SUMMARIZING & GRAPHING DATA

Gautam K Sharan

MD, PDCR, FICRO

Medical Director
Jawaharlal Nehru Cancer Hospital
Bhopal

Secretary
Indian College of Radiation Oncology
Association of Radiation Oncologists of India

48th ICRO SUN PG Teaching Course

CLINICAL TRIALS & CANCER STATISTICS

Max super Speciality Hospital

26/27 October, 2024

What are Statistics?

- Academic study

- The science of assembling and interpreting numerical data (Bland, 2000)
- The discipline concerned with the treatment of numerical data derived from group of individuals (Armitage et al, 2001)
- The science that deals with the collection, tabulation, and systematic classification of quantitative data, especially as a basis for inference and induction (Funk & Wagnalls dictionary)

The Purpose of Statistics

Fundamentals of Statistics

- The five basic words
- The Branches of Statistics
- Source of Data
- Sampling concepts
- Sample selection methods

The Five Basic Words

- **Population**
- **Sample**
- **Parameter**
- **Statistic**
- **Variable**

The Five Basic Words

1. POPULATION

- All the members of a group about which you want to draw a conclusion

2. SAMPLE

- The part of population selected for analysis

The Five Basic Words: Population v Sample

ASPECT	POPULATION	SAMPLE
Definition	The entire group of individuals, objects, or events	A subset or smaller representation
Size	Typically larger and more comprehensive	Smaller, often manageable for data collection
Representiveness	Represents the entire group under study	Represents a subset, potentially introducing sampling bias
Data Collection	Impractical or resource-intensive	Feasible and efficient
Accuracy	More accurate and precise	Results are estimates and subject to sampling error

The Five Basic Words

3. PARAMETER

- A numerical measure that describes a characteristic of a population

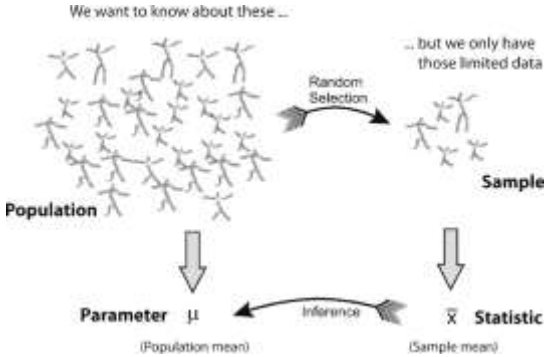
4. STATISTIC

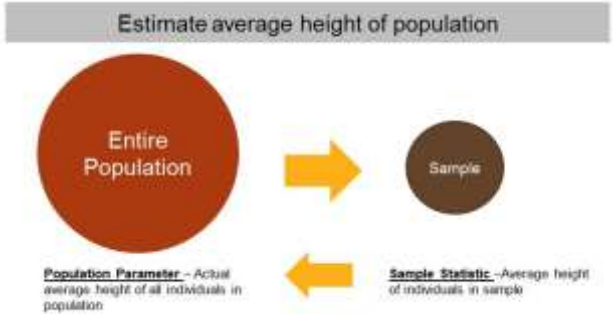
- A numerical measure that describes a characteristic of a sample

Parameter v Statistic

Parameter: True quantities of the (whole) population

Statistic: Quantity calculated from a sample





The Five Basic Words

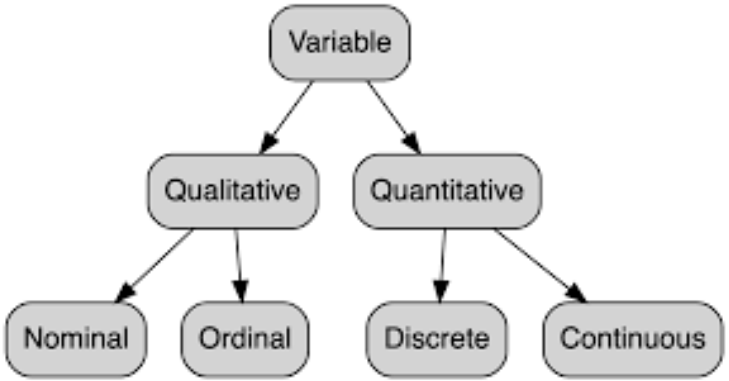
5. VARIABLE

- A characteristic of an item or an individual that will be analyzed using statistics
- A variable is a characteristic that can be measured and that can assume different values
- **Good Variable**: good reliability and validity, low bias, feasibility/practicality, low cost, objectivity, clarity, and acceptance

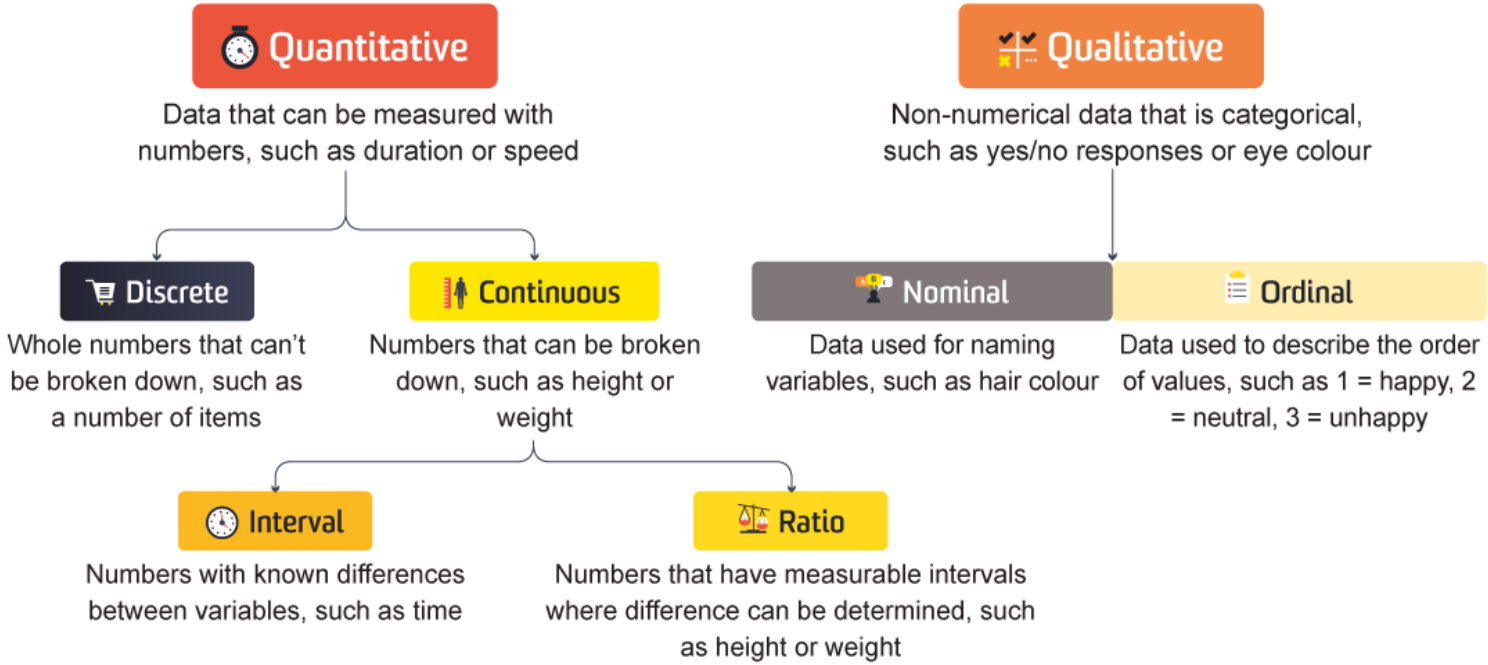
Types of Data: Quantitative vs Categorical Variables

Data is a specific measurement of a variable

1. **Quantitative (Numerical) data** represents amounts
2. **Categorical (Qualitative) data** represents groupings

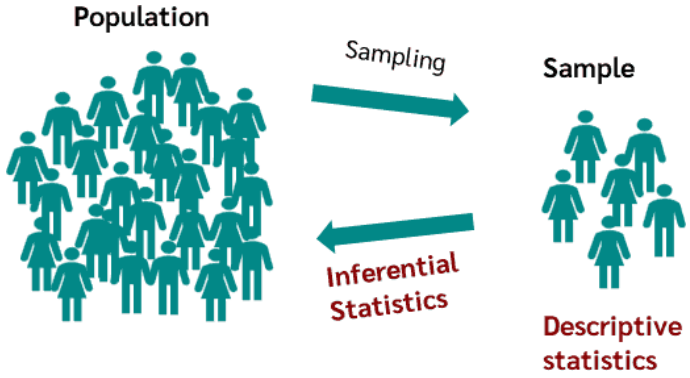


Types of Data

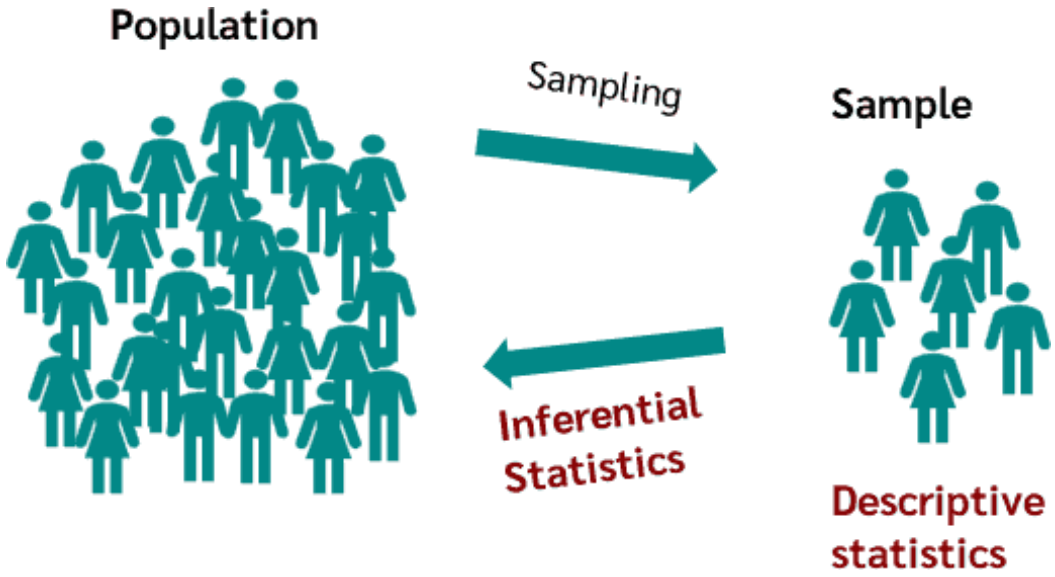


The Branches of Statistics

- Data: items of information
- Use of statistics to describe data:
  Descriptive statistics (minor)
  refer only to actual data
- Use of statistics to draw conclusion or make predictions:
  Inferential statistics (major)
  go beyond the actual data



Descriptive & Inferential Statistics



Descriptive Statistics

- **Descriptive statistics only reflect the data to which they are applied.**
  A descriptive statistic can be:

- A measure of central tendency, like mean, median, or mode: These are used to identify an average or canter point among a data set
- A measure of dispersion or variability, like variance, standard deviation, skewness, or range: These reflect the spread of the data points
- A measure of distribution, like the quantity or percentage of a particular outcome: These express the frequency of that outcome among a data set

Inferential Statistics

- **Inferential statistics techniques include:**

- **Hypothesis tests**, or tests of significance: These involve confirming whether certain results are significant and not simply by chance
- **Correlation analysis**: This helps determine the relationship or correlation between variables
- **Logistic or linear regression analysis**: These methods enable inferring and predicting causality and other relationships between variables
- **Confidence intervals**: These help identify the probability an estimated outcome will occur

Populations, Samples, Elements, Sampling Frame, & Subject

- **Population**: Complete set of people or objects which can be studied
- **Sample**: Subset of population, studied or observed
- Element: A single observation, denoted by **X**
  The number of elements in a population is **N**
  The number of elements in a population is **n**
- Sampling Frame: A listing of all the elements in the population from which sample is drawn
- Subject: A single member of the sample

Errors may happen!

Wrong Number

Not interpreting statistical information properly can lead to disaster. Coca-Cola performed a major consumer study in 1985 and, based on the results, decided to reformulate Coke, its flagship drink. After a huge public outcry, Coca-Cola had to backtrack and bring the original formulation back to market. What a mess!

Let’s review some concepts

Target Population

Population for your study

Sampling Frame

List of every individual in the population from whom the sample will be taken

Sample

Smaller group from the population

Statistic

A value collected from the sample (always an estimate of the true value)

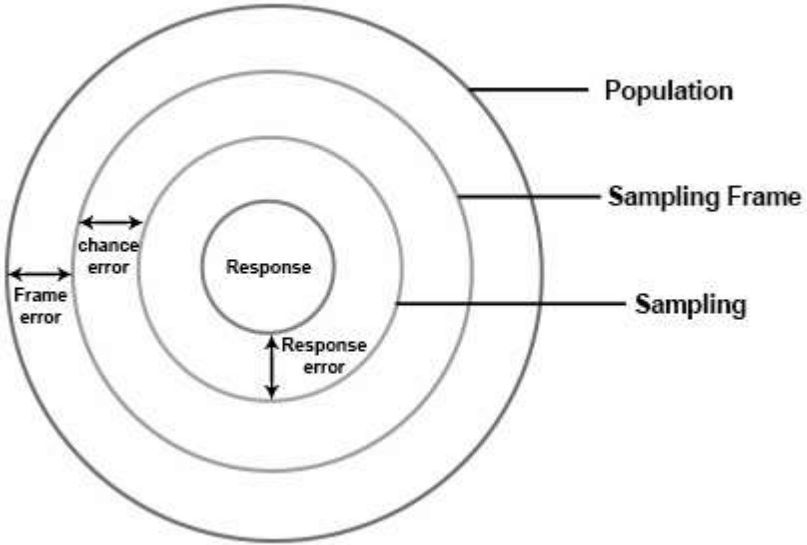
Sampling Variations

Natural differences that occur when we take multiple samples from the same population

Sampling Error

Amount of inaccuracy in estimating some value, which occurs due to considering a small section of the population

Sampling Error



Sampling Error = (Response Error) + (Frame Error) + (Chance Error)

Sources of Data

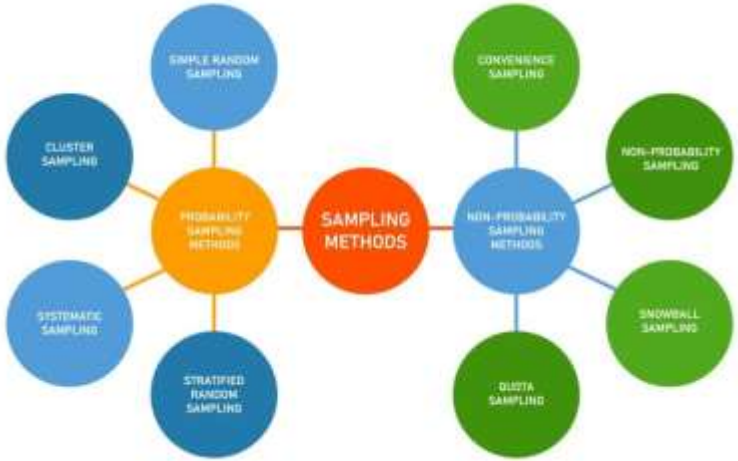
- **Published sources**
- Primary data: published by individual or group that collected the data
- Secondary data: compiled from primary sources

- **Experiments**

- **Surveys**

Sampling

- Process by which members of a population are selected for a sample



Sampling Concepts

Probability Sampling Methods	Non-probability Sampling Methods
Subjects of the population get an equal opportunity to be selected as a representative sample	Subjects of the population DON'T get an equal opportunity to be selected as a representative sample
These are also known as Random sampling methods.	These are also called non-random sampling methods.
These are used for research which is conclusive.	These are used for research which is exploratory.
Produces unbiased result	Produces biased result
There is an underlying hypothesis in probability sampling before the study starts. Also, the objective of this method is to validate the defined hypothesis.	The hypothesis is derived later by conducting the research study in the case of non-probability sampling.

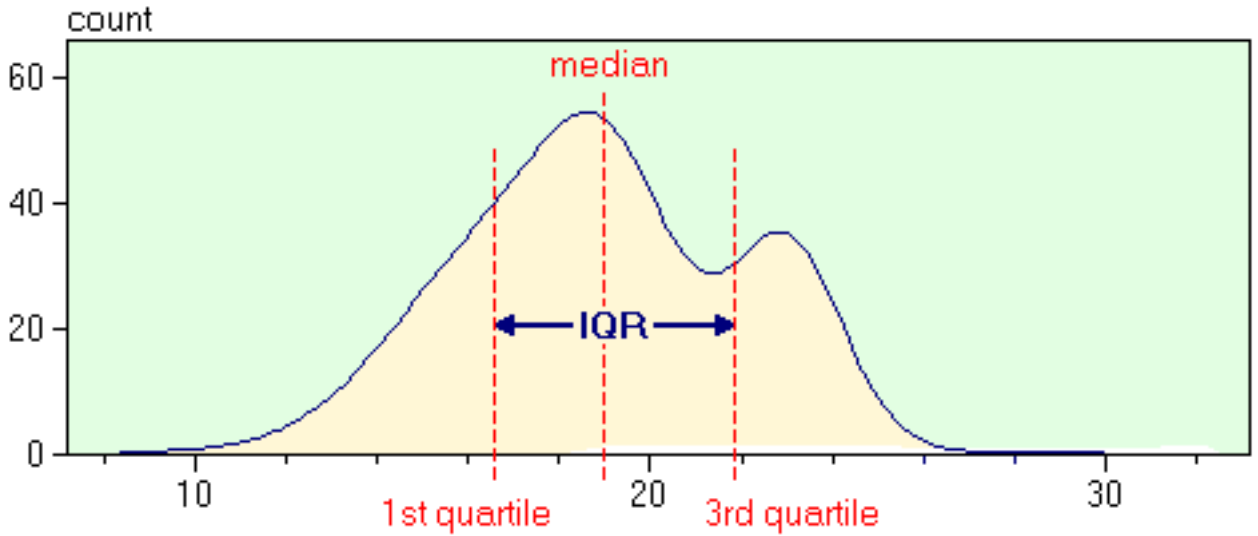
OUTLIER

- An outlier is **an observation that lies an abnormal distance from other values** in a random sample from a population

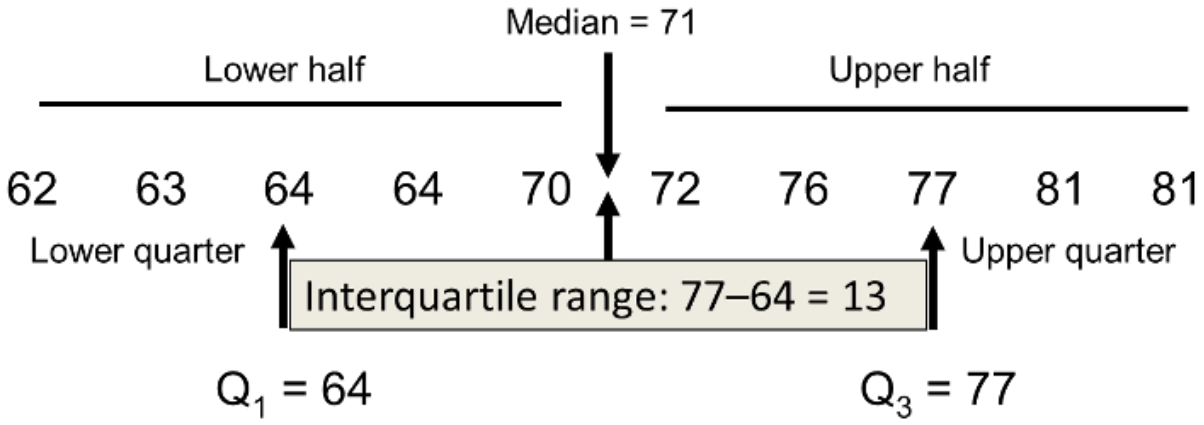
  - **Range = $X_{max} - X_{min}$**
  - Where X_{max} is the largest observation and X_{min} is the smallest observation of the variable values

  - **Interquartile Range**: defines the difference between the third and the first quartile

Interquartile Range



Interquartile Range



Outlier

- **Outlier < Q1 - 1.5(IQR) or**
- **Outlier > Q3 + 1.5(IQR)**

- 2, 4, 5, 5, 6, 11, 11, 13, 14, 25, 30
- **MIN** = 2 **Q1** = 5 **MED** = 11 **Q3** = 14 **MAX** = 30
- **IQR= 9**

Outlier

- **2, 4, 5, 5, 6, 11, 11, 13, 14, 25, 30**

- **Outlier < Q1 - 1.5(IQR)**
- **Outlier < 5 - 1.5(9)**
- **Outlier < 5 - 13.5**
- **Outlier < - 8.5**

Outlier

- **2, 4, 5, 5, 6, 11, 11, 13, 14, 25, 30**

- **Outlier > Q3 + 1.5(IQR)**
- **Outlier > 14 + 1.5(9)**
- **Outlier > 14 + 13.5**
- **Outlier > 27.5**

- **Outlier: 30**

Chart vs Table

- A chart, also known as a graph, is a visual representation of data used to display patterns, trends, and relationships clearly and concisely
- Charts are best used to display patterns, trends, and relationships in the data
- They are particularly useful for identifying trends over time and comparing data points

Chart vs Table

- A table organizes data into rows and columns, making it easy to compare and analyse information
- Tables are best used when precise values need to be displayed and compared
- They are particularly useful for displaying large amounts of data and making detailed comparisons between data points

Bar Chart

- A bar chart (aka bar graph, column chart) plots numeric values for levels of a categorical feature as bars
- Levels are plotted on one chart axis, and values are plotted on the other axis
- Each categorical value claims one bar, and the length of each bar corresponds to the bar's value. Bars are plotted on a common baseline to allow for easy comparison of values

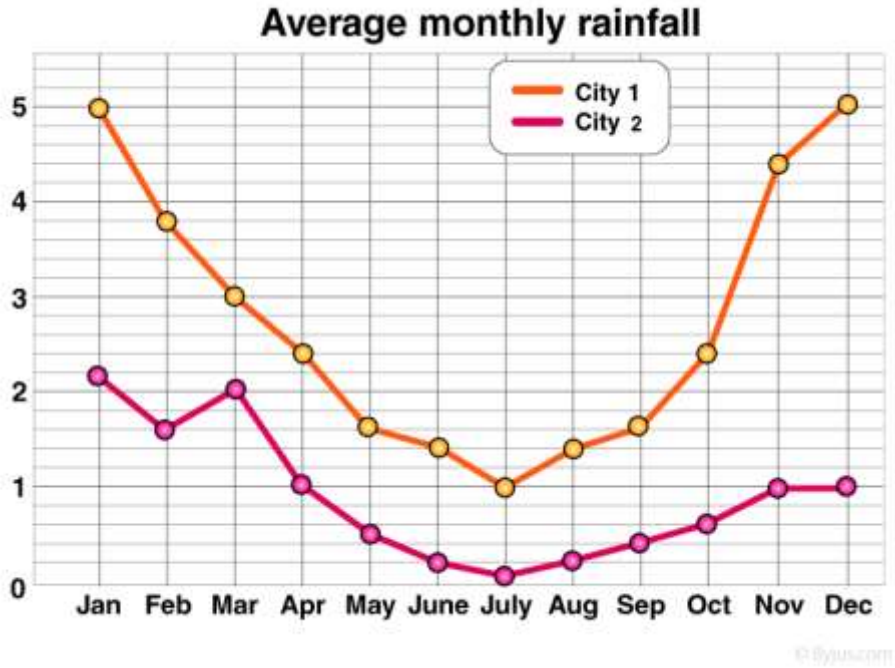
Bar Chart


100
80
60
40
20
0
Quality Score
Their Product
Our Product
100
98
96
94
92
90
Quality Score
Their Product
Our Product

Line Graph

- A line graph or **line chart** or **line plot** is a graph that utilizes points and lines to represent change over time
- It is a chart that shows a line joining several points or a line that shows the relation between the points
- The graph represents quantitative data between two changing variables with a line or curve that joins a series of successive data points

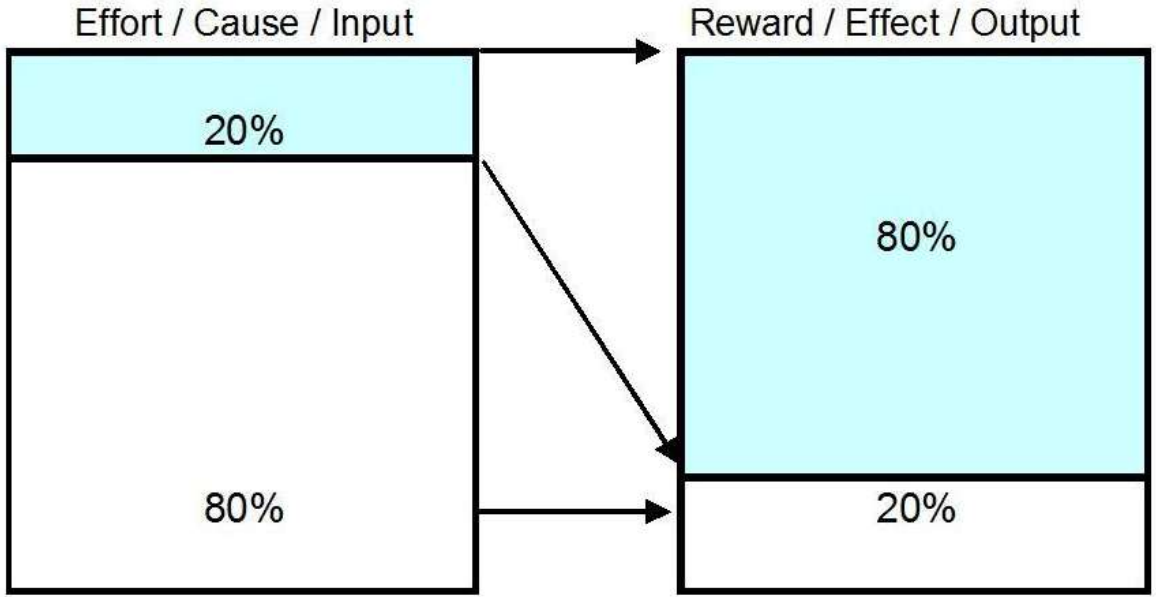
Line Graph


Average monthly rainfall
City 1
City 2
5
4
3
2
1
0
Jan
Feb
Mar
Apr
May
June
July
Aug
Sep
Oct
Nov
Dec

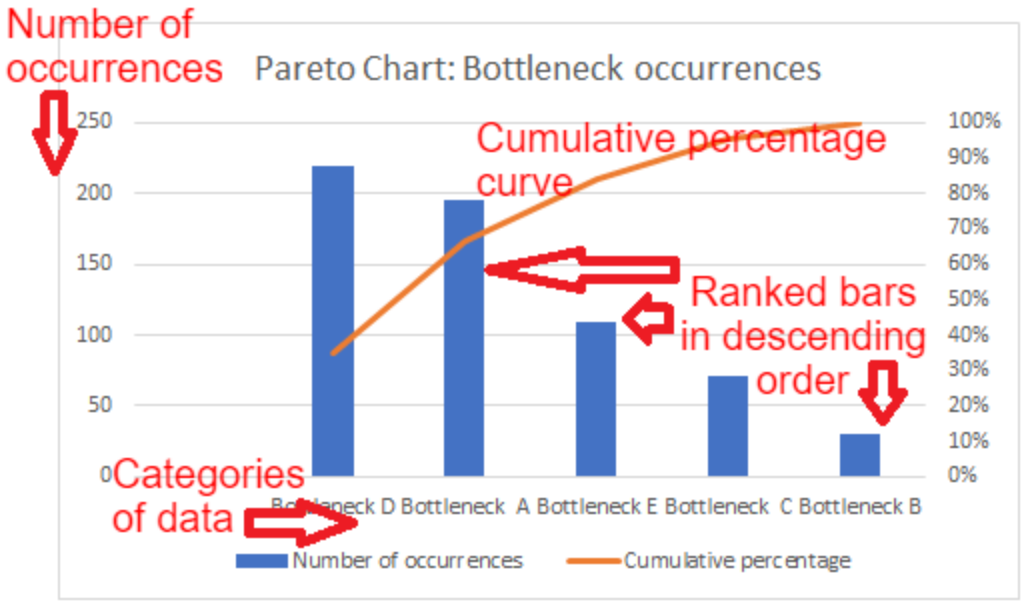
Pareto Chart

- A **Pareto chart** is a type of chart that contains both bars and a line graph, where individual values are represented in descending order by bars, and the cumulative total is represented by the line
- The purpose of using this chart is to represent a set of data in a bar graph chart
- The individual values are represented by the length of the bars and the line shows the combined total
- Values are expressed from the longest bar to the shortest bar in the graph

The Pareto Principle (80/20)



Pareto Chart

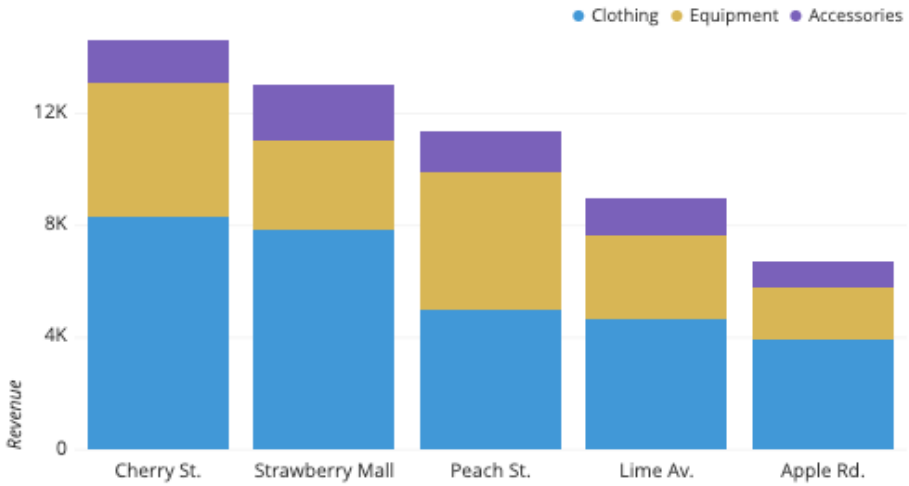


Segmented (Stacked) Column Chart

- A **segmented bar chart** is a type of chart that uses segmented bars that add up to 100% to help us visualize the distribution of categorical data

- Each bar in a standard bar chart is divided into a number of sub-bars stacked end to end, each one corresponding to a level of the second categorical variable

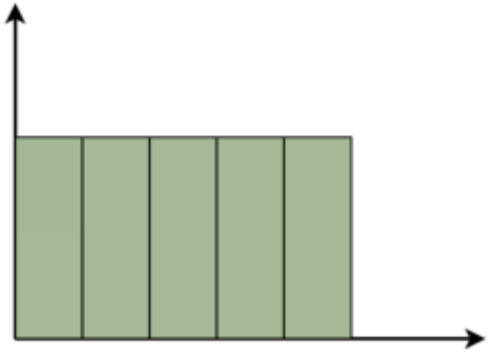
Segmented (Stacked) Column Chart



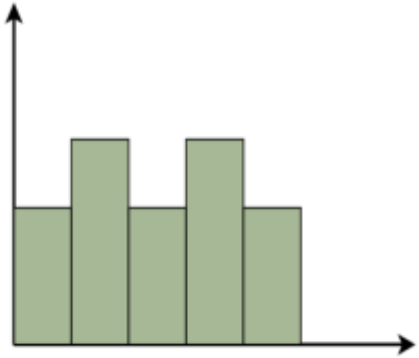
Histogram

- **Histogram** is a graphical representation that condenses data series into easy-to-understand numerical data by **grouping them into logical ranges of varying heights,** often known as bins.

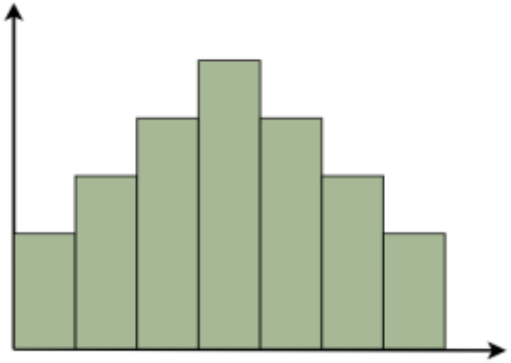
- Unlike bar graphs, which are used for categorical data, **histograms are designed for continuous data**, grouping it into logical ranges or “bins.”



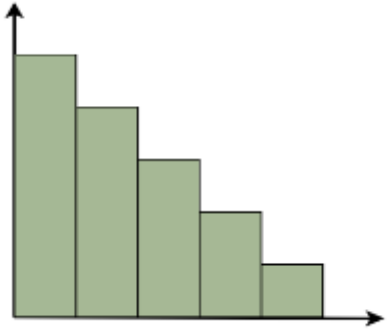
Uniform Histogram



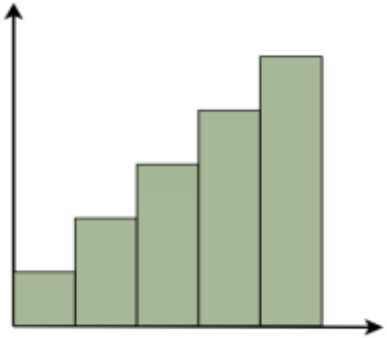
Bimodal Histogram



Symmetrical Histogram



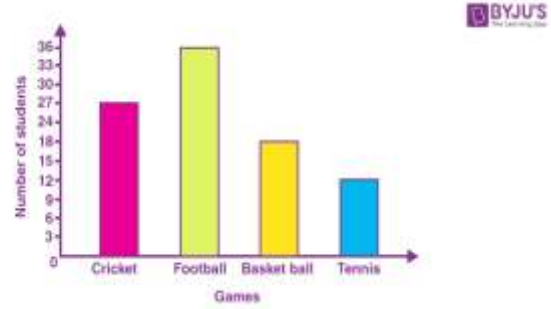
Right Skewed Histogram

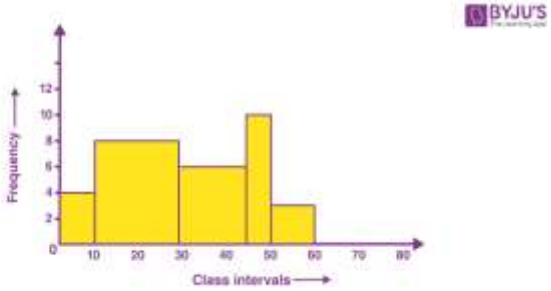


Left Skewed Histogram

Histogram vs Bar Graph

Histogram	Bar Graph
It is a two-dimensional figure	It is a one-dimensional figure
The frequency is shown by the area of each rectangle	The height shows the frequency and the width has no significance.
It shows rectangles touching each other	It consists of rectangles separated from each other with equal spaces.

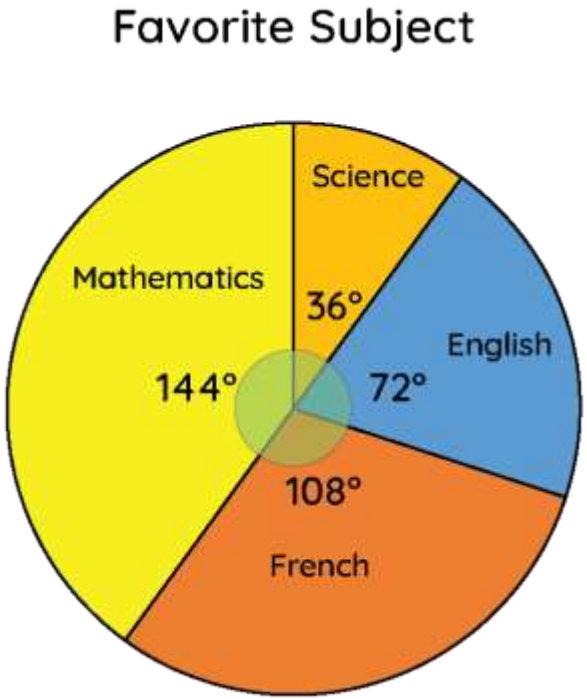




Pie Chart/ Circle Chart

- A **pie chart** is a type of graph that represents the data in the circular graph.
- The slices of pie show the relative size of the data, and it is a type of pictorial representation of data.
- A pie chart requires a list of categorical variables and numerical variables.

Pie Chart/ Circle Chart



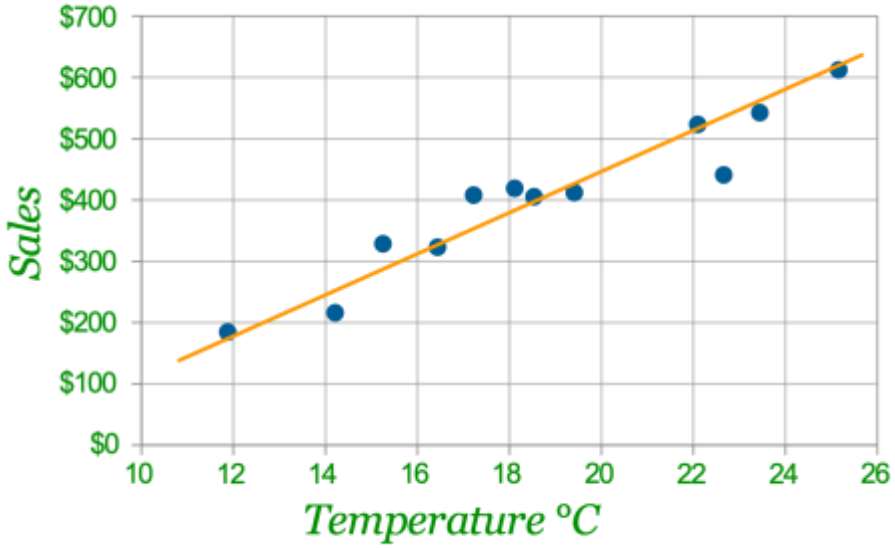
Dot Plot

- Group of data points plotted on a simple scale
- Simplest statistical plots
- Suitable for small to moderate sized data sets

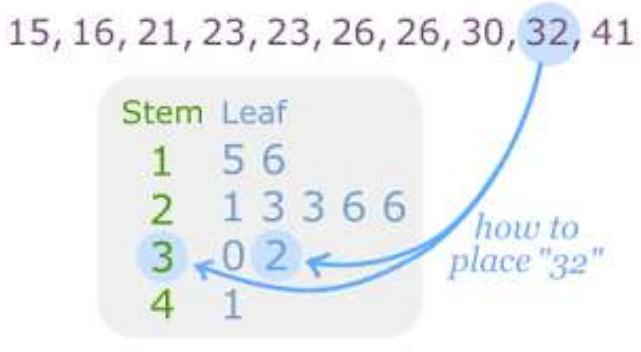
Scatter Plot

- Graphs that present relationship between two variables in a data-set
- For a large set of data points given
- Each set comprises a pair of values
- The given data is in numeric form

Scatter Plot



Stem & Leaf Plot



- Graphical representation used to organize and display quantitative data in a semi-tabular form
- Stem: leading digits, listed vertically
- Leaf: trailing digit, next to stem
- Key is provided to explain what the stem and leaf represent for that particular plot

Phone Call Lengths

Stem	Leaf
0	2 3 5 6
1	0 4 9
2	3 3
3	0 6
4	
6	6

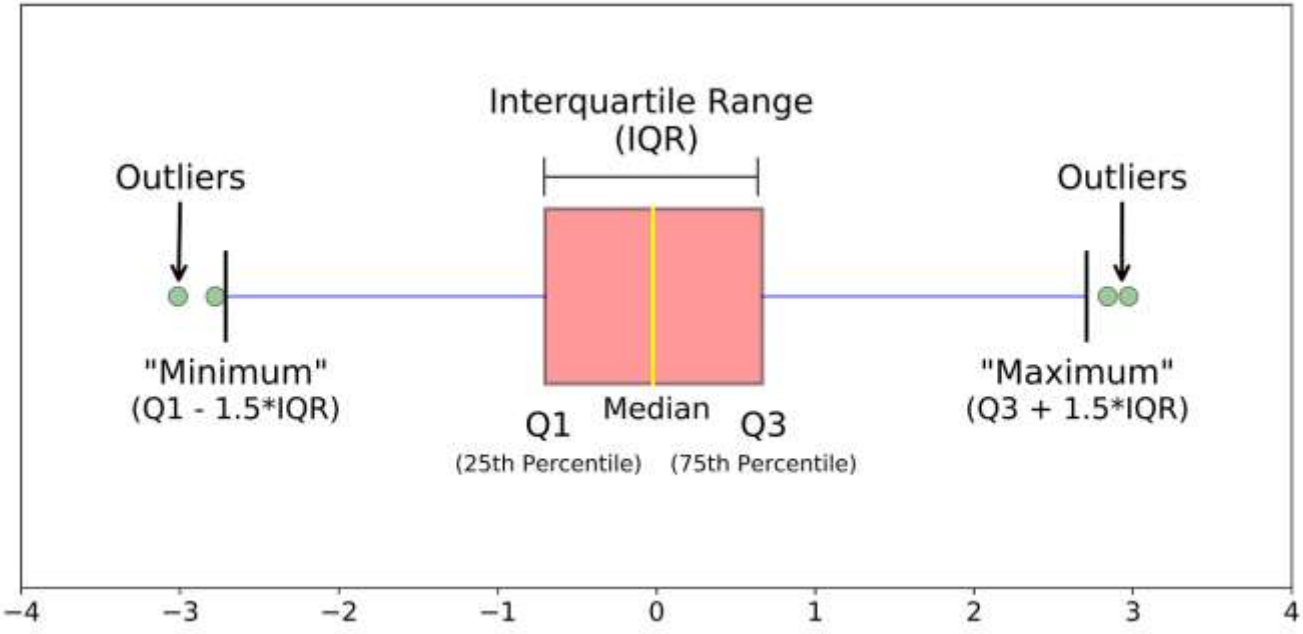
Key : 1|4 = 14 minutes

Box and Whisker Plot

- Visual display of data distribution through their quartiles

- Here are the types of observations one can make from viewing a Box Plot:
- What the key values are, such as: the average, median, 25th percentile, etc
- If there are any outliers and what their values are
- If the data is symmetrical or not
- How tightly is the data grouped
- If the data is skewed and if so, in what direction

Box and Whisker Plot



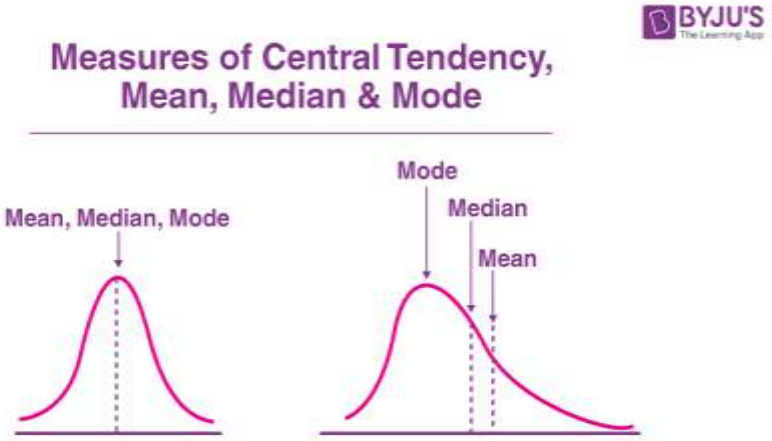
Descriptive Statistics

- **Descriptive statistics only reflect the data to which they are applied.**
  A descriptive statistic can be:

- A measure of central tendency, like mean, median, or mode: These are used to identify an average or canter point among a data set
- A measure of dispersion or variability, like variance, standard deviation, skewness, or range: These reflect the spread of the data points
- A measure of distribution, like the quantity or percentage of a particular outcome: These express the frequency of that outcome among a data set

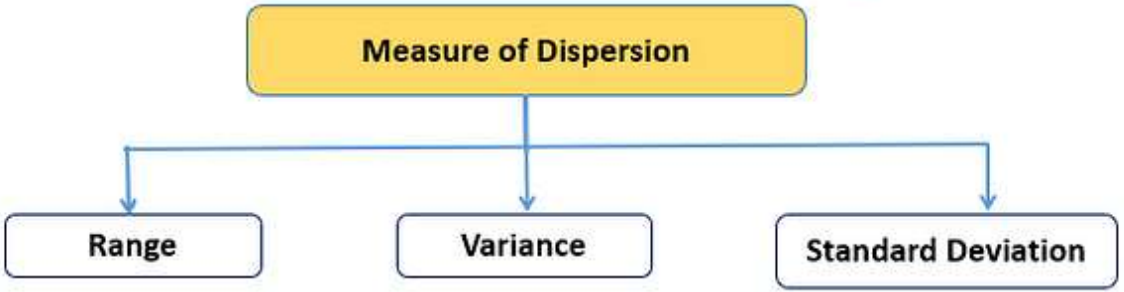
Measures of Central Tendency

- **Mean**: Arithmetic mean
- **Median**: Middlemost value in the ordered list of observations
- **Mode**: Most frequently occurring value

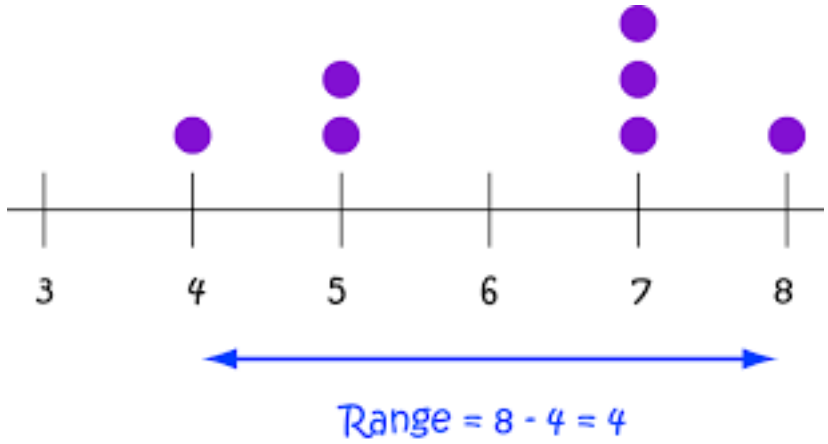


2 × Mean + Mode = 3 × Median

Measures of Dispersion



Measures of Dispersion: Range

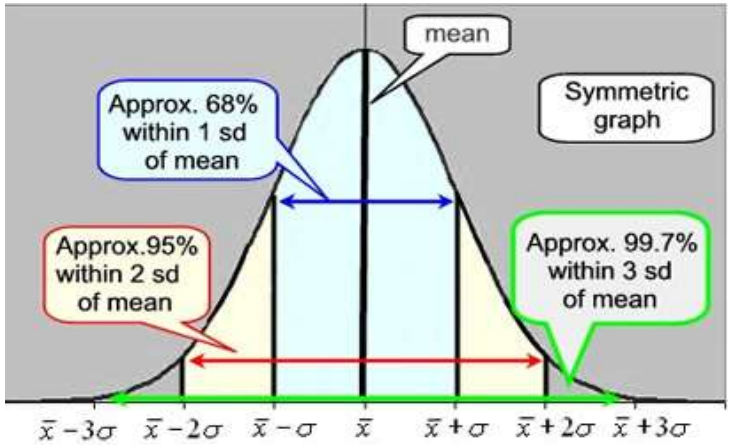


Measures of Dispersion: Variance

- Variance measures how far a set of data is spread out
- Variance is the average of the squared distances from each point to the mean
- σ^2 (doesn't have the same unit of measure as the original data)
- Zero variance: all data values are identical
- Always positive

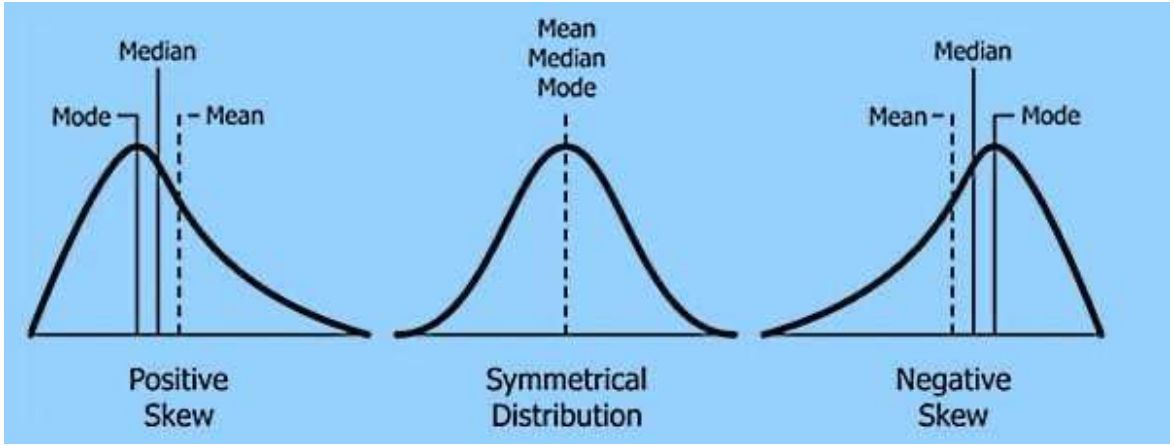
Measures of Dispersion: Standard Deviation

- Standard deviation shows how much variation (dispersion, spread, scatter) from the mean exists
- Represents a "typical" deviation from the mean
- Popular measure of variability because it returns to the original units of measure of the data set
- σ
- Square root of variance

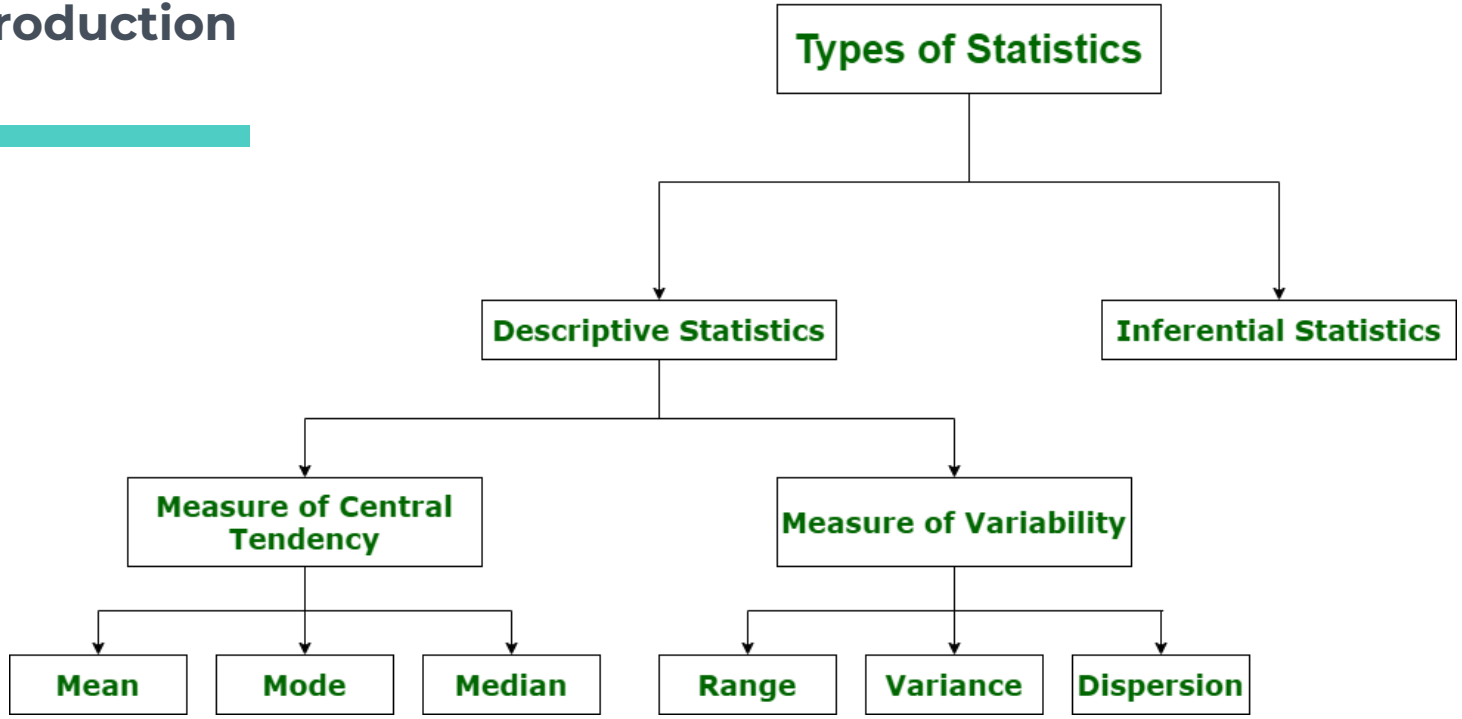


Skewness

- Skewness is the degree of asymmetry or departure from the symmetry of the distribution of a real-valued random variable



Introduction



Thanks!

This is a slide title

1.

Transition headline

Let’s start with the first set of slides