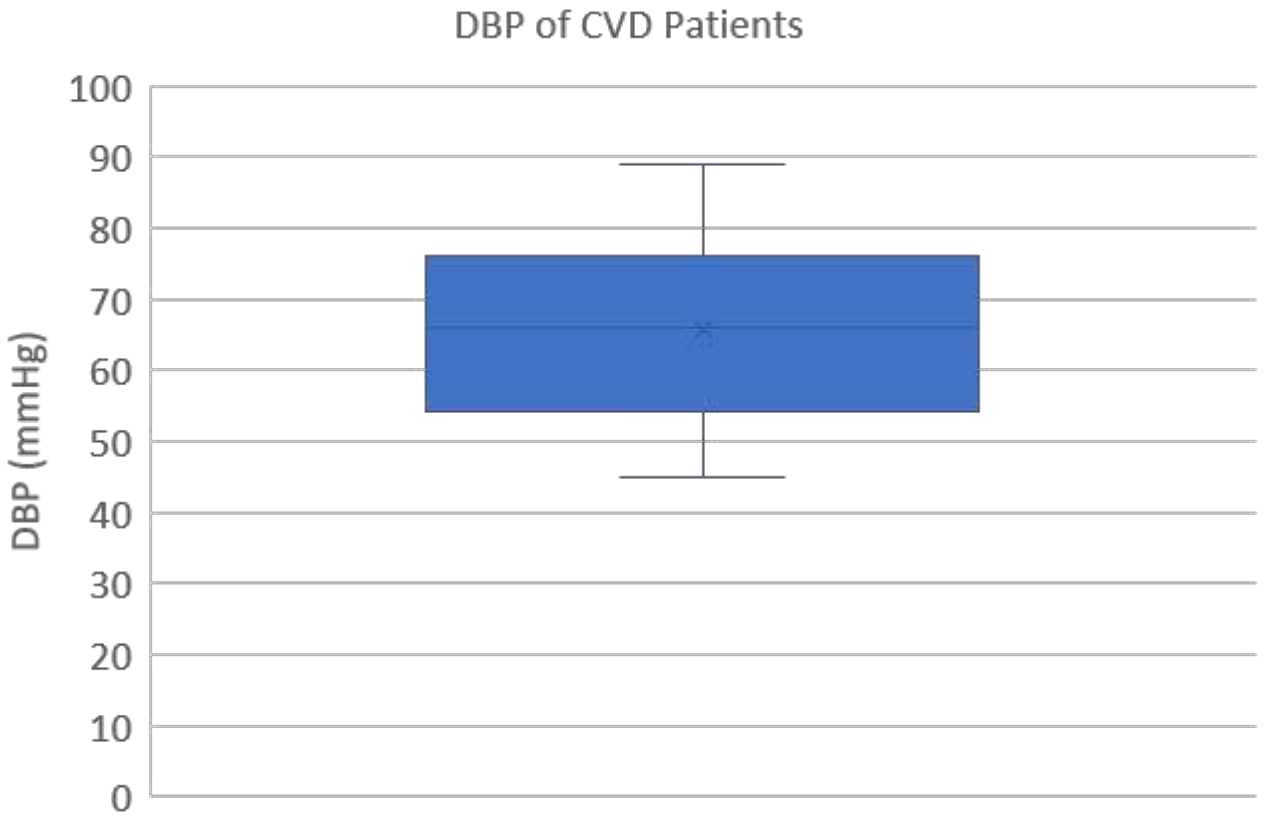


Figure 14: Distribution of Weight of Breast cancer patients

Scatter Plot

- It's used to study the relationship between two graphs
- The pattern of the resulting points represents the correlation between two variables under study

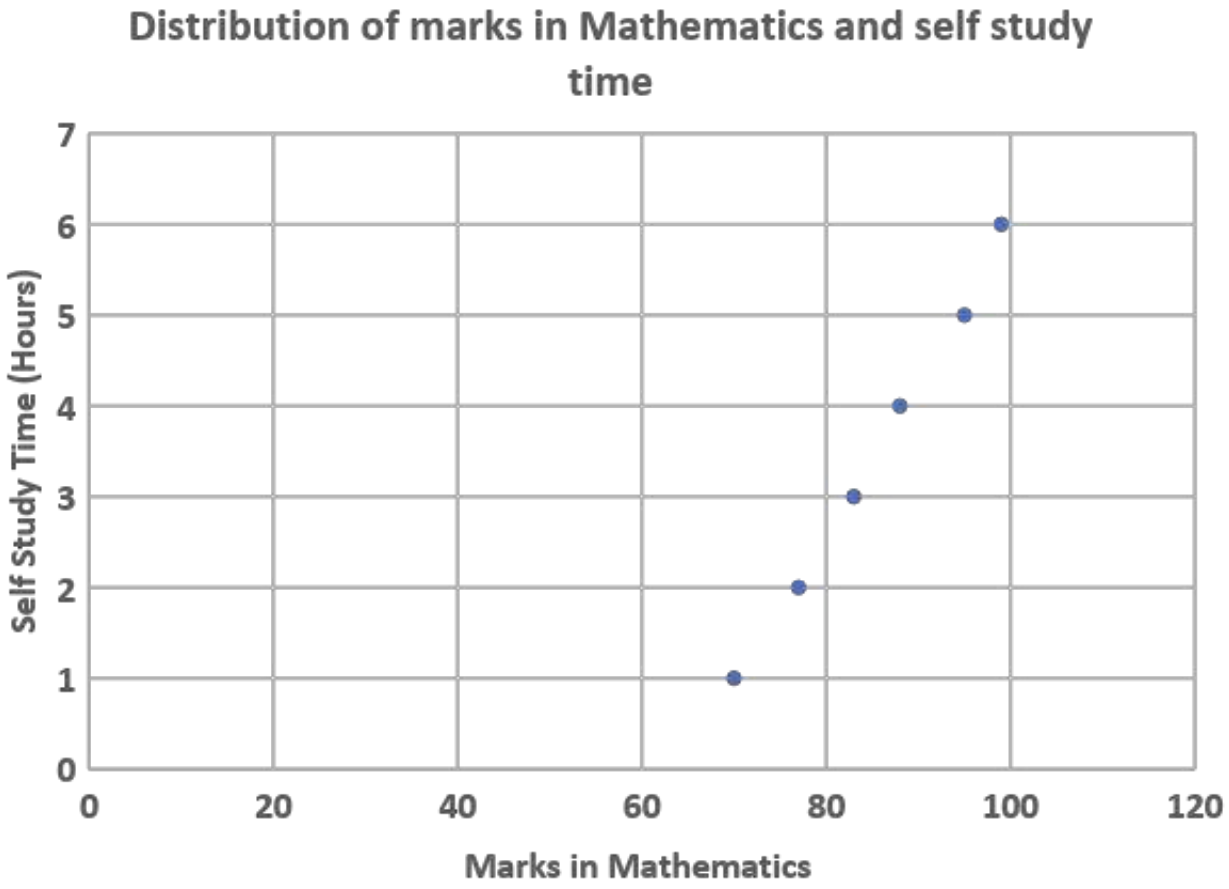


Figure 15: Correlation between Mathematics Marks and Self Study Time of class 10th students

Conclusion

An appropriate and properly prepared graph can be a powerful tool to convey statistical information.

Features of an Idea Graph

- What you aimed to present
- Graph should be Clear
- Define Chart Title & Legends
- Name & Number of each graph
- Label Axes

Which one among the following represents Mean Age of children based upon following dataset:

Age (Years)	10	10	11	12	11	12	13	14	14	15

11.1 years
12.2 years
13.0 years
12.9 years

Which one among the following represents Median weight of children based upon following dataset:

Weight (Kgs)	11	10	14	15	18	16	18	13	14	15

14.0 kgs
15.0 kgs
15.5 kgs
14.5 kgs

In case of nominal data, which measure of central tendency is preferred?

Mean

Mode

Median

Weighted Mean

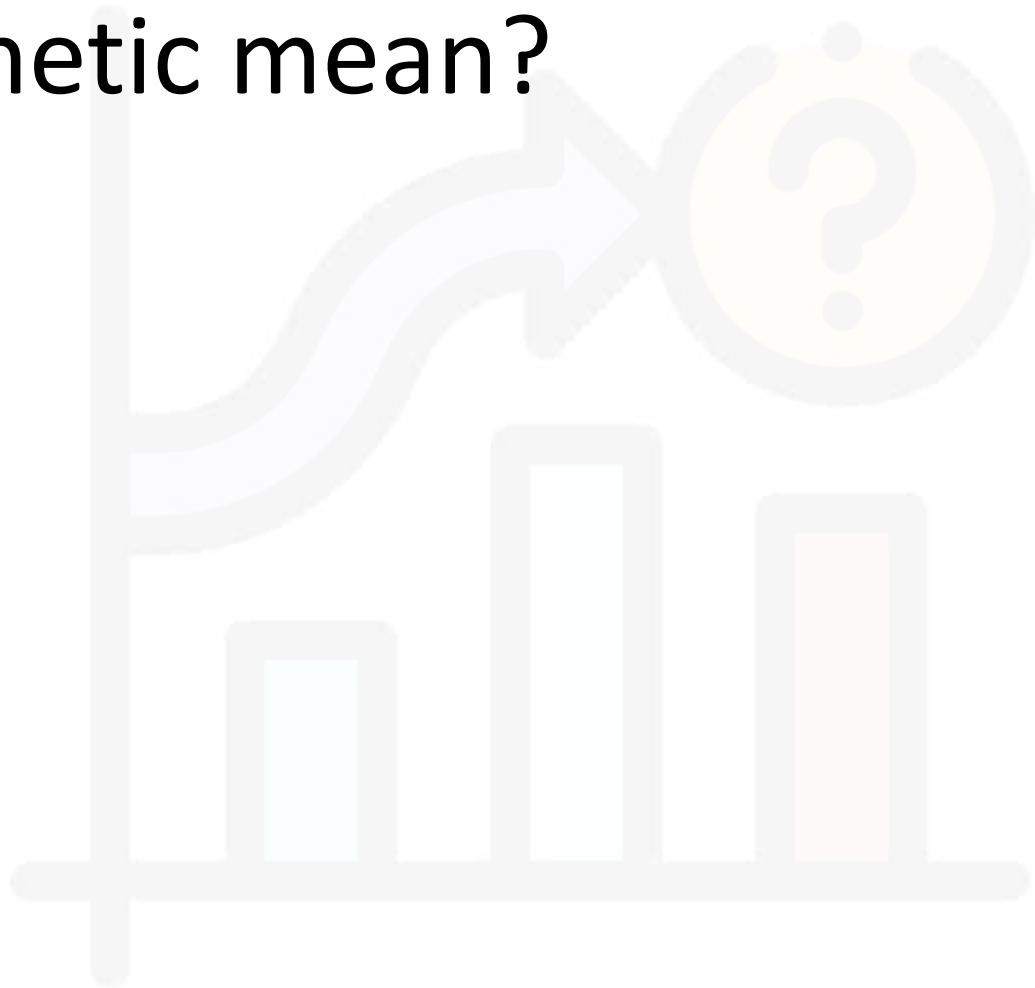
Which measure of dispersion should be reported along with arithmetic mean?

Range

Standard Deviation

Quartile Deviation

Inter Quartile Range



Quartiles divides the data into ___ equal parts?

10

100

4

2

Which one of the following option is correct for Inter Quartile range ?

It gives the range of entire data set

It is the difference between maximum value and the minimum value

It is the difference between first and third quartile

All of the above statements are correct

Which combination among the following is preferred measure of central tendency and dispersion when data consists of outliers?

Median,Range

Median, Inter Quartile Range

Mean, Quartile Deviation

Mean, Standard Deviation

Which chart you will prefer to represent continuous data:

- Histogram
- Bar Diagram
- Pie Chart
- Donut Chart

Which chart you will prefer to represent trend over time:

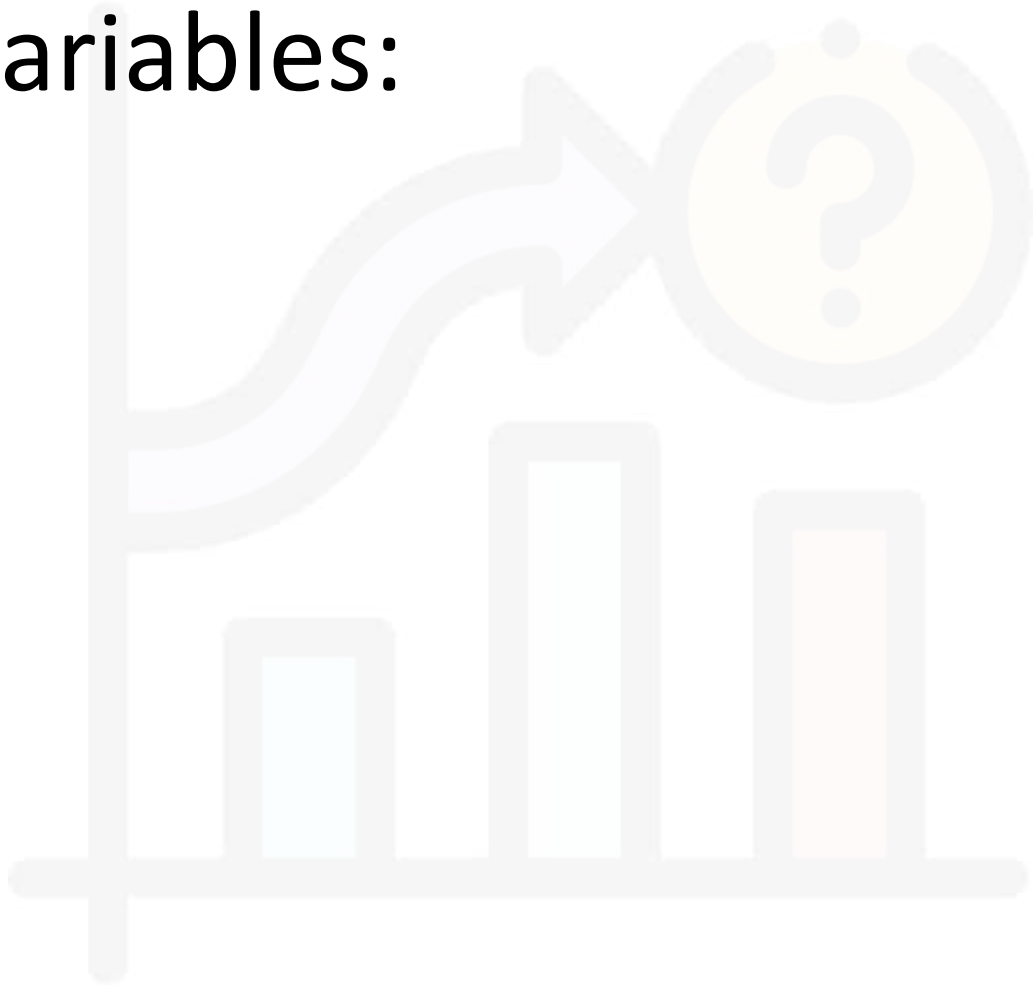
- Histogram
- Bar Diagram
- Line Chart
- Donut Chart

Box and Whisker plots gives us information about all of the below mentioned options except:

- Mean
- Median
- Maximum value
- Quartiles
- Minimum value

Which graph is used to study the relationship between two continuous variables:

- Pie Chart
- Line Chart
- Scatter Plot
- Histogram



Bibliography/Further Readings

- Jerrold H. Zar. Biostatistical Analysis, Fourth Edition, Pearson Education India, 1999.

- S Manikandan. Frequecy Distribution, J Pharmacol Pharmacother.[cited from 2011 Jan-Mar; 2(1): 54–56]. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3117575/

Lets Connect!

draanchalawasthi@gmail.com

https://www.youtube.com/c/sscrindia

+91 750.625.0403