Statistics for Data Science

AccioJob/ Statistics-for-DataScience

Content

- What is Statistics
- Basic Terms of Statistic
- Measures of Central Tendency (Mean, Medium, Mode)
- Measures of Dispersion (Range, Variance, Standard Deviation, CV)
- Box Plot
- Skewness & Kurtosis
- Percentile/ Quantile
- Chebyshev’s Theorem
- Correlation and Covariance

Definition

Concepts

Examples

Charts

Content

1 What is Statistics?

2 Basic Terms and Concepts

3 Measures of Central Tendencies

4 Measures of Dispersion

5 Skewness & Kurtosis

6 Chebyshev’s Theorem

What is Statistics?

- **Science** concerned with **developing and studying methods** for
  - ***Collecting,***
  - ***Analyzing,***
  - ***Interpreting*** and
  - ***Presenting*** of empirical data
- Also called the Science of ***data*** and ***information***
- Form of **mathematical analysis** that uses ***quantified models***, ***representations*** and ***synopses*** for a given set of experimental data or real-life studies
- Studies methodologies to ***gather***, ***review***, ***analyze*** and ***draw conclusions*** from data.
- Two branches -
  - ***Descriptive*** *statistics*
  - ***Inferential*** *statistics*

The word Statistic is derived from Italian word *‘stato’* which means ***state*** and ‘statista’ refers to a person involved in the affair of the state. Therefore, *statistics* originally meant the **collection of facts useful to the statista**.

Data and Information

- **Data** are a collection of unrefined facts and figures that do not have any added interpretation or analysis. These are the observed values of a variable, at a specific observational Unit.

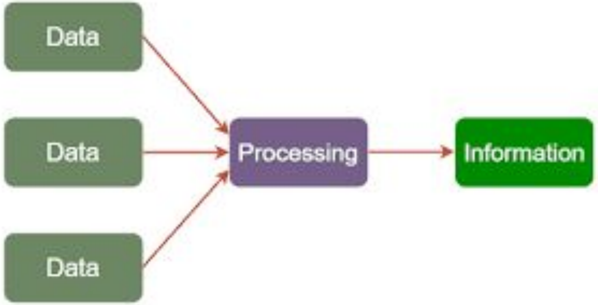
- **Observed data** are a measurement of a variable, in a given state. For instance, the height of a person (at a particular point of his life), the weight of a package, the pressure of a riveting machine, etc., are the observed data.

- Data can be of different types like - Numerical and Categorical. Further they can be Nominal, Ordinal, discrete and Continuous.

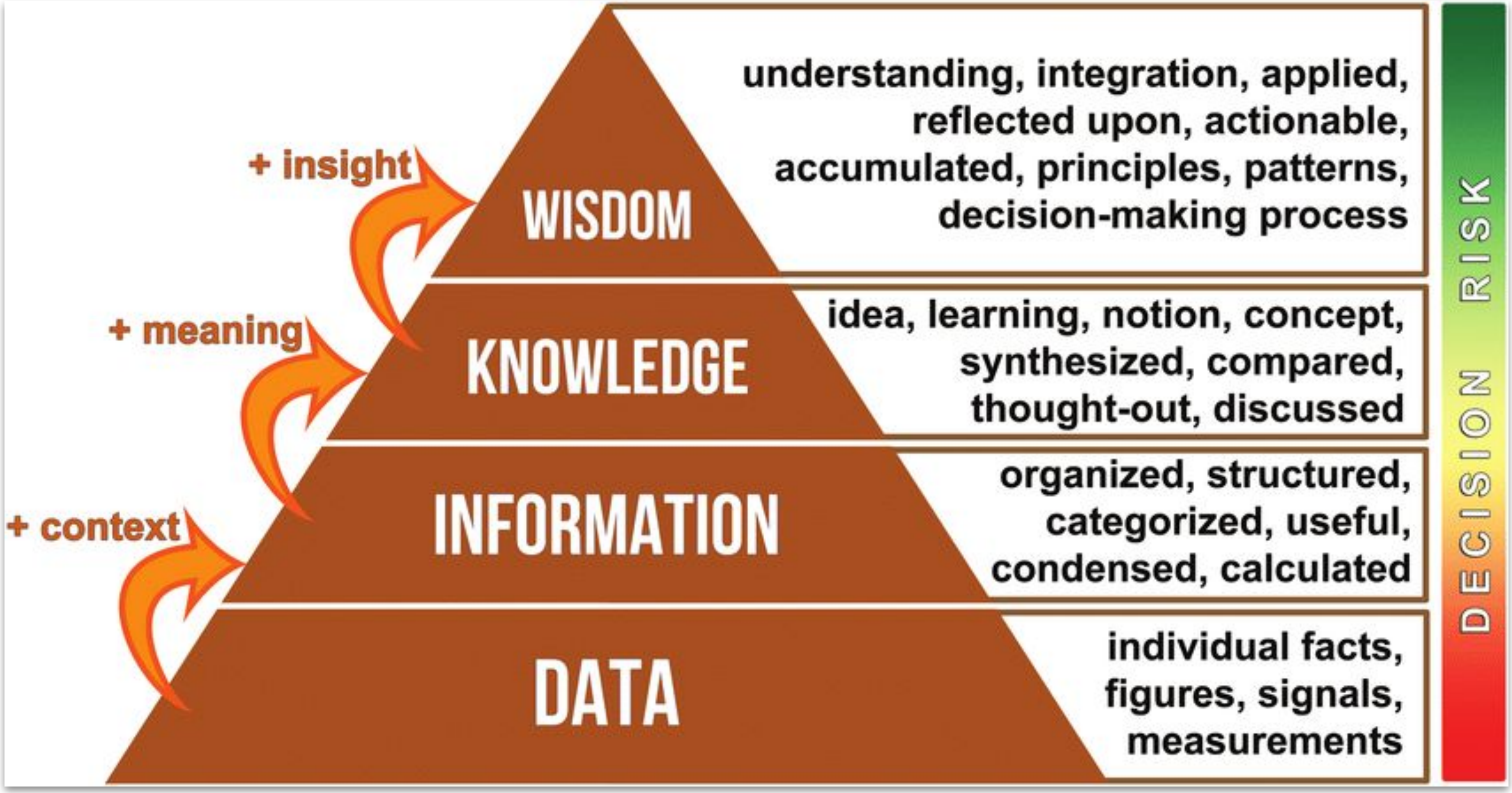
Data, Information and Statistic

- **Information,** retrieved from the data, is a logical interpretation and/or statement which can be observed from the data.

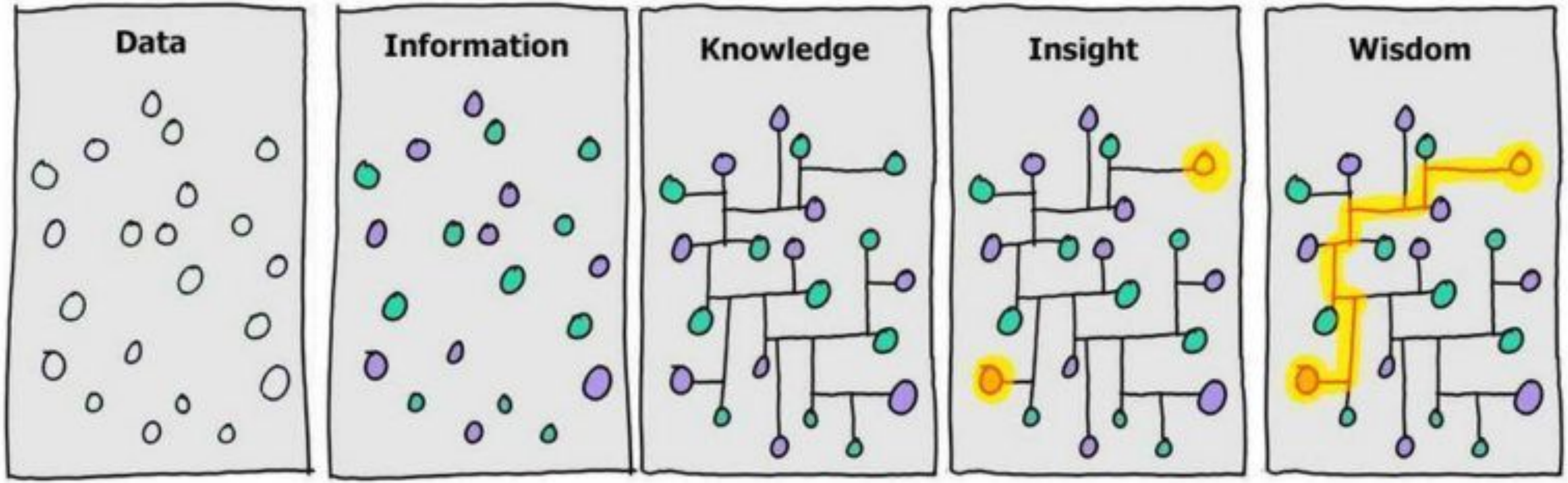
- **Statistic** is a function which takes values observed in a sample as input and gives a number with some meaning/information about the data as output. It is a fully computable quantity which can be obtained from the data. A statistic is based on logical reasoning, e.g., mean gives the average, and standard deviation gives a measure of variation in the sample.



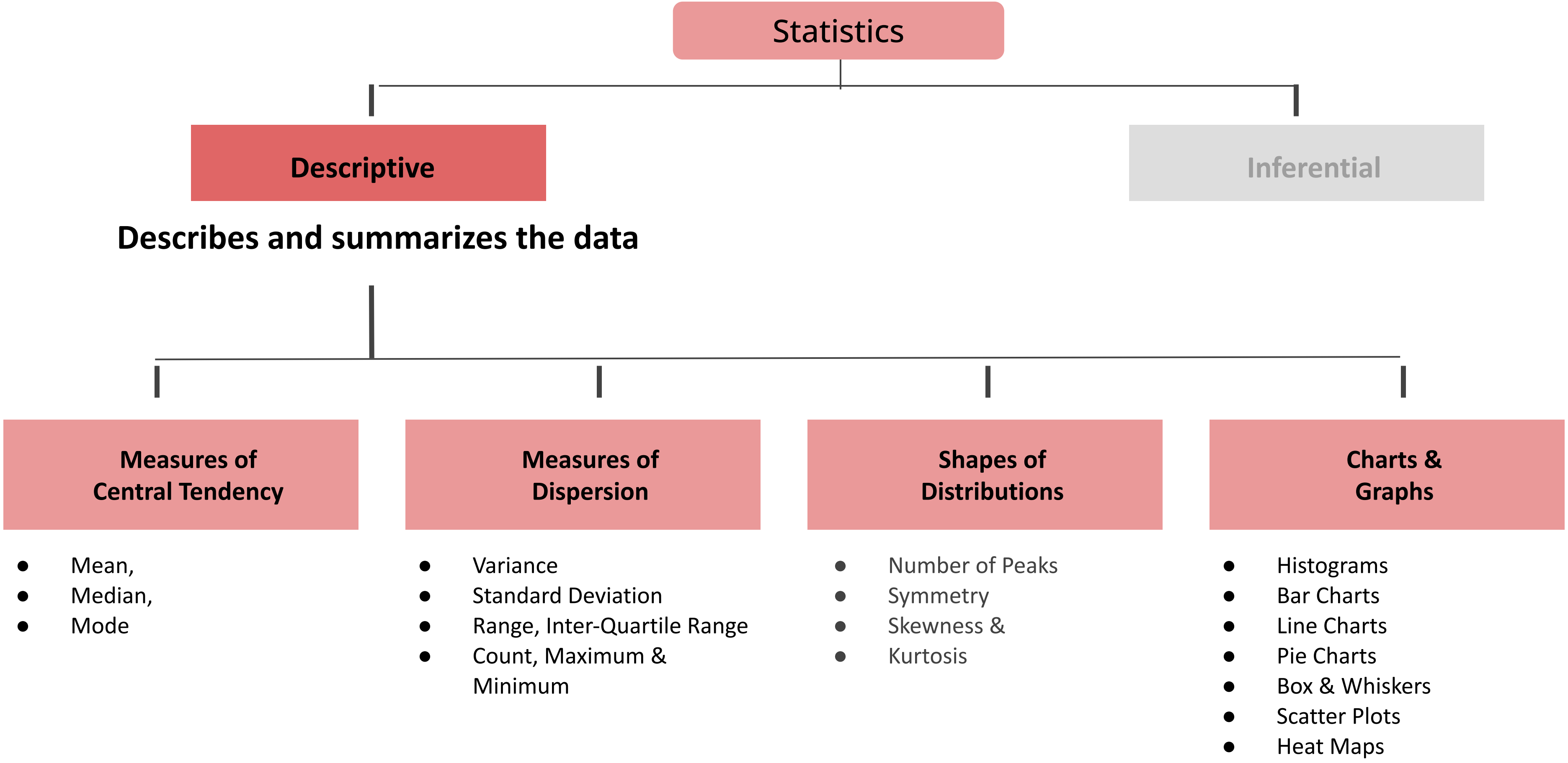
Data, Information, Knowledge and Wisdom Hierarchy


+ insight
+ meaning
+ context
WISDOM
KNOWLEDGE
INFORMATION
DATA
understanding, integration, applied, reflected upon, actionable, accumulated, principles, patterns, decision-making process
idea, learning, notion, concept, synthesized, compared, thought-out, discussed
organized, structured, categorized, useful, condensed, calculated
individual facts, figures, signals, measurements
DECISION RISK

Data, Information, Knowledge Wisdom



Descriptive Statistics



Statistics

Descriptive

Inferential

Describes and summarizes the data

Measures of Central Tendency

- Mean,
- Median,
- Mode

Measures of Dispersion

- Variance
- Standard Deviation
- Range, Inter-Quartile Range
- Count, Maximum & Minimum

Shapes of Distributions

- Number of Peaks
- Symmetry
- Skewness &
- Kurtosis

Charts & Graphs

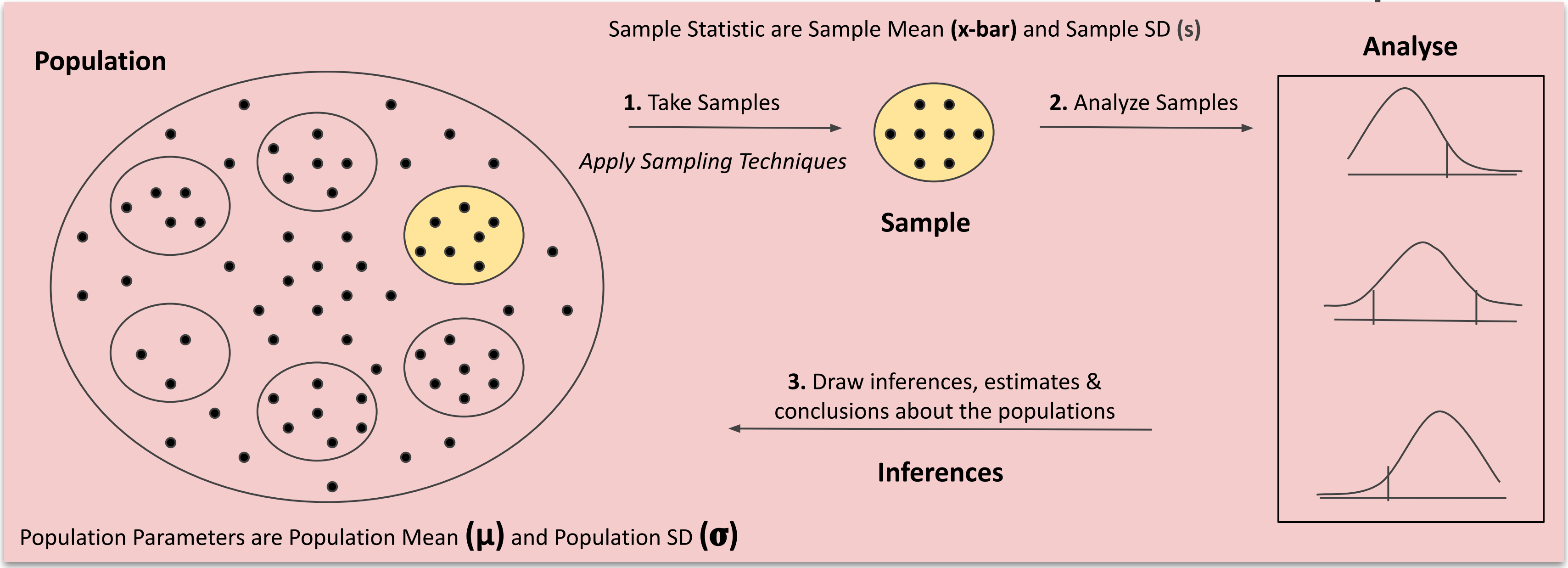
- Histograms
- Bar Charts
- Line Charts
- Pie Charts
- Box & Whiskers
- Scatter Plots
- Heat Maps

Inferential Statistics

Descriptive

Inferential

Use Sample statistics to make predictions about Population



Sample Statistic are Sample Mean **(x-bar)** and Sample SD **(s)**

Population

1. Take Samples

Apply Sampling Techniques

Sample

2. Analyze Samples

Analyse

3. Draw inferences, estimates & conclusions about the populations

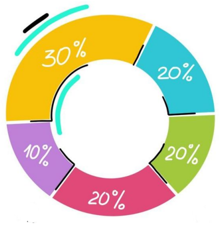
Inferences

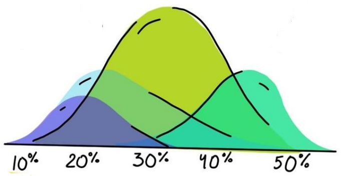
Population Parameters are Population Mean **(μ)** and Population SD **(σ)**

Basic Concepts for Statistics

Types of Statistics

Descriptive Statistics	Inferential Statistics
Method of *organising*, *summarizing* and *presenting data* in an informative way	Make *inferences* and *draw conclusions* about the 'Population' on the basis of 'Sample'
Describes the target 'Population'	Make inferences from 'Samples' and generalize then on the 'Population'
Describe historical data (which already exist)	Make conclusions about the 'Population' which is beyond the data that is available
Tools - Measures of Central Tendency (Mean/ Median/ Mode), Spread of Data (Range, Variance, Standard Deviation)	Tools - Hypothesis, ANOVA, Chi Square, Regression Analysis
Final results are in the form of charts, tables & graphs	Final results are in the form of probability scores





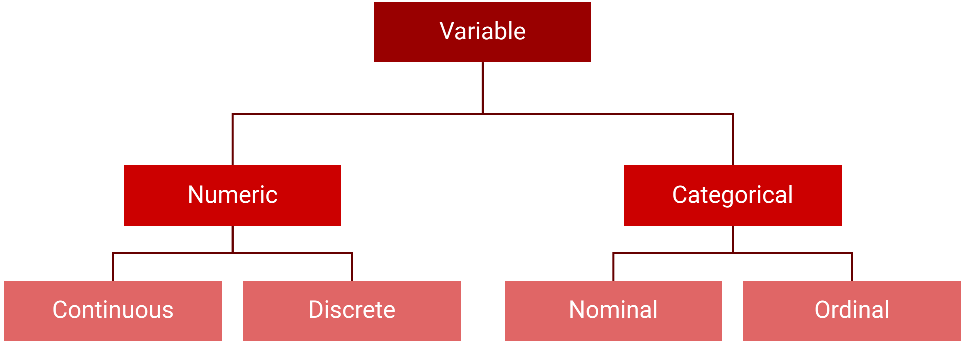
Types of Data/ Variables

Data Types

Qualitative	**Quantitative**
A variable that *describes* a Population/ Sample element. Arithmetic operations such as additions and averaging do not apply to such variables	A variable that *quantifies* a Population/ Sample element. Arithmetic operations such as additions and averaging apply to such variables
Attribute or Categorical Variable	Numeric Variable
Eg - ● Color or Type of Product Category sold ● Boolean (Yes/ No, True/ False) ● Customer Type (Male/ Female/ Kids) ● Ordered data (High/ Medium/ Low; Big/ Small; First/ Second/ Third) ● Location of Delivery of Order	Eg - ● Height Weight of Men/ Women ● Sales Revenue ● Count of footfall in a mall ● No of Beds in Hospital ● No. of Flights took off in each hour of day

Types of Variables

Data Types



Variable

Numeric

Categorical

Continuous

Discrete

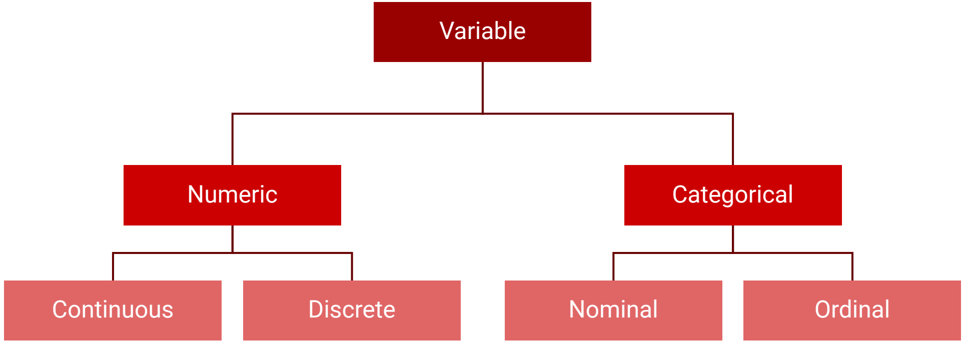
Nominal

Ordinal

Numeric (Quantitative)	**Categorical (Qualitative/ Attribute)**
● Numeric variables have values that describe a **measurable quantity as a number**, like '*how many*' or '*how much*' ● Between any 2 continuous values, there can be infinite number of more data points. ● They are always essentially numeric ● Eg - Sales in $ ($ 530,234.5), Height (5’7.5”), Distance (3.14 kms)	● A variable that ***describes*** a population element. **Arithmetic operations** such as additions and averaging does not **apply** to such variables. ● Categorical variable has values that you can put into a **countable number** of distinct groups based on a characteristic. If the variable has a natural order, it is an ordinal variable

Basic Concepts for Statistics

Types of Variables


Variable

Numeric

Categorical

Continuous

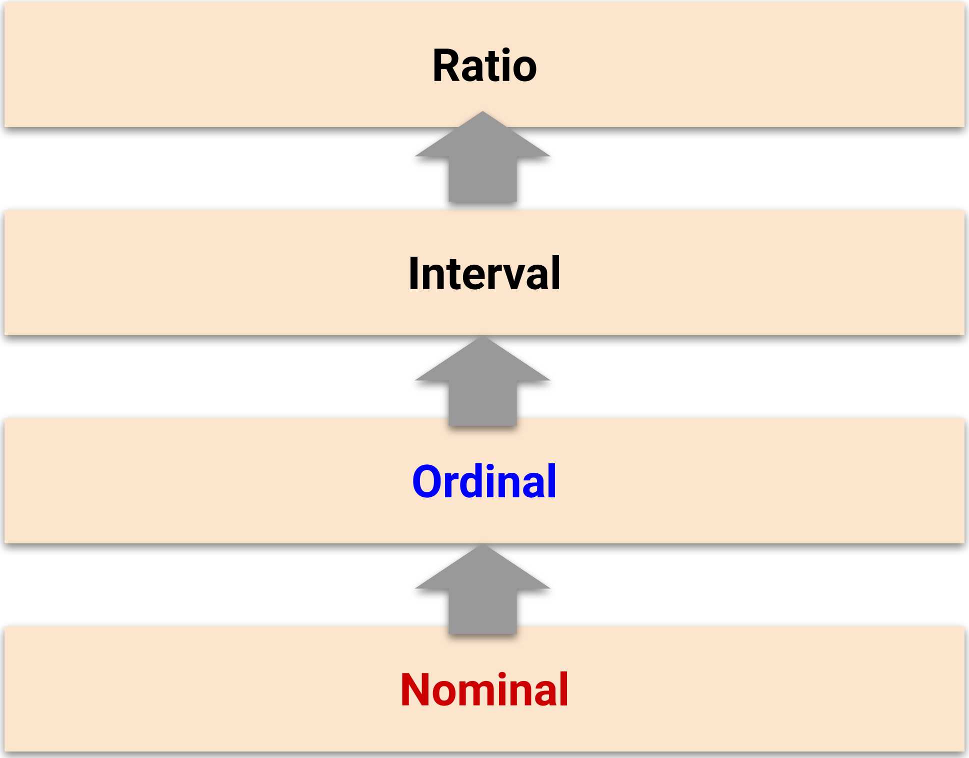
Discrete

Nominal

Ordinal

Numeric (Quantitative)	Categorical (Qualitative/ Attribute)
● **Continuous** Numeric variable: A numerical variable that can take values on a continuous scale (e.g. age, weight). ● **Discrete** Numeric variable: A numerical variable that only takes on whole numbers (e.g. number of visits)	● **Nominal** variables have *two or more categories* and no natural order to these categories. They are **named** (Nominal comes from Latin meaning pertaining to names) ● **Ordinal** variables have *at least three categories* and the categories have a natural order. The categories are ranked but the differences between ranks may not be equal.

Scales of Measurement



Ratio

Interval

Ordinal

Nominal

Difference between measurements and True zero exist.

Difference between measurements but no True zero

Ordered Categorical *(ranks, order or scaling)*

Categorical (No Ordering or direction

Scales of Measurement

Four levels of measurement in research and statistics:

Nominal Scale:

Naming scale, where variables are simply named/ labeled, with no specific order. Categorical. Used to classify. Countable.

Eg - Gender, House Numbers, Names of the Cities, Telephone numbers, Names etc.

Ordinal Scale :

Has all its variables in a specific order, beyond just naming them. In addition to Nominal capabilities it can be Ranked, Ordered or Scaled.

Eg - Heights of people, Floor Numbers,

Scales of Measurement

Four levels of measurement in research and statistics:

Interval Scale :

- Scale in which distances between the consecutive numbers have meaning and data are always numerical.
- 'Interval' indicates 'distance between two entities' which is what Interval scale helps in achieving. Between two points the Interval remains same always
- Central tendency in this scale are Mean, Median, or Mode
- The only drawback of this scale is that there NO pre-decided starting point or a true Zero value.

Eg : temperature, IQ Scale, Credit Score, Time on a clock with hands

Scales of Measurement

Four levels of measurement in research and statistics:

Ratio Scale:

Ratio scale not only produces the order of variables but also makes the difference between variables known along with meaningful ratio and true value of zero. Calculated by assuming that the variables have an option for zero, the difference between the two variables is the same and there is a specific order between the options. It is exactly the same as the interval scale except that the zero on the scale means: DOES NOT EXIST. It means there is absence of property.

With the option of true zero, varied inferential, and descriptive analysis techniques can be applied to the variables.

Eg : Height, Weight are best examples.

In market research, ratio scale is used to calculate market share, annual sales, price of an upcoming product, the number of consumers, etc.

Scales of Measurement

Four levels of measurement in research and statistics:

Data	Nominal	Ordinal	Interval	Ratio
Labeled	Yes	Yes	Yes	Yes
Meaningful Order	No	Yes	Yes	Yes
Measurable Scale	No	No	Yes	Yes
True Zero Starting Point	No	No	No	Yes

Population and Sample

POPULATION

- A **Population** is a collection of all people, items, or events about which you want to make inferences. The size of the population can be finite or infinite

- Not always convenient or possible to examine every member of an entire population.

- **Parameter** is the value that describe the characteristics of the population eg - ***Population Mean*** (μ) & ***Population S.D*** ***(σ)***

Population Parameters (μ,σ)

↓

Sampling

↓

*Sample Statistics **(x(bar),σ**)*

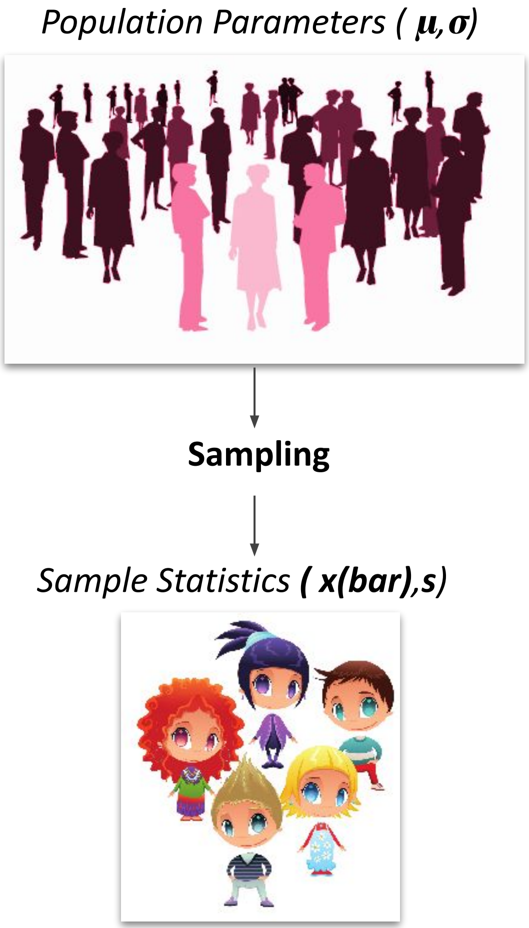
Population and Sample

SAMPLE

- A **Sample** is a representative subset of people, items, or events from a larger population that you collect and analyze to make inferences

- To represent the population well, a sample should be **randomly collected** and **adequately large**.

- **Statistic** is the value that describe the characteristics of the Sample eg - ***Sample Mean* (x(bar))** & ***Sample S.D (s)*** *are called Sample Statistics*



*Population Parameters (**μ,σ**)*

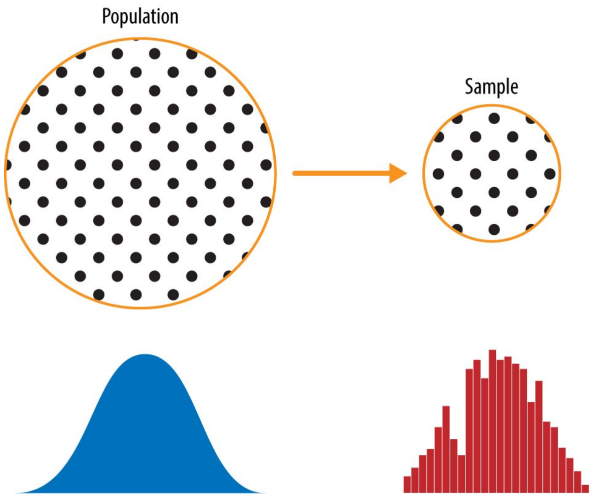
Sampling

*Sample Statistics **(x(bar),s**)*

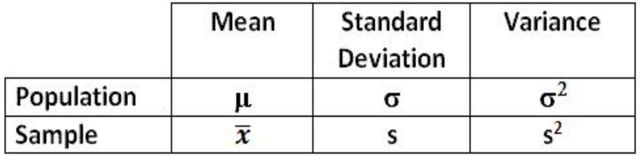
Basic Concepts for Statistics

Population and Sample

Basis	Population	Sample
Meaning	Population refers to the collection of all elements possessing common characteristics, that comprises universe.	Sample means a subgroup of the members of population chosen for participation in the study.
Includes	Each and every unit of the group.	Only a handful of units of population.
Characteristic	Parameter	Statistic
Data collection	Complete enumeration or census	Sample survey or sampling
Focus on	Identifying the characteristics.	Making inferences about population.

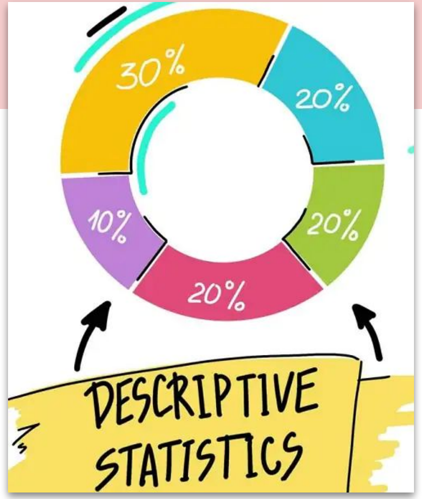


Notations for Population and Sample



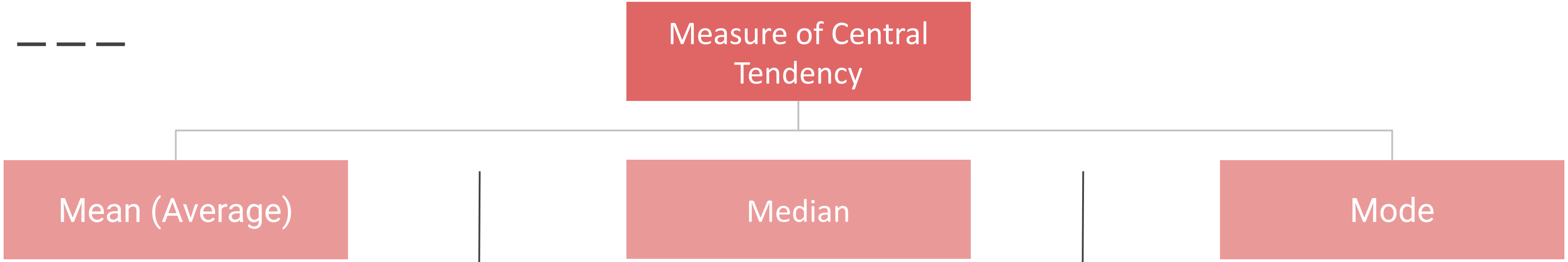
	Mean	Standard Deviation	Variance
Population	μ	σ	σ^2
Sample	$\bar{x}$	s	s^2

Descriptive Statistics



Measures of Central Tendency

Measure of Central Tendency is a single value that attempts to describe a set of data by identifying a central position within that set of data. There are 3 measures of Central Tendency -



Measure of Central Tendency

Mean (Average)

Median

Mode

- The arithmetic mean of a set of observations is their **average**.
- It is equal to the **sum of all observations divided by the number of observations** in the set.

- The median is an **observation in the center of the data** set **arranged** in ascending or descending order.
- **One-half of the data lie above this observation**, and one-half of the data lie below it.

- The mode of the data set is the **value that occurs most frequently.**

The mean, median and mode are all valid measures of central tendency, but under different conditions, some measures of central tendency become more appropriate to use than others.

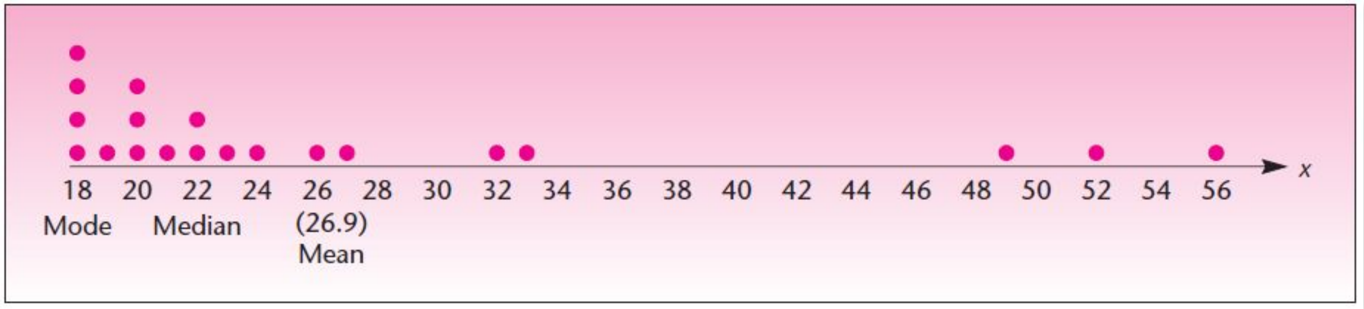
Measures of Central Tendency

$$\text{Mean} = \frac{(\mathbf{18+18+18+18}+19+20+20+20+21+\mathbf{22+22}+23+24+26+27+32+33+49+52+56)}{20}$$

Mean = 26.9

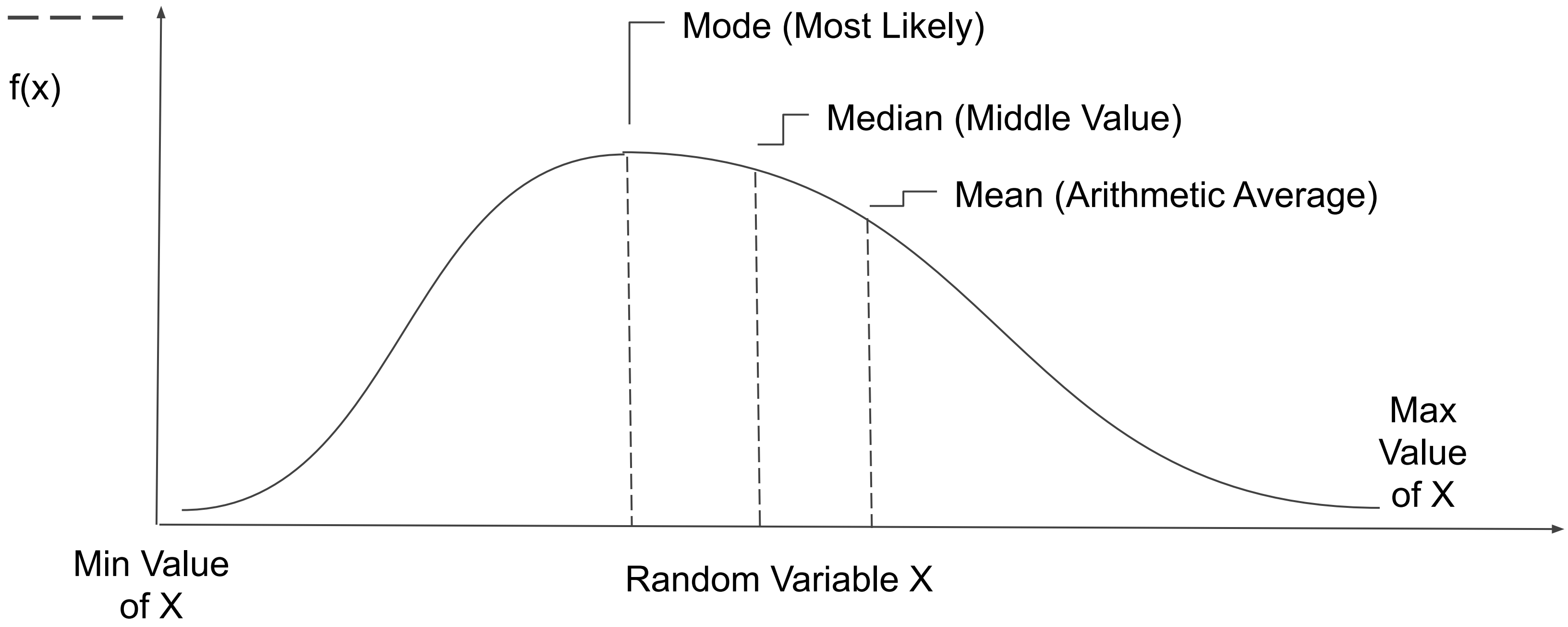
Median = 22

Mode = 18



Value	**Frequency**
18	4 (///)
19	1 (/)
20	3 (///)
21	1 (/)
22	2 (//)
23	1 (/)
24	1 (/)
26	1 (/)
27	1(/)
32	1 (/)
33	1 (/)
49	1 (/)
52	1 (/)
56	1 (/)

Measures of Central Tendency for a distribution

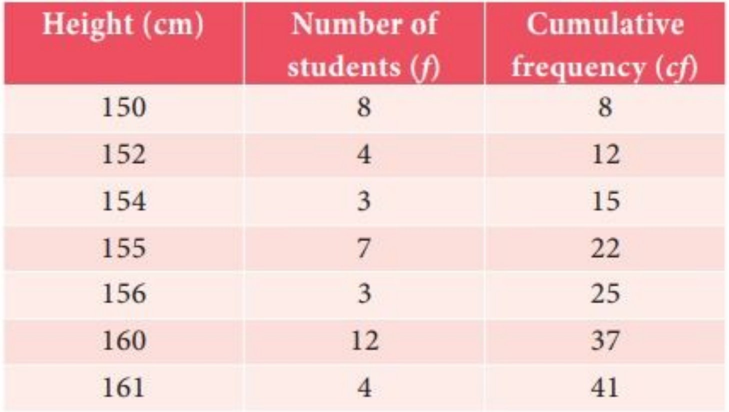


Frequency Distribution

- A frequency distribution is a **collection of observations produced by sorting observations** into classes and showing their frequency (f) of occurrence in each class.
- When observations are sorted into classes of single values, the result is referred to as a frequency distribution for **ungrouped data.**
- When observations are sorted into classes of more than one value, the result is referred to as a frequency distribution for **grouped data.**

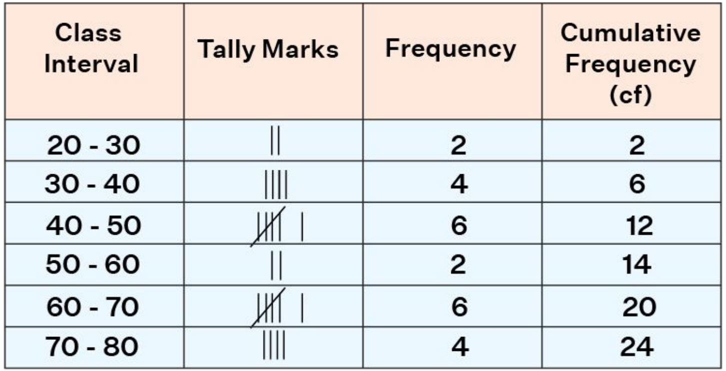
Guidelines to create frequency tables

- Each observation should be included in one, and only one, class.
- List all classes, even those with zero frequencies
- All classes should have equal intervals



Height (cm)	Number of students (*f*)	Cumulative frequency (*cf*)
150	8	8
152	4	12
154	3	15
155	7	22
156	3	25
160	12	37
161	4	41

Frequency Distribution of Ungrouped Data



Class Interval	Tally Marks	Frequency	Cumulative Frequency (cf)
20 - 30	||	2	2
30 - 40	||||	4	6
40 - 50	𝍸 |	6	12
50 - 60	||	2	14
60 - 70	𝍸 |	6	20
70 - 80	||||	4	24

Frequency Distribution of Grouped Data

Histogram and Frequency Polygon

Histogram

A histogram is a graph that uses bars to portray the frequencies or the relative frequencies of the possible outcomes for a quantitative variable.

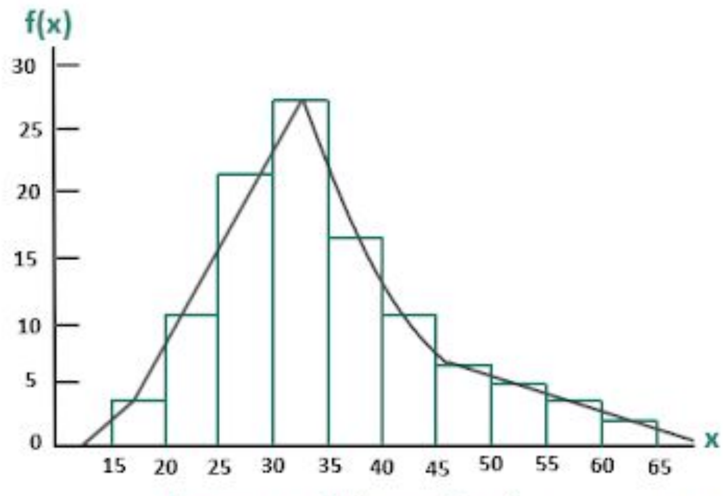
Bins are created on X-Axis and frequencies are plotted on Y-Axis

Unique things about Histogram

- It is a frequency distribution chart
- Looks like a bar chart but it is different
- There is no gaps between the bars representing continuity
- Created for one continuous variable ONLY
- Frequencies are plotted for bins/ intervals

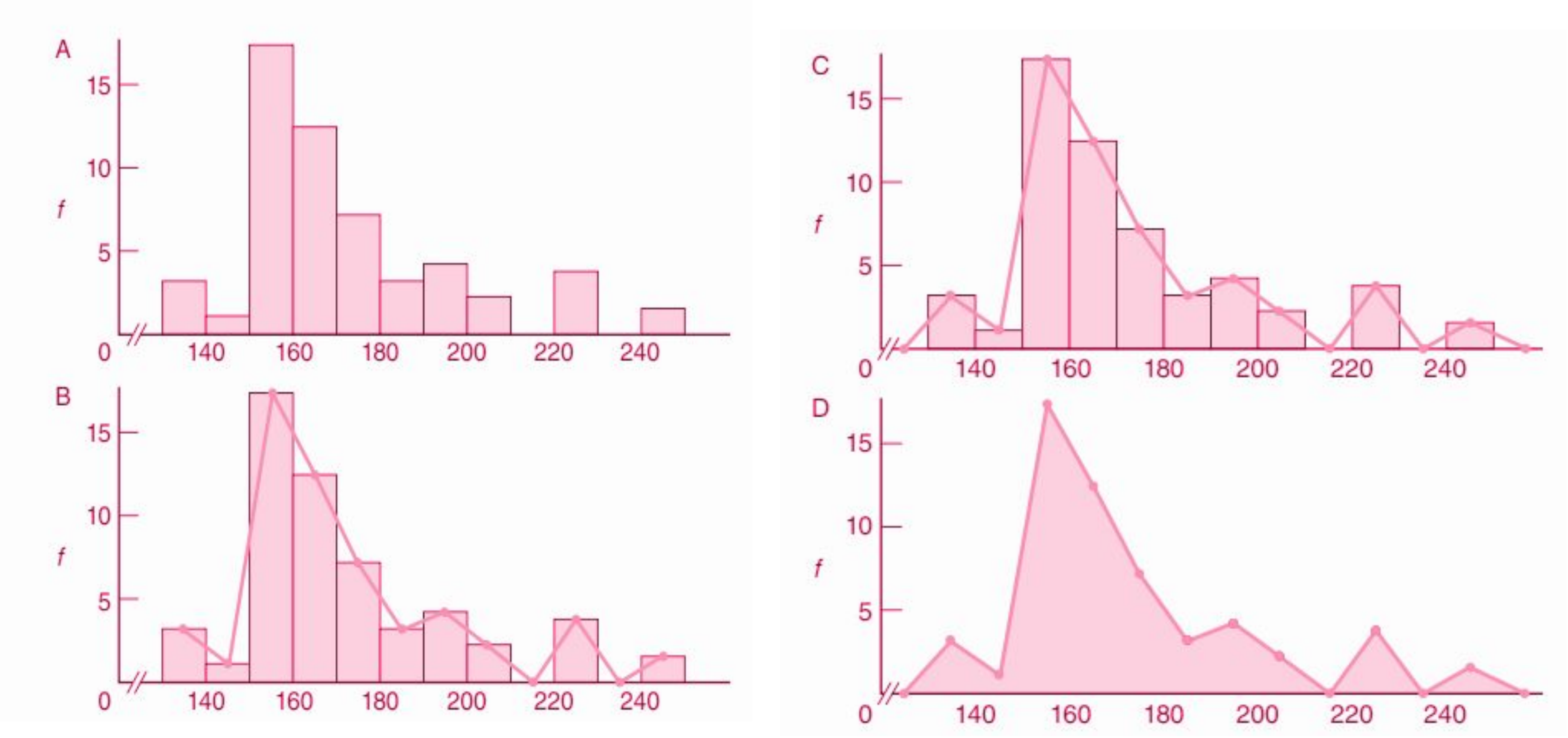
Frequency Polygon

A line graph for quantitative data that also emphasizes the continuity of continuous variables. It is created by joining the mid points of the Histogram.



Histogram and Frequency Polygon

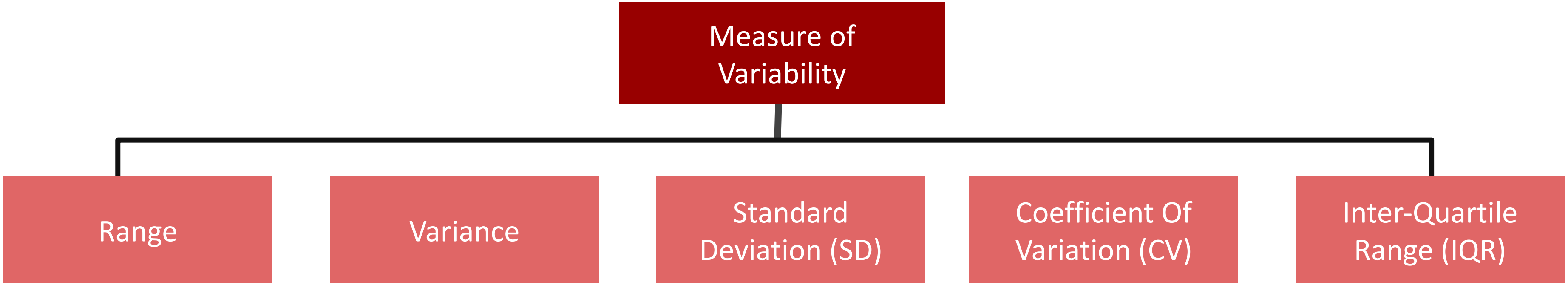
Transition from Histogram to Frequency Polygon



Measures of Variability/ Dispersion

How spread out are the values ?

- Measure of Variability is a '*Summary Statistics*' that attempts to describe the **Spread/ Diversity/ Variability/ Dispersion** in the distribution of data
- Defines how far away the data points tend to fall from the center
- It is used to identify the ***'Risk component'*** in the data
- We talk about variability in the context of a distribution of values
- *Low dispersion* indicates that the data points tend to be clustered tightly around the center (mean). *High dispersion* signifies that they tend to fall further away



Measure of Variability

Range

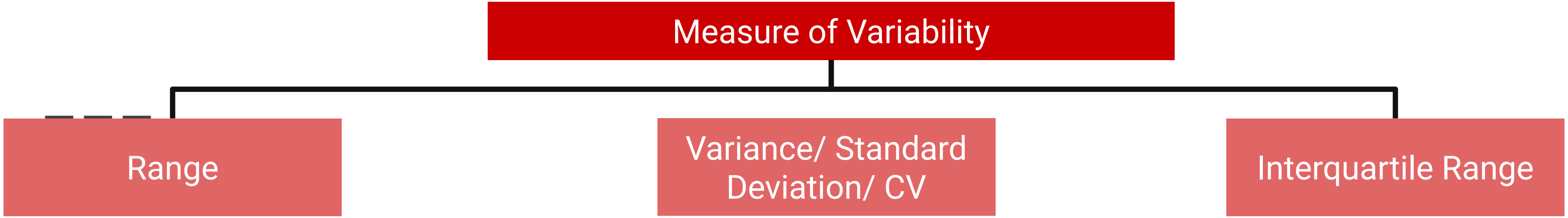
Variance

Standard Deviation (SD)

Coefficient Of Variation (CV)

Inter-Quartile Range (IQR)

Measures of Variability



Measure of Variability

Range

Variance/ Standard Deviation/ CV

Interquartile Range

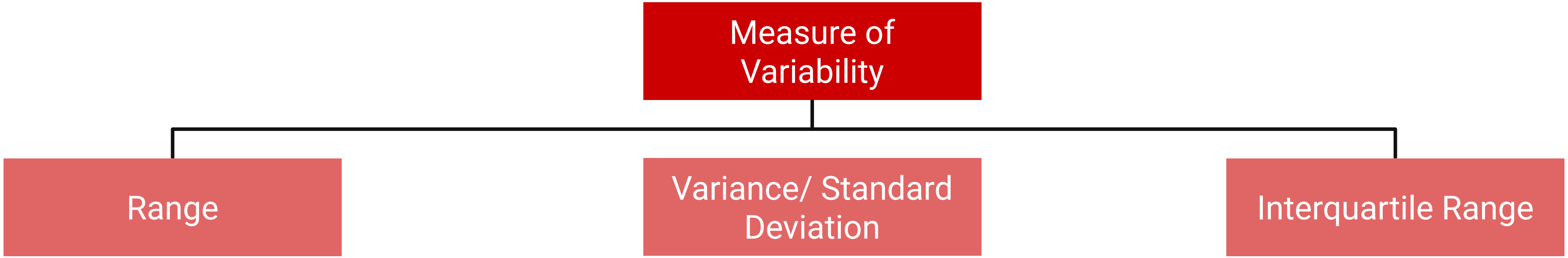
- The Range of a set of observations is the **difference between the largest and the smallest observation**.
- It uses the most extreme observations to define the scope of the distribution

- The Variance of a set of observations is the **Average Squared Deviation of the data points from their Mean.**
- Standard Deviation is the **Square Root of Variance.**
- Coefficient of Variation (CV) is the measure of Relative Variability. Used when we have to compare the SD of 2 or more Samples
- All of these are more useful as they use the information contained in all the observations of the sample dataset or population

- Interquartile Range is the **difference between the Third and First Quartile**.
- Interquartile range is more resistant to extreme observations compared to Range

Measures of Variability

Measure of Variability is a value that attempts to describe the **Diversity** or **Variability** in the distribution of data. It can also be used to identify the risk component in the data. Following are the different Measures of Variability used -



Measure of Variability

Range

Variance/ Standard Deviation

Interquartile Range

Range =
Max. Value (-)
Min Value

Population Variance

$$\sigma^2 = \frac{\sum_{i=1}^{N}(x_i - \mu)^2}{N}$$

Sample Variance

$$s^2 = \frac{\sum_{i=1}^{n}(x_i - \overline{x})^2}{n - 1}$$

Population SD

$$\sigma = \sqrt{\frac{\sum(x_i - \mu)^2}{N}}$$

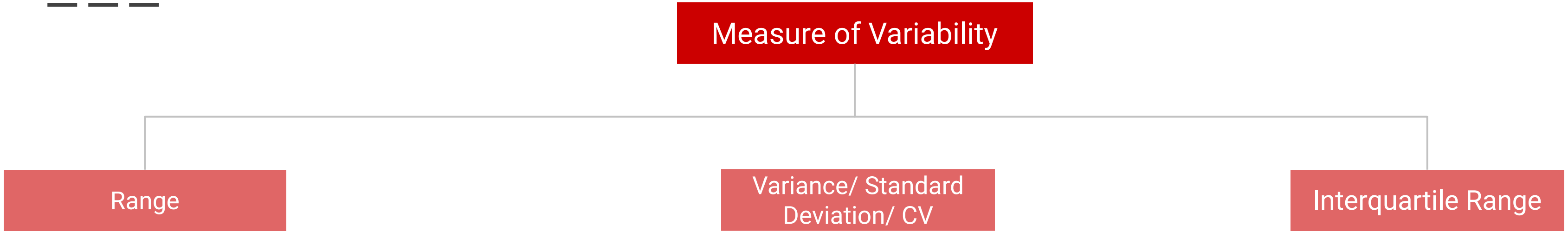
Sample SD

$$s = \sqrt{\frac{\sum\left(x - \overline{x}\right)^2}{n - 1}}$$

Interquartile Range =
Third Quartile [Q3] (-) First Quartile [Q1]

Measures of Variability

Measure of Variability is a value that attempts to describe the diversity or variability in the distribution of data. Following are the different Measures of Variability used -



Range =
Max. Value (-) Min Value

Population Variance

$$\sigma^2 = \frac{\sum_{i=1}^{N}(x_i - \mu)^2}{N}$$

μ is population mean

Sample Variance

$$s^2 = \frac{\sum_{i=1}^{n}(x_i - \overline{x})^2}{n - 1}$$

x(bar) is Sample mean
$S^2$is the sample SD

Coefficient of Variance (CV)

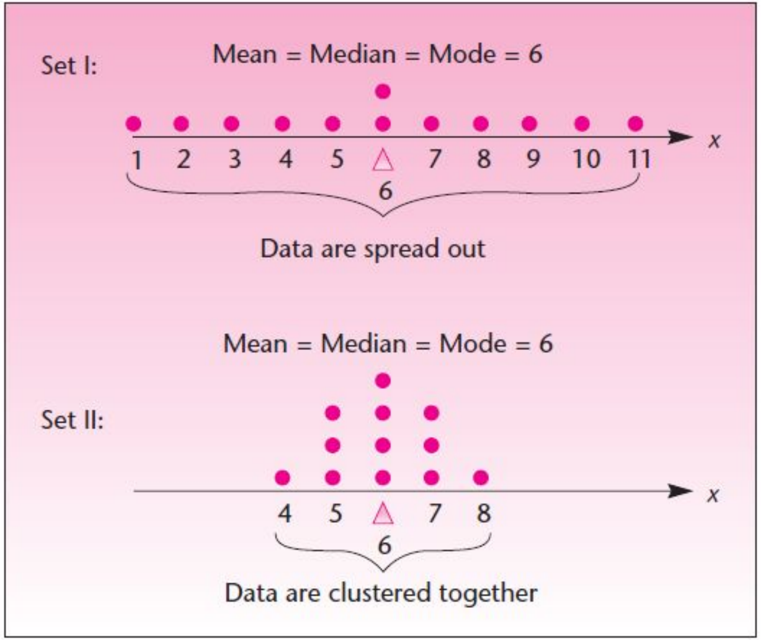
- Measures the relative variability
- Use when we compare the SD of 2 or more Samples
- Coefficient of Variation = (Standard Deviation / Mean) * 100
- CV = SD/ $\overline{X}$

Interquartile Range =
Third Quartile [Q3] (-) First Quartile [Q1]

Interquartile Range is used to identify the outliers. Values falling outside the range of 1.5 times the IQR usually are outliers

Measures of Variability

Set-I	1	2	3	4	5	6	6	7	8	9	10	11
Set-II	4	5	5	5	6	6	6	6	7	7	7	8

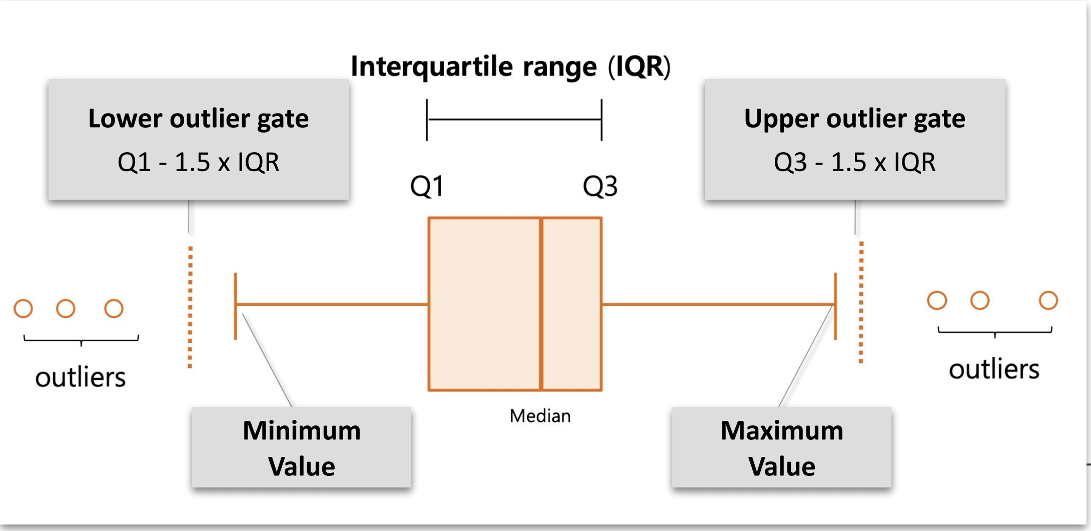


Observations

- The measures of Central Tendencies are exactly the same (=6) for both the sets of data
- Set-I data is spread out and Set-II data is closely clustered
- We can say that both the sets have same *Central Tendencies* but different *Variabilities* or *dispersion*

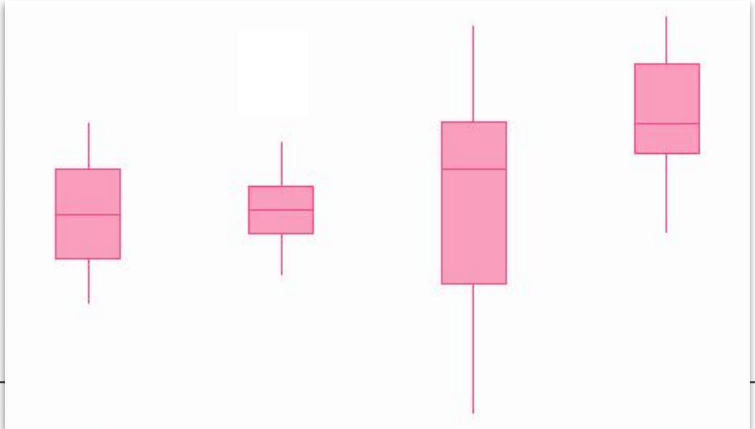
BoxPlot

- Boxplot is a graph that gives a good indication of **how the values are spread out** in dataset
- **Standardized way** of displaying the distribution of data based on - **Minimum, First quartile (Q1), Median, Third quartile (Q3), and Maximum**.
- The box in the boxplot is your InterQuartile Range (IQR). It contains 50% of your data. By comparing the size of these boxes, you can understand your data's variability. More dispersed distributions have wider boxes.



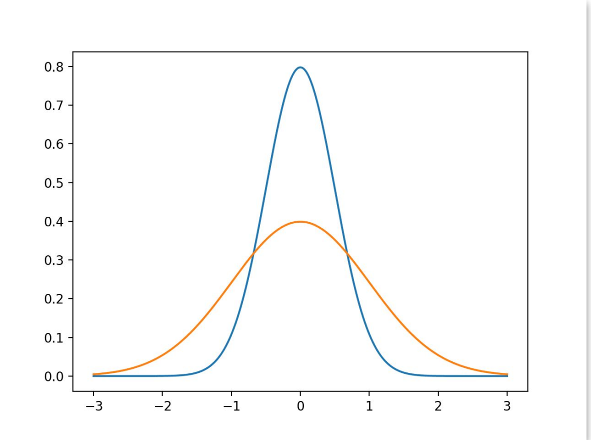
BoxPlot

- Additionally, find where the median line falls within each interquartile box. If the median is closer to one side or the other of the box, it’s a skewed distribution. When the median is near the center of the interquartile range, your distribution is symmetric.
- To find outliers, you’ll need to know your data’s IQR, Q1, and Q3 values. Take these values and input them into the equations below.
- Lower Outliers Gate: Q1 – 1.5 * IQR
- Upper Outliers Gate: Q3 + 1.5 * IQR



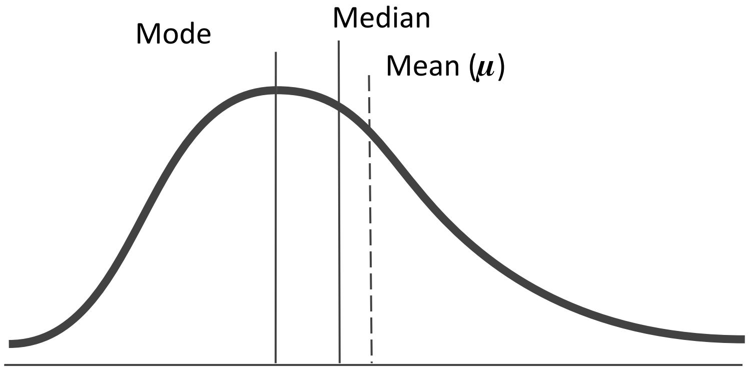
Understanding Variability and why it is important

- Variability is **pervasive.** You cannot escape it.
- While the mean is relevant, people often react to variability even more. When a distribution has lower variability, the values in a dataset are more consistent.
- However, when the variability is higher, the data points are more dissimilar and extreme values become more likely. Consequently, understanding variability helps you grasp the likelihood of unusual events.
- Extreme values can cause problems in some situations. Whether it is extreme temperature or extreme speed, we feel discomfort at extreme values than mean values.
- Some variation is inevitable, but problems occur at the extremes. Distributions with greater variability produce observations with unusually large and small values more frequently than distributions with less variability
- Central tendencies does not provide you the complete picture



Skewness & Kurtosis

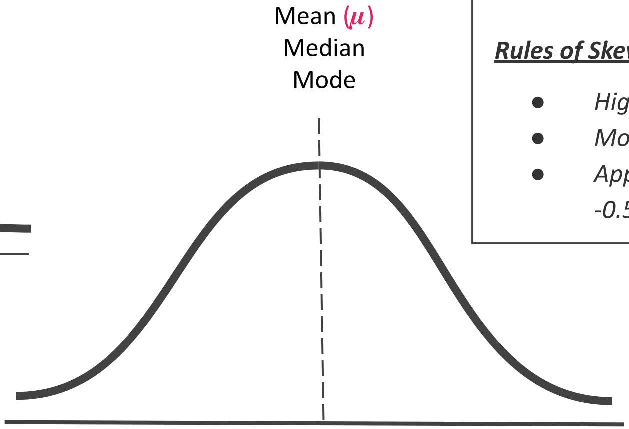
Skewness is the measure of the **degree of asymmetry** of frequency distribution.



Positively Skewed/ Right Skewed

Mean > Median > Mode

Skewness is +ve Number



Symmetrical Distribution (Normal)

Mean = Median = Mode
Skewness is Zero

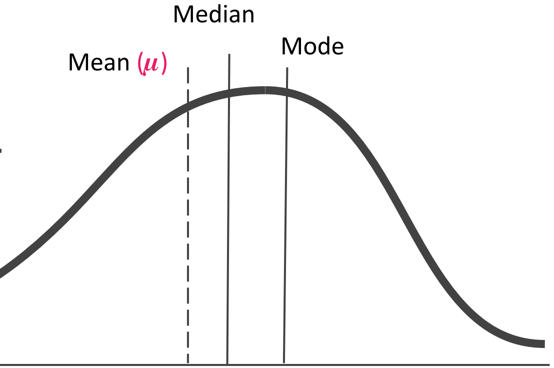
Formula for calculating the Skewness for a population is

$$\sum_{i=1}^{N}\left[\frac{x_i - \mu}{\sigma}\right]^3 \Big/ N.$$

Rules of Skewness

- *Highly Skewed when <(-1) or >1*
- *Moderately Skewed when <(-0.5) or > 0.5*
- *Approximates Normal or Symmetric when between -0.5 and 0.5*

Two distributions that have the same mean, variance, and skewness could still be significantly different in their shape. We may then look at their **Kurtosis**.

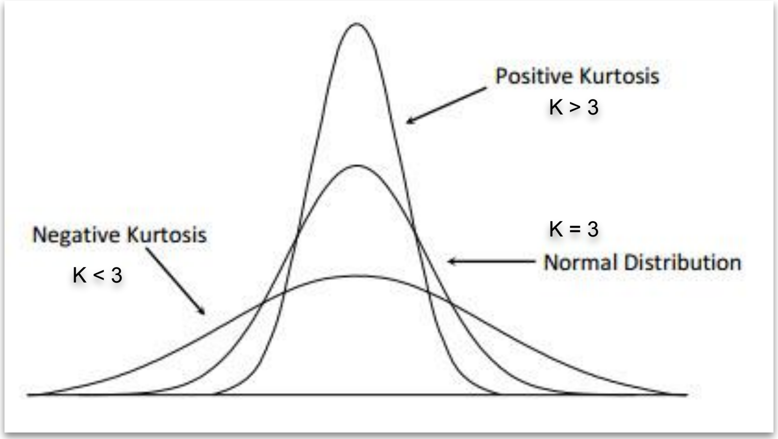


Negatively Skewed/ Left Skewed

Mean < Median < Mode / Skewness is a -ve Number

Skewness & Kurtosis

Kurtosis is the measure of peakedness of a distribution.



Formula for calculating the Kurtosis for a population is

$$\sum_{i=1}^{N}\left[\frac{x_i - \mu}{\sigma}\right]^4 \Big/ N.$$

- Larger the Kurtosis more peaked the distribution
- Kurtosis is calculated and reported either as an absolute or a relative value
- Absolute Kurtosis is always a positive number and Relative Kurtosis can be a negative number. But usually we work with **Relative Kurtosis** only
- Absolute Kurtosis for a Normal Distribution is 3.
- **Relative Kurtosis = Absolute Kurtosis (-) 3.**
- Negative Kurtosis implies a flatter distribution compared to Normal distribution and it is called ***Platykurtic*** distribution
- Positive Kurtosis implies more peaked distribution compared to Normal distribution and it is called ***Leptokurtic*** distribution

In general we look at the Skewness and treat the data. Kurtosis is not used much in data modelling to treat data.

Normal Curve (Bell Curve)

Relation between Mean and Standard Deviation.

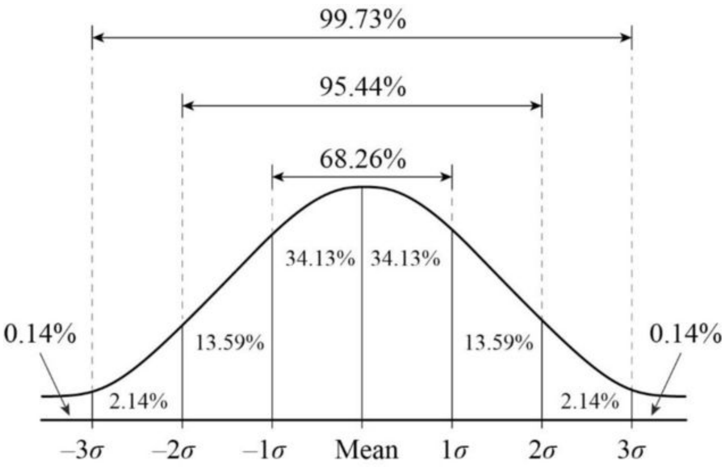
Normal distribution, also known as the Gaussian distribution, is a probability distribution that is symmetric about the mean, showing that data near the mean are more frequent in occurrence than data far from the mean. In graph form, normal distribution will appear as a bell curve.

Standard Normal Distribution has two parameters: the Mean(x-bar) and the Standard Deviation (SD).

For a Normal Distribution -

- 68% of the observations are within +/- one SD of the mean
- 95% are within +/- two SD, and
- 99.7% are within +- three SD

Normal distribution model is motivated by the Central Limit Theorem (which is the backbone of inferential statistics). This theory states that averages calculated from independent, identically distributed random variables have approximately normal distributions, regardless of the type of distribution from which the variables are sampled

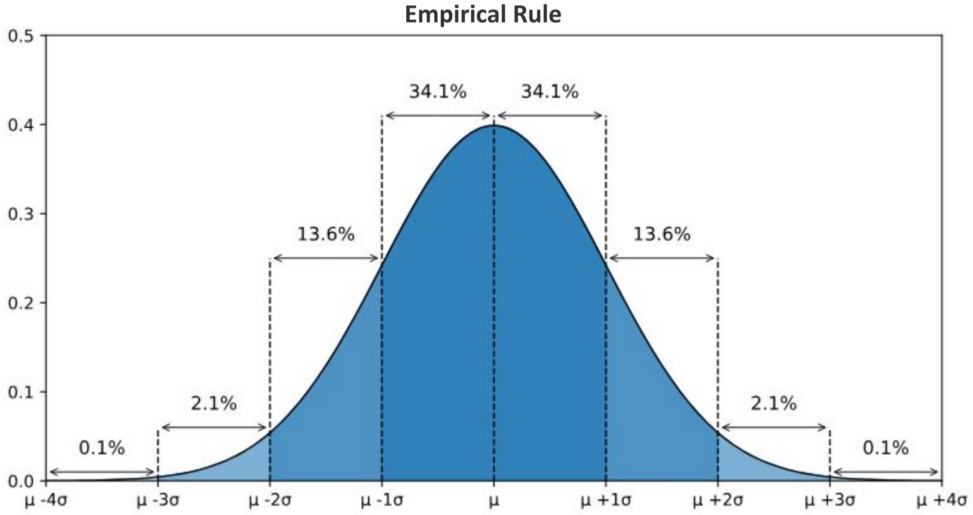


Empirical Rule/ 3 Sigma Rule

Relation between Mean and Standard Deviation.

The **Empirical Rule** states for a Symmetric data (normal distribution or mountain shaped) -

- **~ 68%** of the observations lie within **1** S.D. of the mean.
- **~ 95%** of the observations lie within **2** S.D. of the mean.
- A vast majority of the observations (all, or all most all) will be within **3** S.D. of the mean

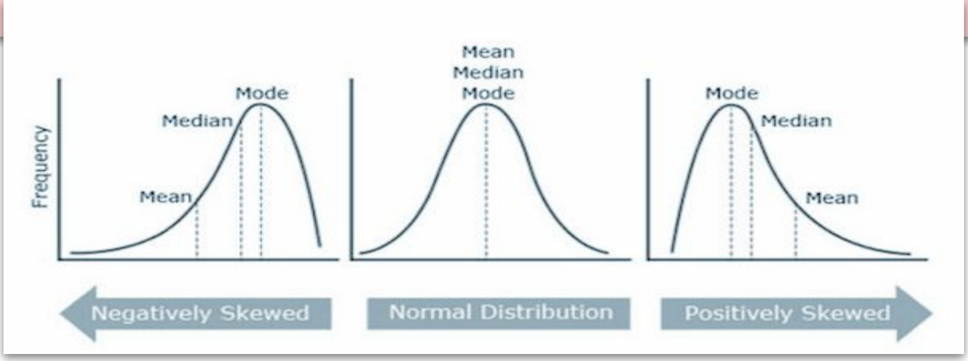


Which is the best Measure - Mean Median and Mode

The 3 main measures of central tendency are **best used in combination** with each other to understand the data because they have **complementary strengths and limitations**. But sometimes only 1 or 2 of them are applicable to your dataset. The measure of central tendency to be used in a situation **depends on -**

- Level of measurement
- Type of data
- Type of distribution
- Skewness of the data
- Presence of outliers

More importantly the **purpose of calculating** the measure cannot be ignored i.e. **'business goals'** need to be kept in mind. Thus the choice of using mean or median should primarily driven by the goals of analysis.



Which is the best Measure - Mean Median and Mode

Based on the level of measurement

- The **Mode** can be used for any level of measurement, but it’s most meaningful for **nominal** and **ordinal** levels.
- The **Median** can only be used on data that can be **ordered** – that is, from ordinal, interval and ratio levels of measurement.
- The **Mean** can only be used on **interval** and **ratio** levels of measurement because it requires equal spacing between adjacent values or scores in the scale.

Level	Measure
Nominal	Mode
Ordinal	Mode, Median
Interval Ratio	Mode, Median and Mean

Distribution	Measure
Normal	Mode, Median, Mean
+ve Skewed	Median
-ve Skewed	Median

You should also consider the the type of distribution to decide which measure of central tendency to be used -

- For **Normally distributed data,** all three measures of central tendency will give you the same answer so they can all be used.

- In **Skewed distributions,** the **Median is the best** measure because it is unaffected by extreme outliers or non-symmetric distributions of scores. The **Mean** and **Mode** can vary in skewed distributions.

Mean Median and Mode - A comparison

Measure of Central Tendency	Pros	Cons
Mean	● Based on all observations ● Easy to calculate and understand ● Least affected by sampling fluctuations hence more stable ● Arrangement of data not needed	● Can be used only for quantitative data ● Affected by extreme values ● Graphical representation not possible
Median	● Not affected by extreme values ● Can be calculated for Quantitative and Qualitative data ● Graphical representation possible	● Affected by sampling fluctuations ● Does not use all values to calculate ● Arranging data is mandatory (ascending or descending)
Mode	● Not affected by extreme values ● Can be calculated for Quantitative and Qualitative data ● Graphical representation possible	● Does not use entire data ● Not necessary a unique value ● Not always descriptive of entire dataset

Data visualization - definition and importance

Data Visualization

Data Visualization is a **graphical representation** of data and plays a vital role in understanding information in a better way. It is a way to represent data in visual content.

Why Data Visualization

- Bird's eye view of the entire data
- With big data, charts provide the big message in concise manner
- Understanding the outliers become easy
- Maintains audience interest with charts
- Human brain understand visuals easily, absorbs message from visuals quickly and retains them for a longer period of time

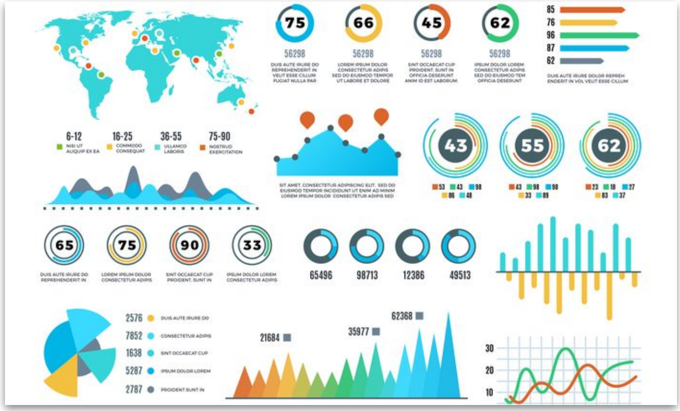
Choosing the right chart

- Choosing the right visualization is paramount whether you’re presenting to a leader or client. Incorrect representation can lead to a wrong message or wrong decision taken by the stakeholder.

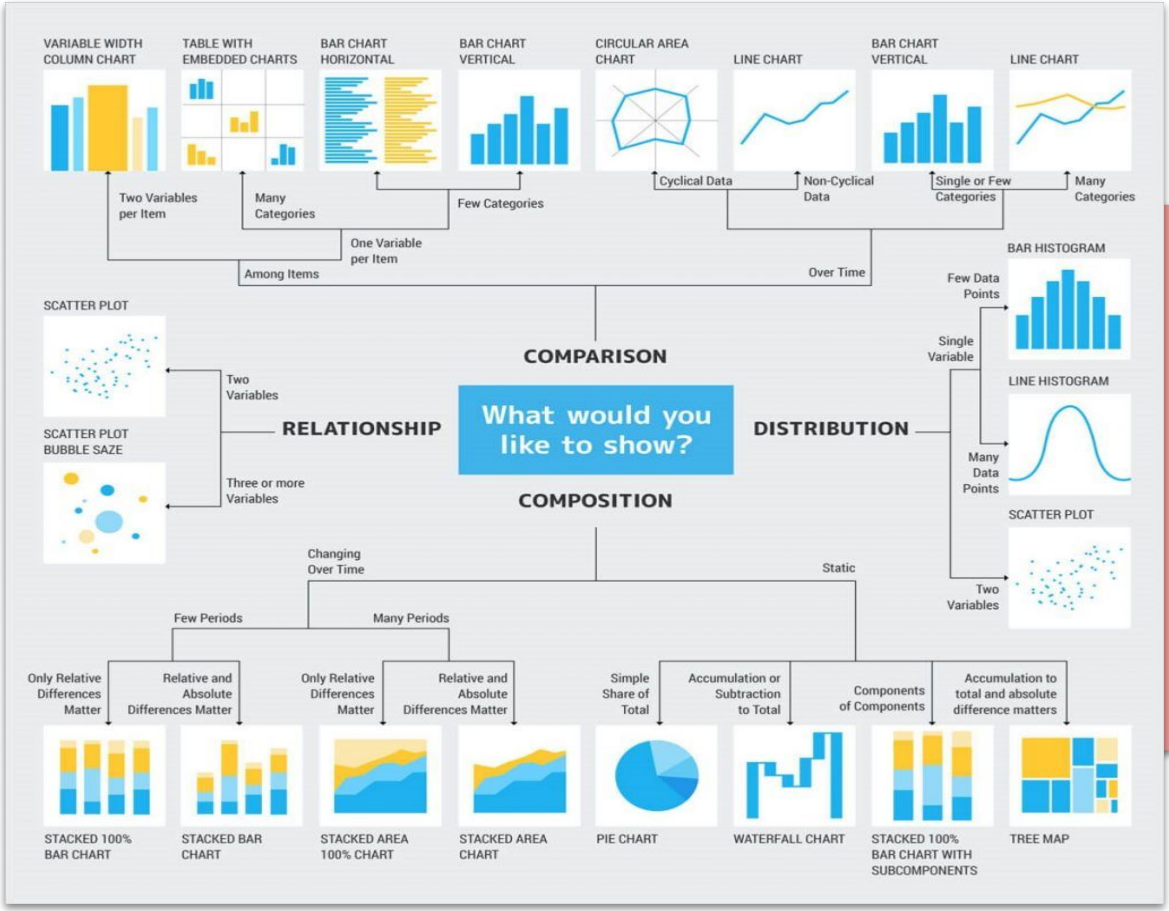
- While preparing the charts, your focus should be on conveying the right message to your audience in an optimal way.

The right chart depends on the message you want to deliver to the audience. A data analyst usually would want to disseminate the below messages -

a) Comparison (includes discrete and continuous comparisons)
b) Distribution
c) Composition (Break up of whole)
d) Relationship



Types of Charts



Comparison charts - Facilitate comparison of values for discrete or continuous variables. The purpose to show how a variable is with respect to other or over the period of time.

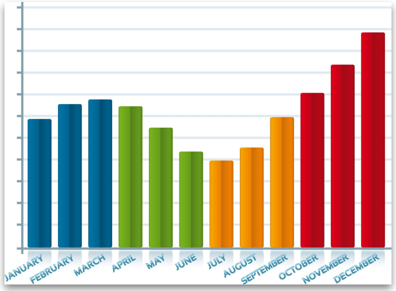
Distribution charts - Used to show how variables are distributed over time or otherwise, helping identify outliers and trends. This includes histogram, scatter and even box plot.

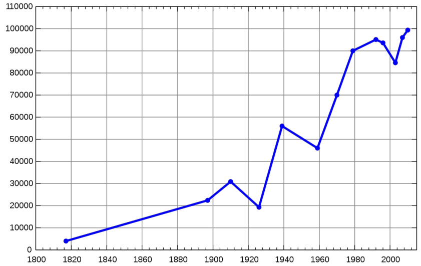
Composition charts - Composition charts are used to display parts of a whole and change of variable over the period of time.

Relationship charts - Used to show a connection or correlation between two or more variables over parameters.

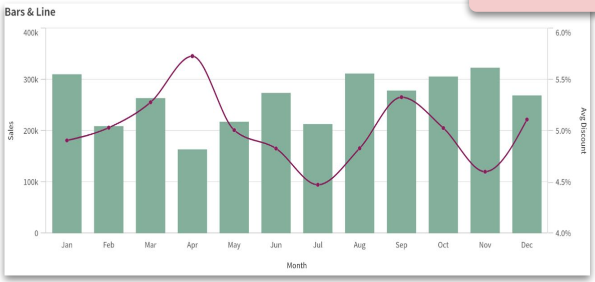
Comparison Charts

Bar charts (Horizontal / Vertical) - In a vertical bar chart the X-axis represents a categorical variable, while the Y-axis represents value of the variable. The length of bars gives the idea of maximum and minimum value with respect to the category. The horizontal bar chart reverses the the orientation. Rest everything remains the same. There are different variations of bar chart which are beyond the scope of this training.





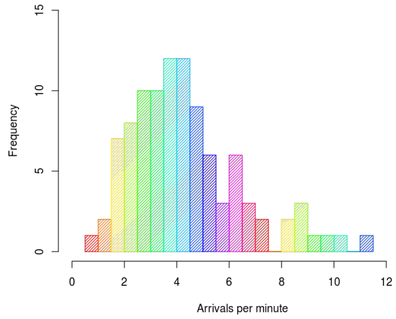
Line Chart - A line chart graphically displays changes in variables continuously over time. Each line graph consists of points that connect data to show a trend (continuous change). Line graphs also have an x-axis and a y-axis. Time is shown on the x- axis in most cases. Line and Bar chart can also be used together to create **combo charts** which show 2 variables on different scales.

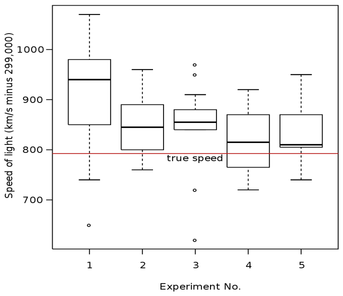


Note - *Time variable can be shown as a bar or a line chart depending on the analysis and the message you want to deliver. Time can be used as a discrete variable (time buckets to facilitate comparison) or continuous variable (over time line to show trend over the period of time).*

Distribution Charts

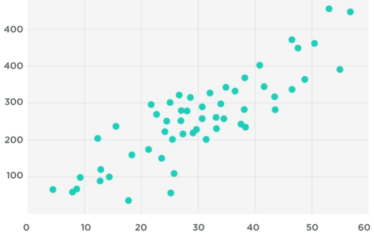
Histograms - Chart that groups a numeric (continuous) data into bins, displaying the bins as segmented columns. They're used to depict the distribution of a variable. Histograms are plotted from frequency distributions and are used to create the distribution plots as we increase the size of the bins max or infinity. Note that Histograms are different from bar charts fundamentally.





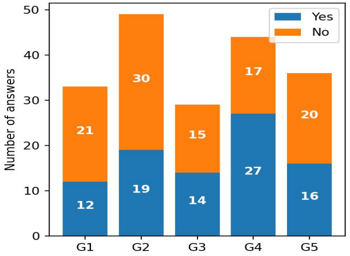
Box Plot charts - Used to show the spread and the intensity (density) of the variable across the 4 quartiles. It is one unique chart that is used to understand the reason of the variance and also locate the outliers in the data. Box plot created over the axis/ scale can also help you understand if the data is skewed or not skewed.

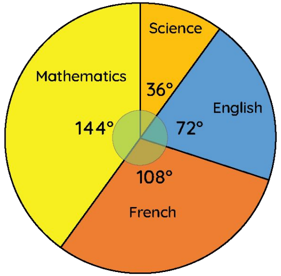
Scatter plots - Used to shows the relation between 2 continuous variables, each plotted on each axis. The spread of the data points show the broad trend across the 2 variables hence can be used to estimate the correlation between the variables. The scatter plot is also used to create a Regression model used to predict the dependent variable by modelling relation with independent variable.



Composition Charts

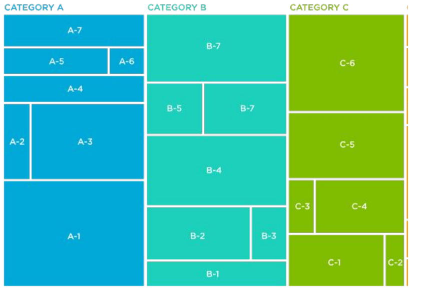
Stack bar charts - A stacked bar chart is a form of bar chart that shows the composition and comparison of a few variables, either relative (called **100% stacked bar chart**) or absolute, over time. Also called a stacked bar or column chart, they look like a series of columns or bars that are stacked on top of each other.





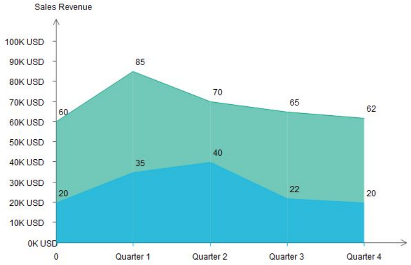
Pie charts - Pie was invented in early 1800's. It is a chart that shows the parts as a whole or the composition. It can show numbers in %, and as values. The total of the divided segments equals 100 %. There are different variables of pie chart like - doughnut chart, 3D pie chart, semi circle pie chart and even irregular pie chart.

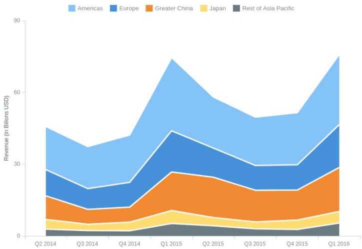
Tree maps - In a vertical bar chart the X-axis represents a categorical variable, while the Y-axis represents value of the variable. The length of bars gives the idea of maximum and minimum value with respect to the category. The horizontal bar chart reverses the the orientation. Rest everything remains the same. There are different variations of bar chart which are beyond the scope of this training.



Composition Charts

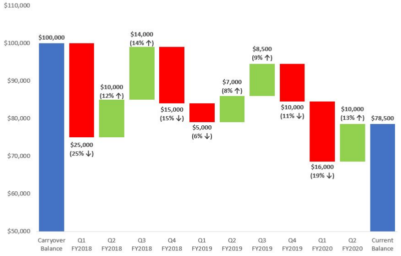
Area chart - An area chart is a line chart with the areas below the lines filled with colors. It displays graphically quantitative data. The area between axis and line are commonly emphasized with colors, textures and hatchings. Commonly one compares two or more quantities with an area chart. It represents the change in one or more quantities over time.





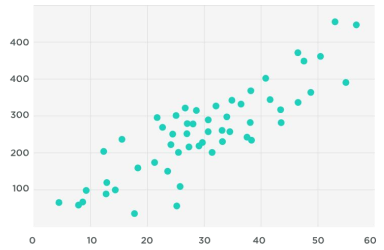
Stacked Area chart - In a stacked area chart, all of the lines are stacked on top of each other over a straight baseline at the bottom of the stack. The height of each series is determined by the value in each data point.A typical use case for Stacked Area Charts is analyzing how each of several variables and their totals vary. The Stacked Area Chart type only works well when more than two series are present.

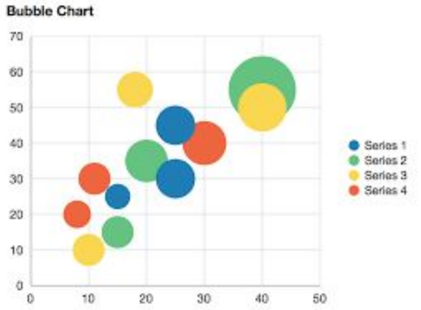
Waterfall chart - A waterfall chart is a specific type of bar chart that reveals the story behind the net change in something's value between two points. It helps in understanding the cumulative effect (running total) of sequentially introduced positive or negative values. The columns are color coded so you can quickly tell positive from negative numbers. Column bar chart shows the total in the end.



Relationship Charts

Scatter plots - Used to shows the relation between 2 continuous variables, each plotted on each axis. The spread of the data points show the broad trend across the 2 variables hence can be used to estimate the correlation between the variables. The scatter plot is also used to create a Regression model used to predict the dependent variable by modelling the relation with independent variable.

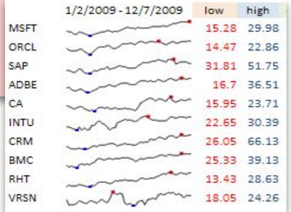




Scatter Plots with bubble - We can add another 3rd dimension to represent another variables to depict the size of the scatter points. These are represented as bubbles in a scatter plot. Hence this can help in the analysis of more than 2 variables. You can also another another categorical variable by adding color to the bubbles and this will lead to an addition of 4th variable.

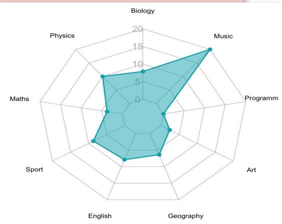
Other interesting Charts

Spark lines - A sparkline is a very small line chart, typically drawn without axes or coordinates. It presents the general shape of the variation (typically over time). Used a lot in stock market charts to show the general trend of the stocks.

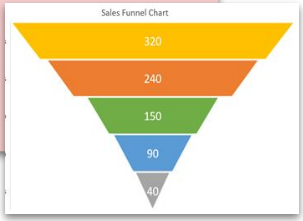


	1/2/2009 - 12/7/2009	low	high
MSFT		15.28	29.98
ORCL		14.47	22.86
SAP		31.81	51.75
ADBE		16.7	36.51
CA		15.95	23.71
INTU		22.65	30.39
CRM		26.05	66.13
BMC		25.33	39.13
RHT		13.43	28.63
VRSN		18.05	24.26

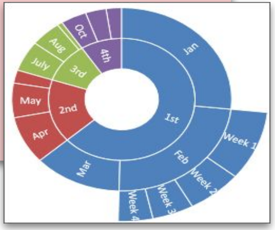
Radar Chart - A radar chart shows multivariate data of three or more quantitative variables mapped onto an axis. It looks like a spider's web, with a central axis that has at least three spokes, called radii, coming from it.



Funnel Chart - Funnel charts show values across multiple stages in a process. Funnel chart show the number of sales prospects at each stage in a sales pipeline.



Sunburst - Sunburst diagrams shows hierarchy through a series of rings, that are sliced for each category node. Each level of the hierarchy is represented by one ring or circle with the innermost circle as the top of the hierarchy.



Percentile/ Quantile

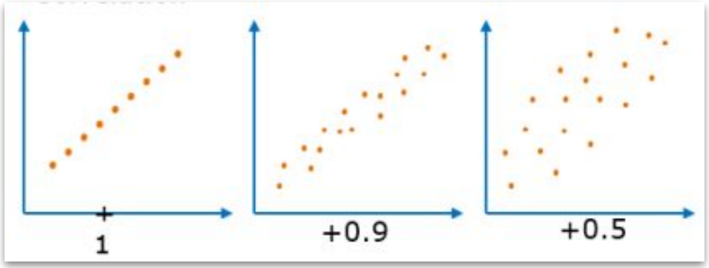
- A **percentile** is a term used in statistics to express how a score compares to other scores in the same set
- A **percentile** (or a centile) is a measure used in statistics indicating the value below which a given percentage of observations in a group of observations falls
- Percentiles indicate the percentage of scores that fall below a particular value.
- Percentile gives the position (not the value) of the participant after sorting them in order
- One of the way to identify an outliers in the data
- *For eg - the 20th* ***percentile*** *is the value (or score) below which 20% of the observations may be found.*
- *The* ***percentile*** *value for the topper in any competitive exam is 100th as 100% of the participants fall below him*

Correlation - Describing the relationships

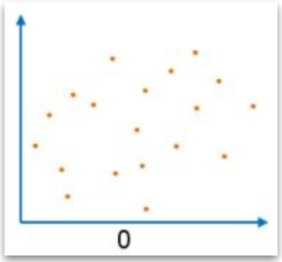
“Correlation” is a statistical term describing the **degree** to which **two variables move in coordination with one-another.** It measures both the **direction** and the **strength** of the relation. If the two variables move in the same direction, they are said to have a **positive correlation**. If they move in opposite directions, then they have a **negative correlation**. Understanding that relationship is useful because we can use the value of one variable to predict the value of the other variable. **Correlation** is usually used in the concept of ‘Multicollinearity’ to identify relation between 2 independent variables and identify the variables that can be eliminated in Regression analysis.

- The **Correlation Coefficient's** (represented by **r**) values range between **-1.0** and **1.0.**
- **r = -1 means** perfectly Negative correlation
- **r = 0 means** No correlation. No linear relation exists
- **r = 1 means** perfectly Positive correlation

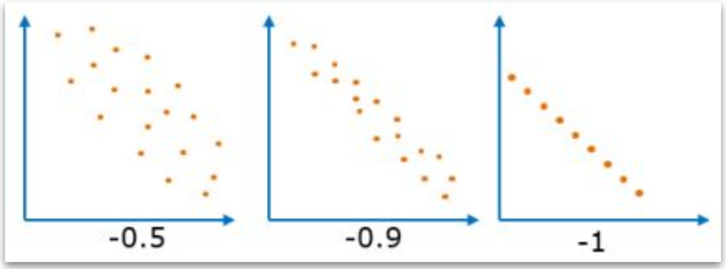
Positive Correlation



No Correlation



Negative Correlation



Correlation ≠ Causation

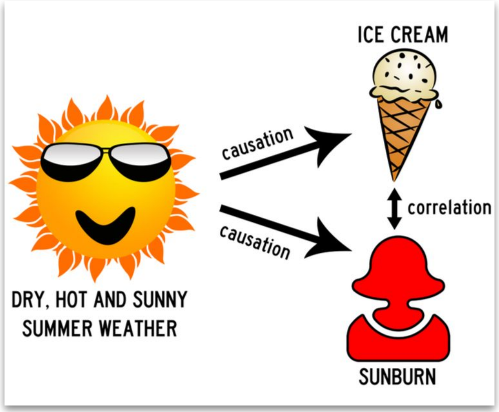
- Correlation does not measure the ‘Causation’ i.e. if there exist a ‘cause’ and ‘effect’ relation between the variables.

- Correlation describes the **degree of association between two or more variables.** It indicates how closely the variables move together or change in relation to each other.

- **While correlation can suggest a potential relationship, establishing causation requires additional evidence and careful consideration of other factors.**

- Causation, on the other hand, refers to a **cause-and-effect relationship** between variables. It suggests that one **variable directly influences or causes a change in another variable.**

- Correlation can arise due to various factors, such as coincidence, common underlying causes or variable, or other variables not considered in the analysis. Therefore, establishing causation requires further investigation to determine the mechanisms and potential confounding factors involved.



Correlation ≠ Causation - Examples

Examples

1. Let's say you are studying the relationship between exercise and weight loss. You collect data from a group of individuals and plot a scatter plot. As you analyze the data, you observe a strong negative correlation between the ‘number of hours’ spent exercising per week and the individuals' weight. This means that as the number of hours of exercise increases, the weight tends to decrease. However, correlation alone does not prove causation. It could be that those who exercise more also tend to make healthier food choices, which could be the actual cause of weight loss.

2. Ice cream sales and crime rates: There is a correlation between ice cream sales and crime rates. This means that when ice cream sales go up, crime rates tend to go up. However, this does not mean that ice cream sales cause crime rates to go up. There is a third variable, such as hot weather, that could be causing both ice cream sales and crime rates to go up.

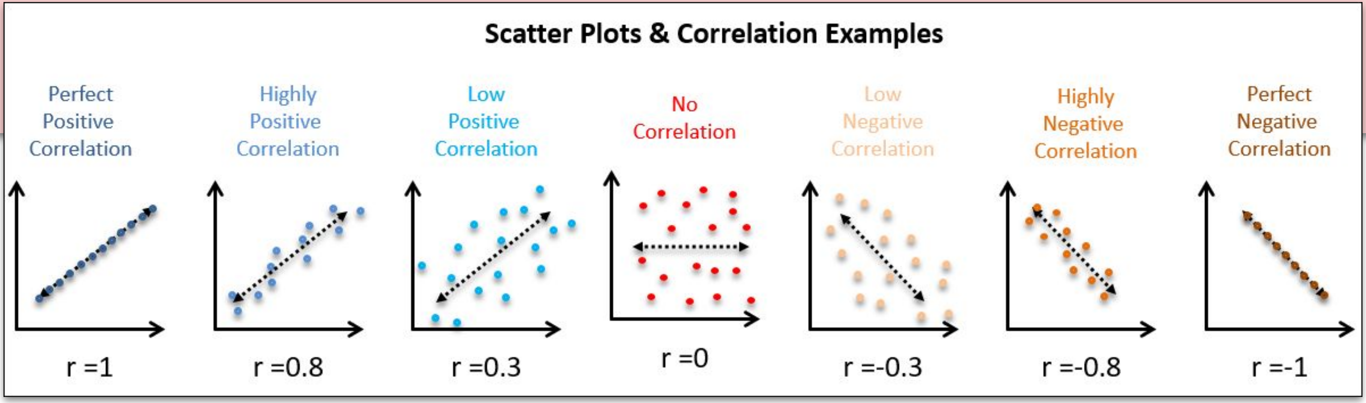
3. Time spent on social media and depression: There is a correlation between time spent on social media and depression. This means that people who spend more time on social media tend to be more depressed. However, this does not mean that social media causes depression. There are other factors that could be contributing to this correlation, such as personality traits or social isolation.

Correlation and Scatter plots

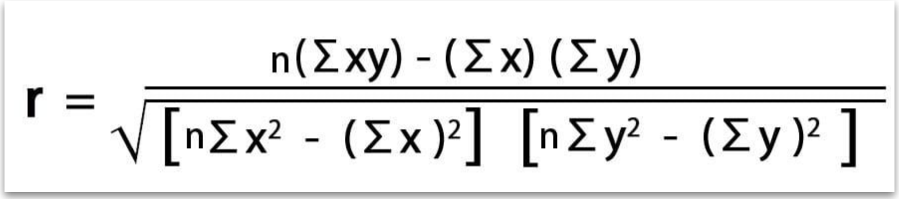
- A scatter plot is a graphical representation of paired data points from two variables. It consists of a horizontal x-axis representing one variable and a vertical y-axis representing the other variable. Each data point is plotted as a dot on the graph, with its position determined by the values of the two variables.

- Scatter plots allow us to observe the pattern and distribution of the data points, as well as any potential relationship between the variables.

- By combining the scatter plot with correlation analysis, we can visually and quantitatively assess the relationship between variables. Scatter plots and correlation analysis provide valuable insights into the relationship between variables, allowing researchers and analysts to understand the strength and nature of the association.



Computation for Correlation



$$\mathbf{r} = \frac{n(\Sigma xy) - (\Sigma x)(\Sigma y)}{\sqrt{\left[n\Sigma x^2 - (\Sigma x)^2\right]\left[n\Sigma y^2 - (\Sigma y)^2\right]}}$$

Where -

r = Correlation Coefficient

n = number of pairs of observations

$\boldsymbol{\Sigma}$ x = Sum of x observation

$\boldsymbol{\Sigma}$ y = Sum of y observation

$\boldsymbol{\Sigma}$ x^2 = Sum of square of x observation

$\boldsymbol{\Sigma}$ y^2 = Sum of square of y observation

Covariance

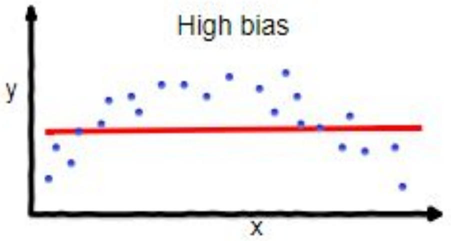
- **Covariance** measures how the two variables move concerning each other and is an extension of the concept of variance (which tells about how a single variable varies). It can take any value from - infinity to infinity.
- The higher the value of covariance, the more dependent the relationship is. A positive number signifies positive Covariance, that **means that an increase or variation will correspondingly increase in the other variable**, provided other conditions remain constant.
- A negative number signifies negative Covariance, which **denotes an opposite relationship between the two variables.** Though Covariance is perfect for defining the type of relationship, it is **not good for interpreting its magnitude.**

Bias and Variance

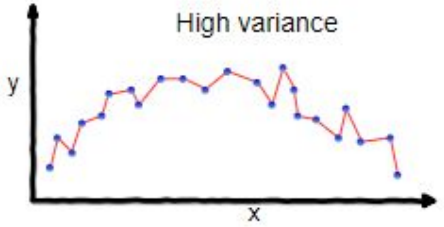
Bias is the difference between the Predicted Value and the Expected Value (average prediction of our model). Model with high bias pays very little attention to the **training data and oversimplifies the model.** It always leads to high error on training and test data. The model **makes certain simplified assumptions when it is trained on the data provided.** Bias is with respect to the training data. When it is introduced to the testing/ validation data, these assumptions may not always be correct.

Variance is the variability of model prediction for a given data point (or value) which indicates the spread of our data. Model with high variance pays a lot of attention to training data and does not generalize on the data which it hasn't seen before. As a result, **such models perform very well on training data but has high error rates on test data**. **Variance is with respect to the Test data.**

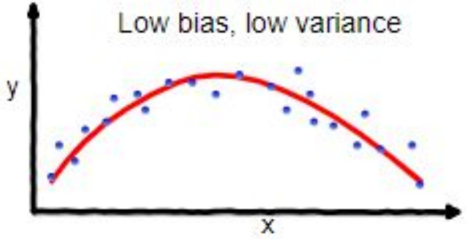
Noise is the Irreducible error i.e. the error that can't be reduced by creating good models. No matter how good we make our model, our data will have certain amount of noise or irreducible error that can not be removed.



Underfitting data



Overfitting data



Balanced data

Bias and Variance

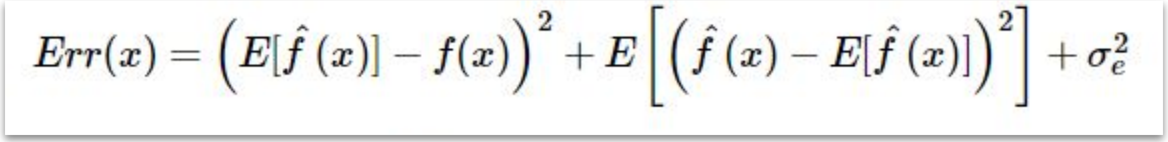
Let us say we are trying to predict Y as a function of X. The relationship between the two is defined by the function below. Here **‘e’ is the error term which is Normally distributed with a Mean 0.**

Y=f(x) + e

The expected Error (e) and the expected squared error at a point x is given below

Error(x) = E[(Y - f(x))^2]

Further it can be decomposed into - Bias, Variance and Irreducible error

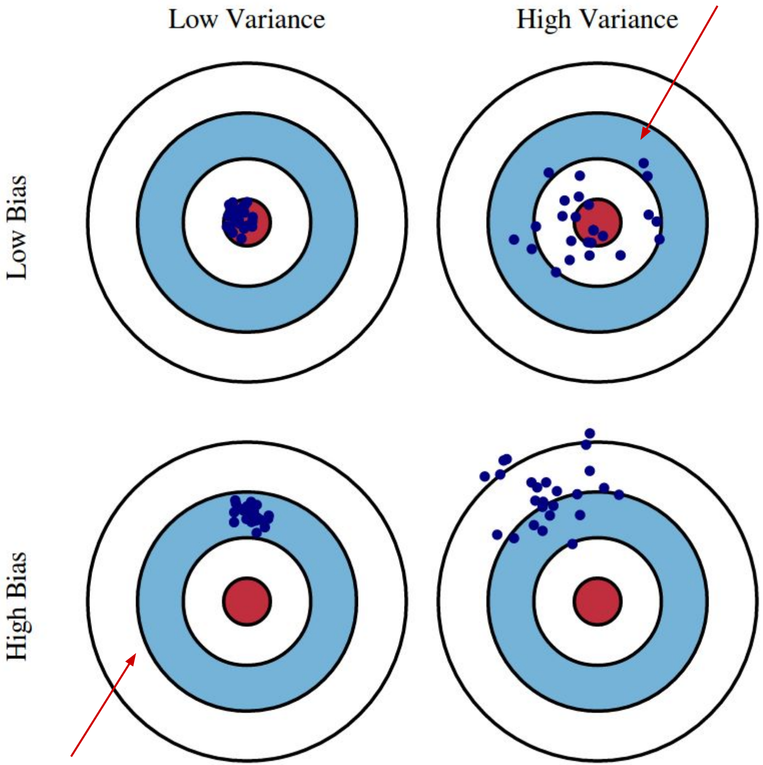


$$Err(x) = \left(E[\hat{f}(x)] - f(x)\right)^2 + E\left[\left(\hat{f}(x) - E[\hat{f}(x)]\right)^2\right] + \sigma_e^2$$

Error(x) = Bias^2 + Variance + Irreducible Error (Noise)

Bias and Variance

Overfitting

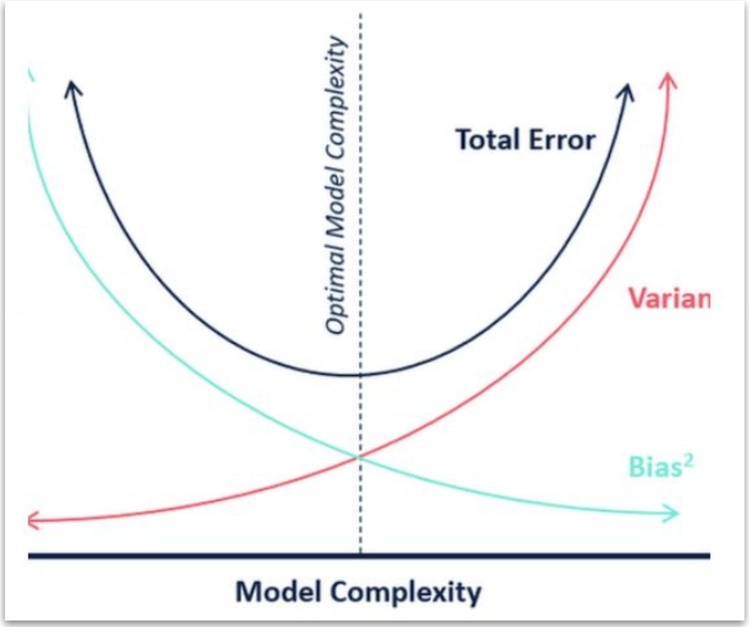


Underfitting

- In supervised learning algorithms, **underfitting** occurs when a model is unable to capture the underlying pattern of the data. Such models usually have **high bias** and **low variance**. It happens when we have very less amount of data to train an accurate model or when we try to fit a linear model with a nonlinear data. Such models are very simple to capture the complex patterns in data like Linear and logistic regression
- **Overfitting** happens when our model captures the noise along with the underlying pattern in data. It happens when we train our model with a noisy dataset. These models have **low bias and high variance.** These models are very complex like Decision trees which are prone to overfitting.

Bias and Variance Tradeoff?

- If our model is too simple and has very few parameters then it may have high bias and low variance.
- On the other hand if our model has large number of parameters then it's going to have high variance and low bias.
- So we need to find the right/good balance without overfitting and underfitting the data. An optimal balance of bias and variance would never overfit or underfit the model



Non-Parametric *vs* Parametric Tests (1)

Non-Parametric	Parametric
Does not make any assumption about the distribution. They are also called ***'Distribution Free Tests'***.	Makes Assumptions about the distribution of the data. Assumes **Normally distributed data**
Used for data that are **other than Normal distribution**	Used in the data that is **Normally distributed**
● Deals with **Medians** ● Used when **Sample size is small** ● Used with **Ordinal** or **Ranked** data ● Presence of **outliers**	● Deals with **Means** and **Variance**. ● Used to **estimate at least one 'Population Parameter'** (Mean or S.D.) from the Sample Statistic (Sample mean or Sample S.D.)

Non-Parametric *vs* Parametric Tests (2)

Non-Parametric	Parametric
Data is **Categorical** (Nominal or Ordinal)	Data is on the **Continuous Scale** of measurement (interval or Ratio)
No knowledge about the population or its Parameters and still it is required to test the Hypothesis of the Population	**Information about the Population is completely known** by means of Population parameters. Thus the statistic test is called Parametric test
H0 is **free from population parameters**	**H0** is **made on the Parameters** of the population distribution

Non-Parametric *vs* Parametric Tests (3)

Non-Parametric	Parametric
Non-parametric tests are **not powerful**. There is a loss of information when we calculate the Ranks. **The impact of the magnitude of the data is lost**	These tests are **Powerful** as the Parameters/ Statistics are calculated from the Original data. These tests **include the impact of the magnitude of the data**
● Chi-Square, ● Mann Whitney U Test, ● Wilcoxon Sign Ranks Test ● Kruskal-Wallis Test	● Z-Test, ● t-Test, ● F-Test, ● ANOVA