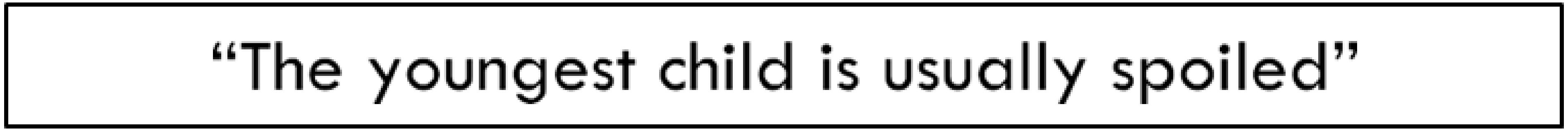
Exploratory Data Analysis

Penyusun Modul: Chairul Aulia

Editor: Rina Fitriyani, Silfa Rahma Aulia

Exploratory Data Analysis



“The youngest child is usually spoiled”

Anecdotal evidence is evidence collected in a casual or informal manner and relying heavily or entirely on personal testimony based on data that is unpublished and usually personal.

Exploratory Data Analysis

By those standards, anecdotal evidence usually fails, because...

1. **SMALL NUMBER OF OBSERVATIONS**

   the difference is probably small compared to natural variation

2. **SELECTION BIAS**

   People who join a discussion of this question might be interested because they relate to the topics

3. **CONFIRMATION BIAS**

   People who believe the claim might be more likely to contribute examples that confirm it.
   People who doubt the claim are more likely to cite counterexamples

4. **IN-ACCURACY**

   Anecdotes are often personal stories, and often misremembered, misrepresented, repeated inaccurately, etc

Exploratory Data Analysis

To address the limitations of anecdotes, we will use the tools of statistics, which include,

DATA COLLECTION

Use data from a large national survey that was designed explicitly with the goal of generating statistically valid inferences about the population

DESCRIPTIVE STATISTICS

Generate statistics that summarize the data concisely, and evaluate different ways to visualize data

HYPOTHESIS TESTING

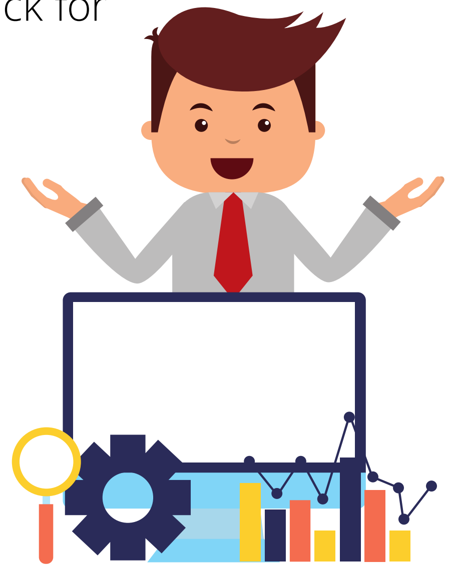
See apparent effects, like a difference between two groups, we will evaluate whether the effect might have happened by chance.

EXPLORATORY DATA ANALYSIS

Look for patterns, differences, and other features that address the questions we are interested in. At the same time we will check for inconsistencies and identify limitations

ESTIMATIONS

Use data from a sample to estimate characteristics of the general population.



Exploratory Data Analysis

Exploratory Data Analysis does two main things:

1. It helps clean up a dataset

For data scientists, cleaning and prepping the dataset before using them to build a machine learning model is one of the first step of building a good model.
In fact, you will mostly spend your time on doing data cleansing and data analysis than building the model

2. It gives you a better understanding of the variables and the relationships between them.

Doing EDA will give you many insights on the project by learning the very useful **Feature Importance**

Exploratory Data Analysis

Why is Feature Importance so Useful?

Feature Importance is extremely useful for the following reasons:

Data Understanding

Building a model is one thing, but understanding the data that goes into the model is another. Like a correlation matrix, feature importance allows you to understand the relationship between the features and the target variable. It also helps you understand what features are irrelevant for the model.

Model Improvement

When training your model, you can use the scores calculated from feature importance to reduce the dimensionality of the model. The higher scores are usually kept and the lower scores are deleted as they are not important for the model. This not only makes the model simpler but also speeds up the model's working, ultimately improving the performance of the model.

Model Interpretability.

Feature Importance is also useful for interpreting and communicating your model to other stakeholders. By calculating scores for each feature, you can determine which features attribute the most to the predictive power of your model.

Wait!

Before we delve further into the unknown field of Exploratory Data Analysis, we need to properly learn about the fundamental of it

The Statistics!

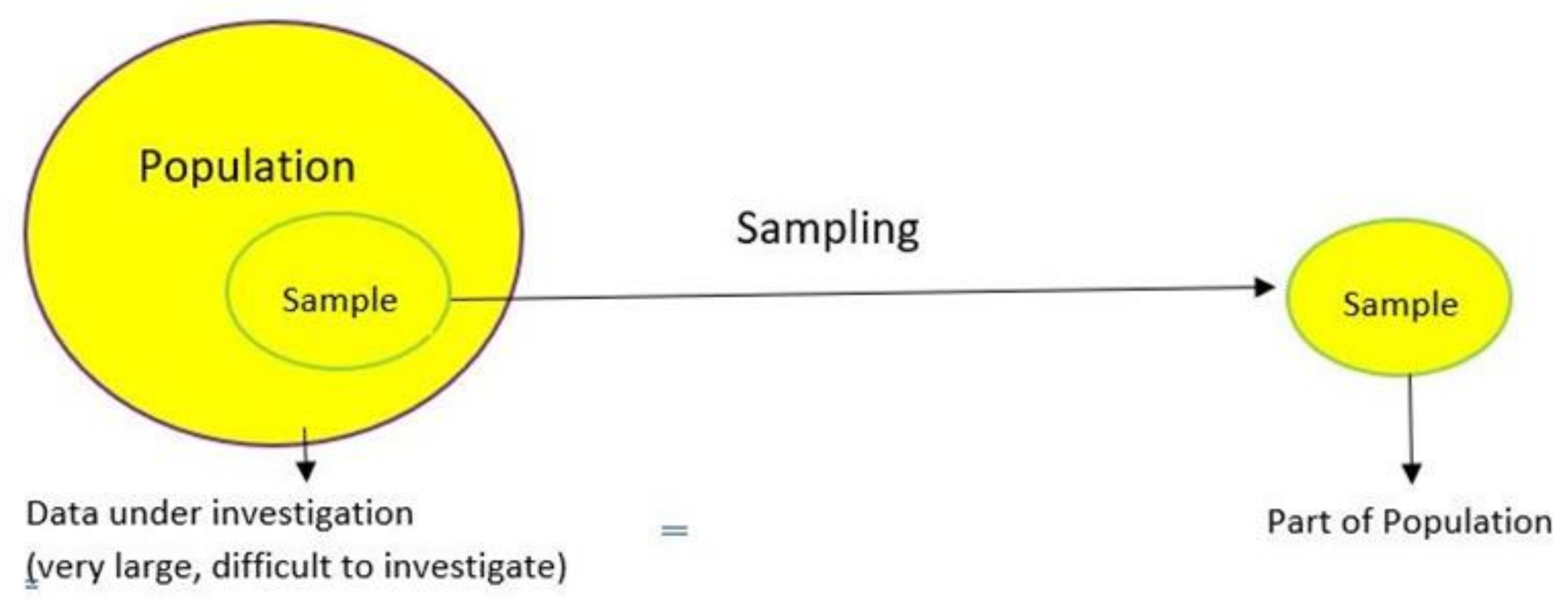
Statistics

What is Statistics?

Statistics is the study and manipulation of data, including ways to gather, review, analyze, and draw conclusions from data. The two major areas of statistics are descriptive and inferential statistics. Statistics can be used to make better-informed business and investing decisions.

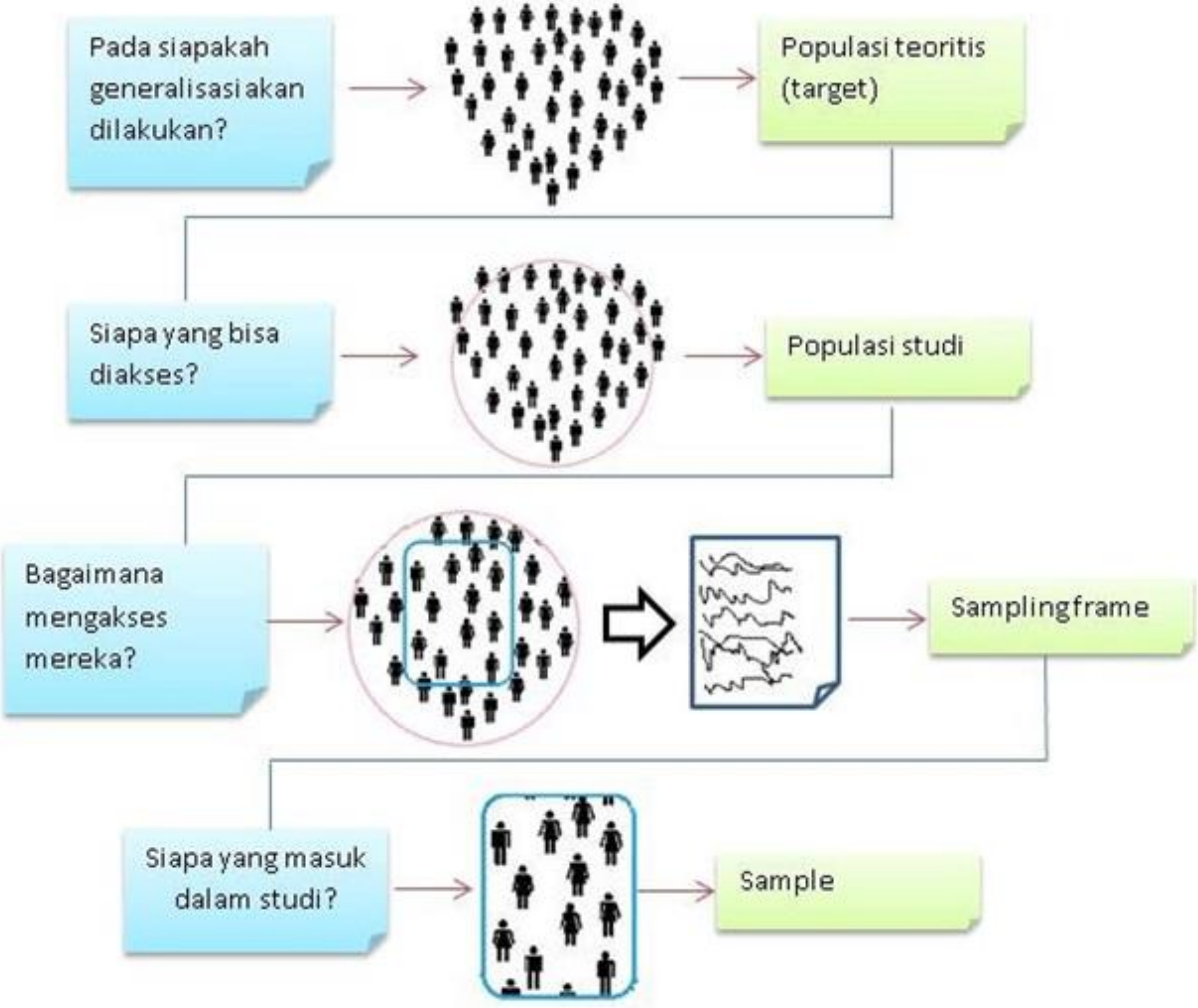
Sampling

Margono (2004), Population is the whole data which is the center of attention of a researcher in the specified scope and time. While, sample is part of the population selected and whose characteristics will be researched (Djarwanto, 1994: 43).



Source :
https://www.analyticsvidhya.com

Sampling



Pada siapakah generalisasi akan dilakukan?

Populasi teoritis (target)

Siapa yang bisa diakses?

Populasi studi

Bagaimana mengakses mereka?

Samplingframe

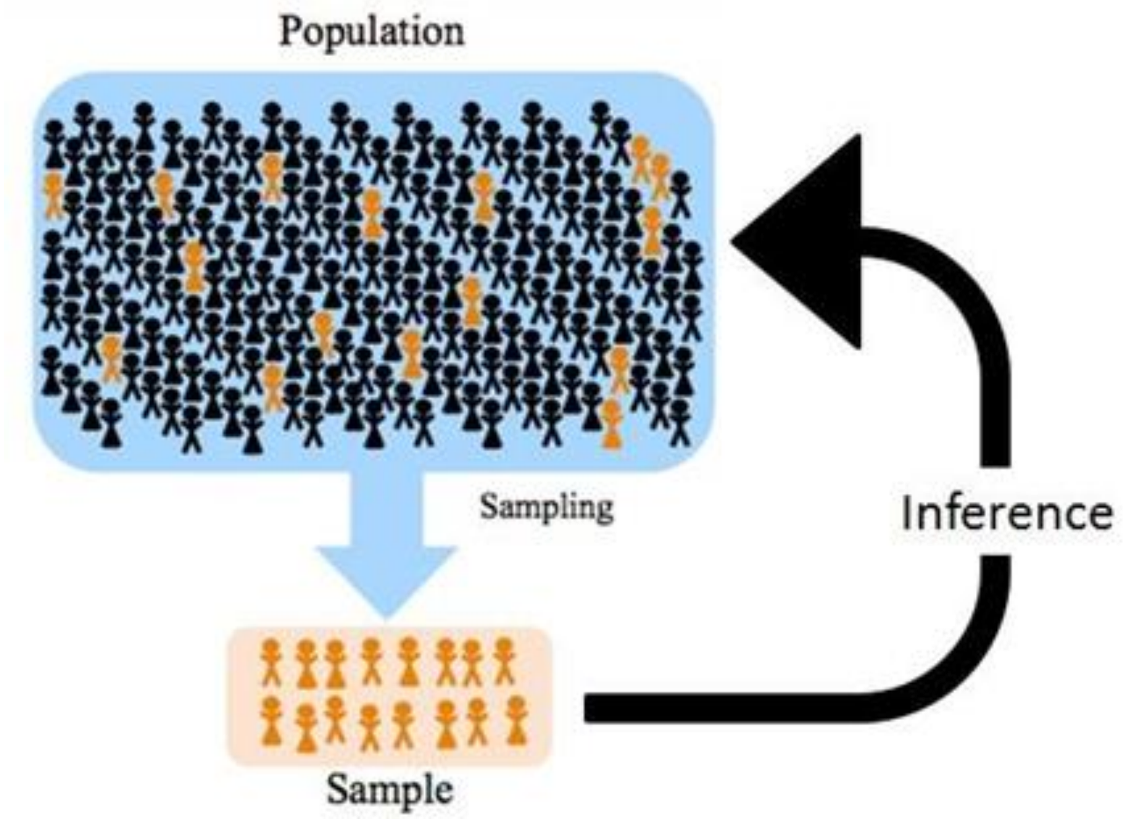
Siapa yang masuk dalam studi?

Sample

Source : https://www.tau-data.id

Sampling

Why do we need Sampling?

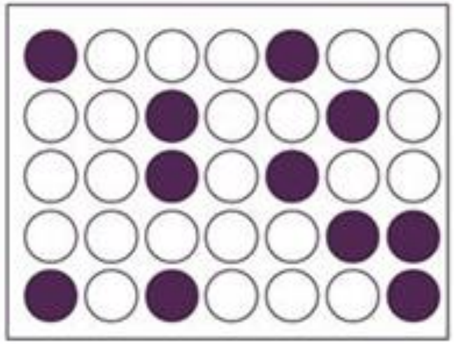


Source:
https://towardsdatascience.com/

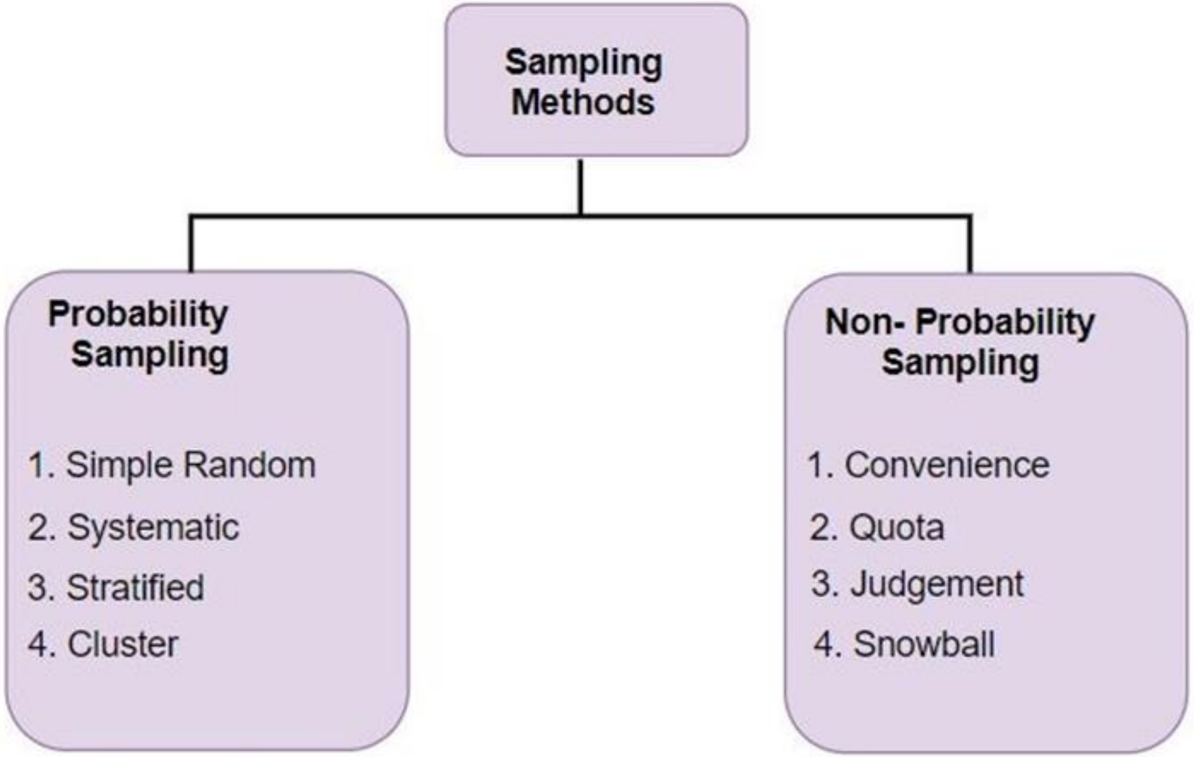
Population and Sample like organisms and organs.

Sampling Methods

Randomization



Source: https://towardsdatascience.com/



Sampling Methods

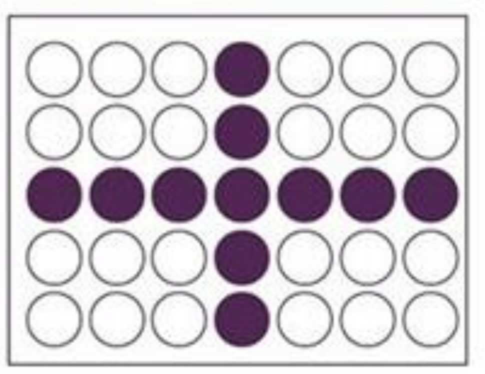
Probability Sampling

1. Simple Random
2. Systematic
3. Stratified
4. Cluster

Non- Probability Sampling

1. Convenience
2. Quota
3. Judgement
4. Snowball

Non Randomization

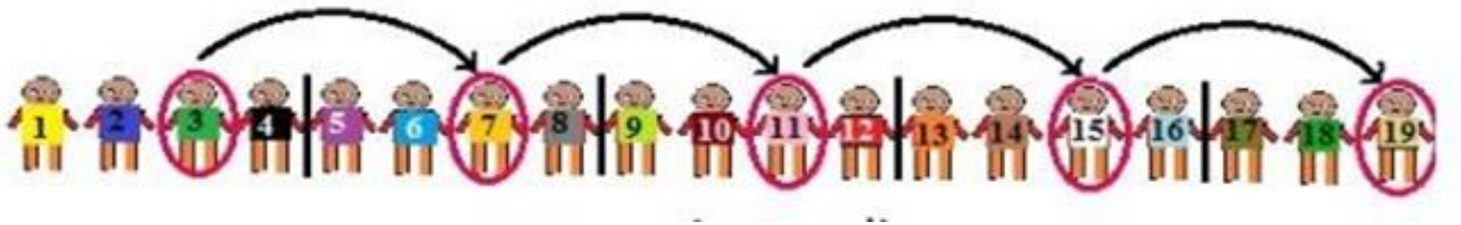


Source: https://towardsdatascience.com/

Source : https://www.analyticsvidhya.com

Sampling Methods

Source: https://towardsdatascience.com/

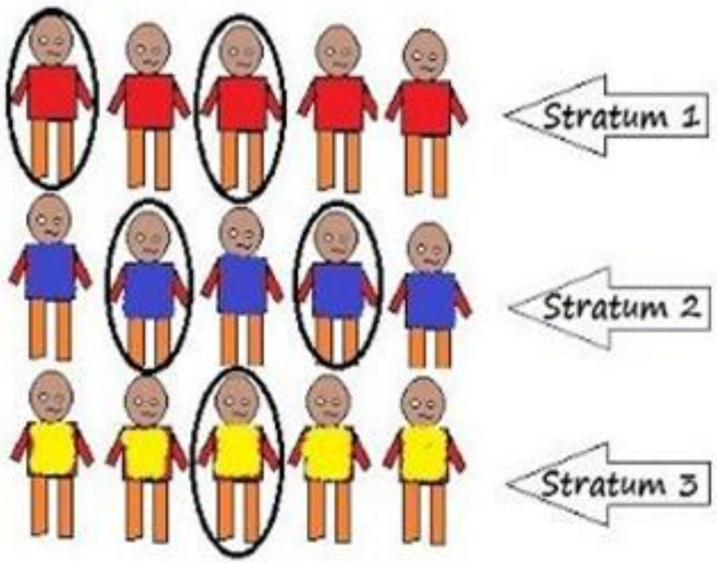


Source: https://towardsdatascience.com/

Systematic Clustering: Selection of elements is systematic and not random except the first element.

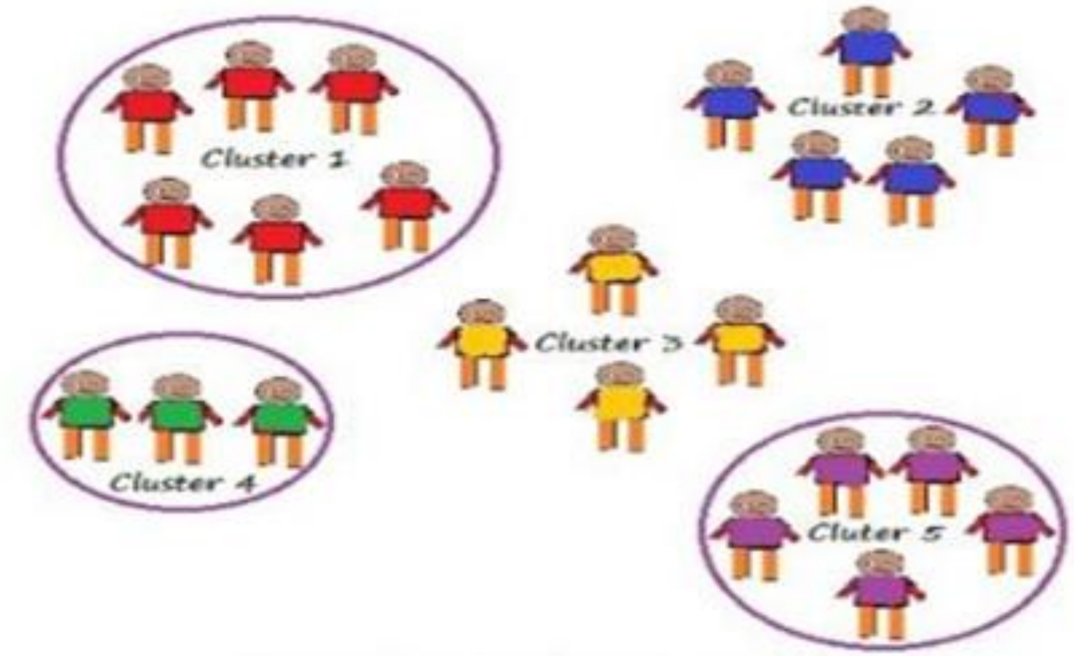
Simple Random Sampling: Everyelement has an equal chance of gettingselected to be the part sample.

Sampling Methods



Source: https://towardsdatascience.com/

Stratified Sampling: This technique divides the elements of the population into small subgroups (strata) based on the similarity in such a way that the elements within the group are homogeneous andheterogeneous among the other subgroups formed. And then the elements are randomly selected from each of these strata.

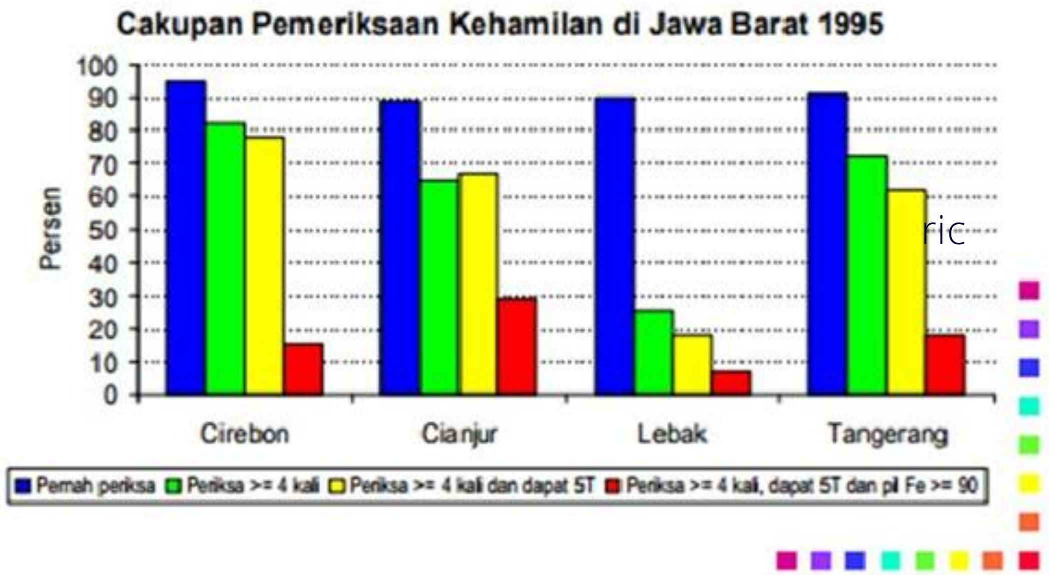


Source: https://towardsdatascience.com/

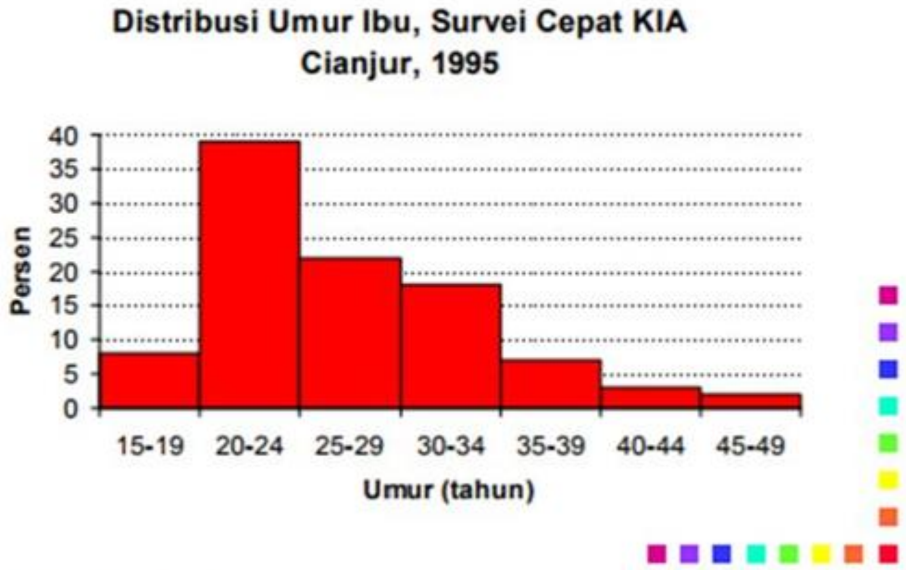
Cluster Sampling: The entire population is divided into clusters or sections and then the clusters are randomly selected. All the elements of the cluster are used for sampling.

Histogram

Histogram is a graphical display of data using bars of different heights.



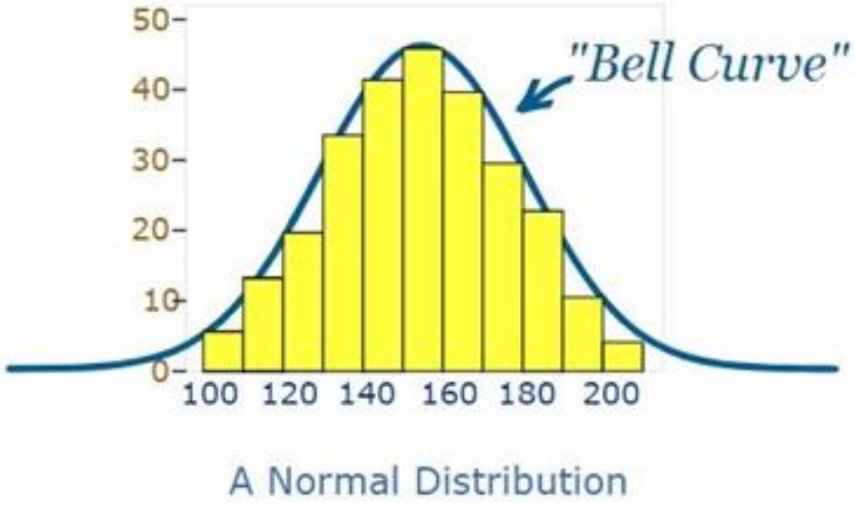
© Jurusan Biostatistik FKM-UI, 2005

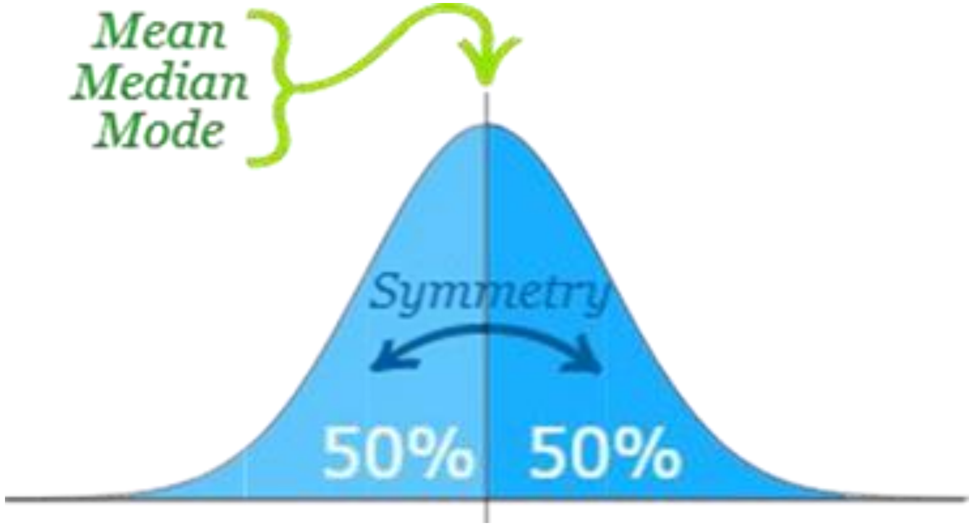


© Jurusan Biostatistik FKM-UI, 2005

Normal Distribution

The normal distribution is a symmetric distribution where most of the observations cluster around the central peak and the probabilities for values further away from the mean taper off equally in both directions.





Source : https://www.mathsisfun.com/

The Normal Distribution has:

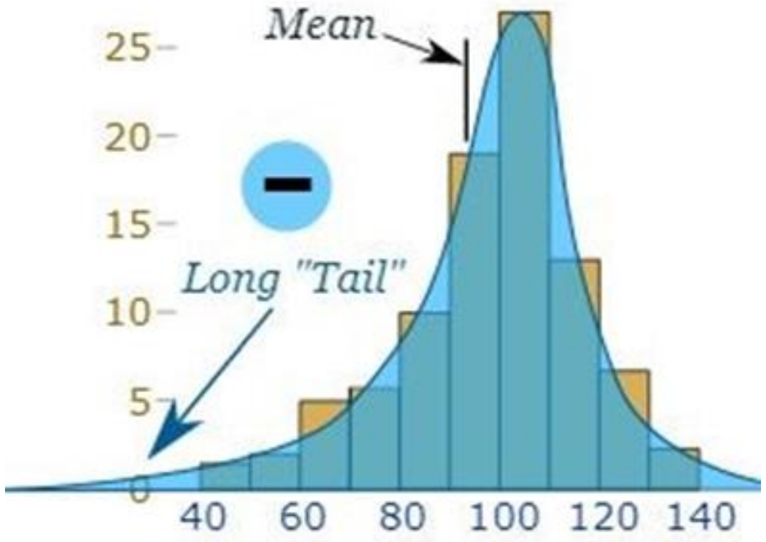
Mean = Median = Mode

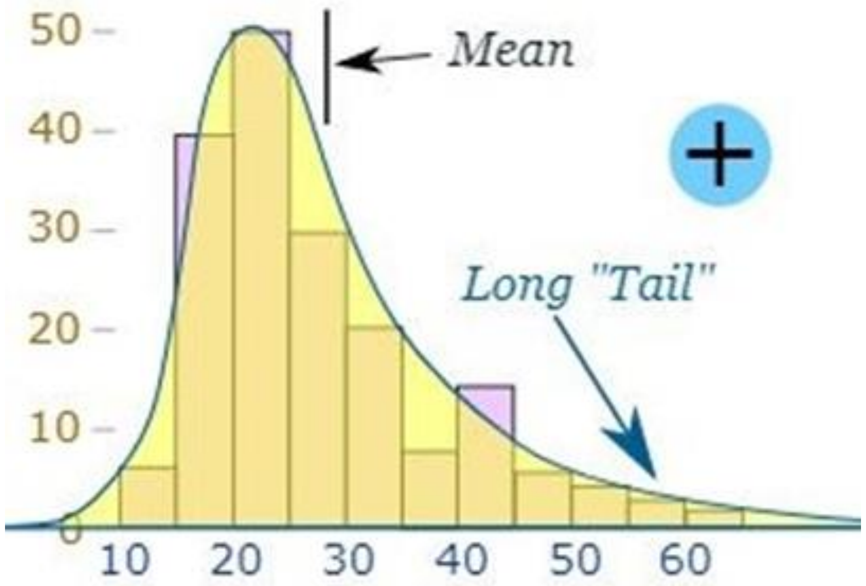
Symmetry about the center

50% of values less than the mean and 50% greater than the mean

Skewed Distribution

Skewed distribution is neither symmetric nor normal because the data values trail off more sharply on oneside than on the other.





Source : https://www.mathsisfun.com/

Negative Skewed

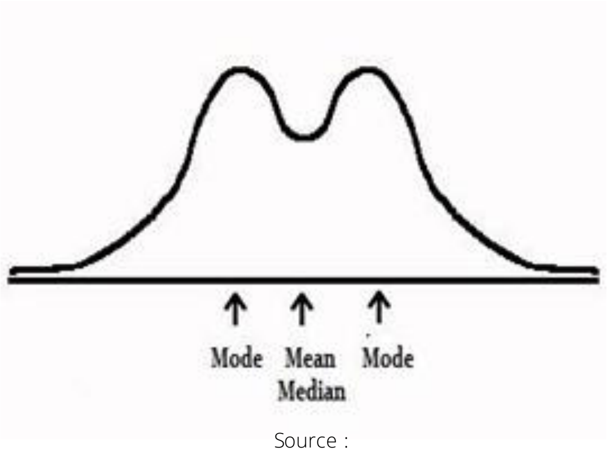
Mean < Median < Mode

Positive Skewed

Mean > Median > Mode

Distribution

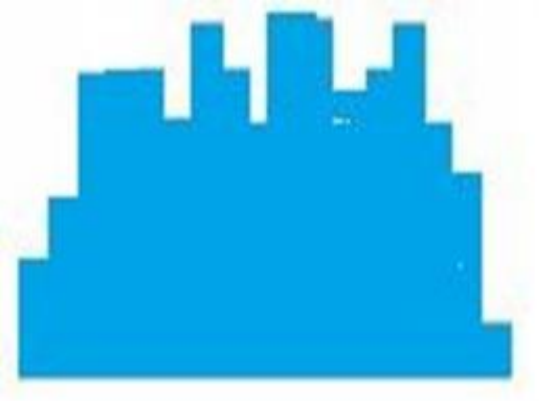
Binomial Distribution



Source : https://www.statisticshowto.com/

Bimodal distribution is a probability distribution with two different modes. Bimodal distribution can occur because the results of twoprocesses with different distributionsare combined in one data set.

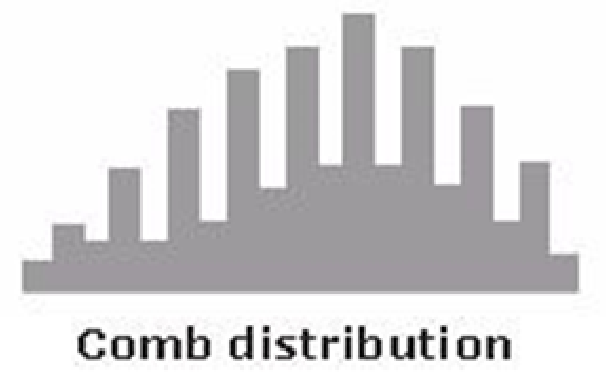
Multimodal Distribution



Source : https://www.statisticshowto.com/

Multimodal distribution is adistribution of probabilities which has two or more modes. In other words, multimodaldistribution is a number of processes with a normal distribution combined. Multimodal distribution is known as Plateau Distribution, when there are more than a few peaks that are close together.

Comb Distribution

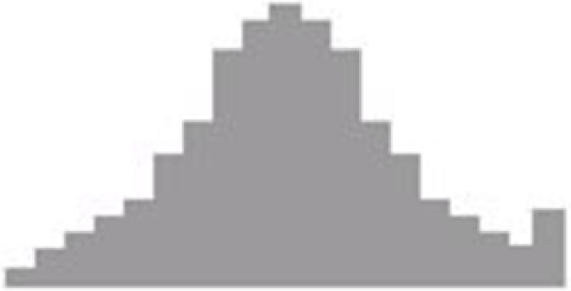


Source : https://www.statisticshowto.com/

This distribution often results from rounded-off data and/or an incorrectly constructed histogram.

Distribution

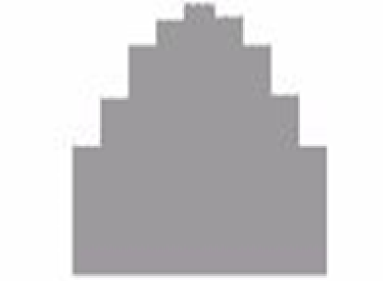
Edge Peak Distribution



Source: https://asq.org/

The edge peak distribution looks like the normal distribution except that it has a large peak at one tail. Usually this is caused by faulty construction of the histogram, with data lumped together into a group labeled “greater than.”

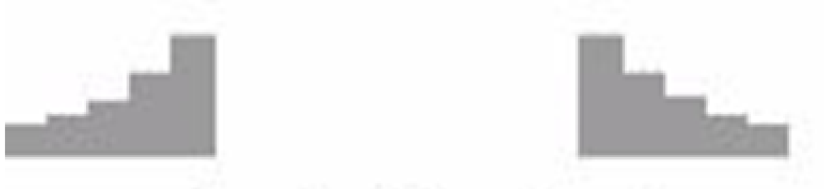
Truncated or heart-cut Distribution



Source: https://asq.org/

The truncated or heart-cut distribution looks like a normal distribution with the tails cut off.

Dog Food Distribution



Source: https://asq.org/

The dog food distribution is missing something—results in the average

Descriptive Statistics

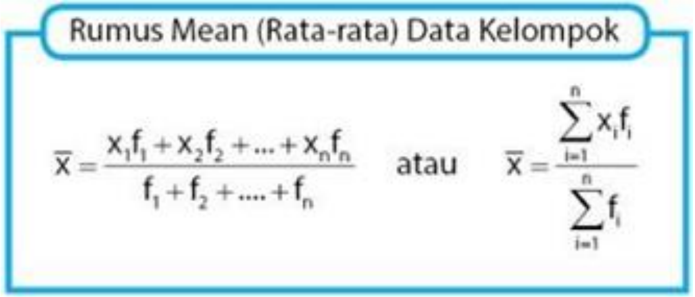
There are two Commonly Used Measures :

1. Measures of Central Tendency

Typically describes the center of the data. These one number summary is of three types.

Mean

Mean is a number around which the entire data set is spread.



Rumus Mean (Rata-rata) Data Kelompok

$$\bar{x} = \frac{x_1f_1 + x_2f_2 + \dots + x_nf_n}{f_1 + f_2 + \dots + f_n} \quad \text{atau} \quad \bar{x} = \frac{\sum_{i=1}^{n} x_if_i}{\sum_{i=1}^{n} f_i}$$

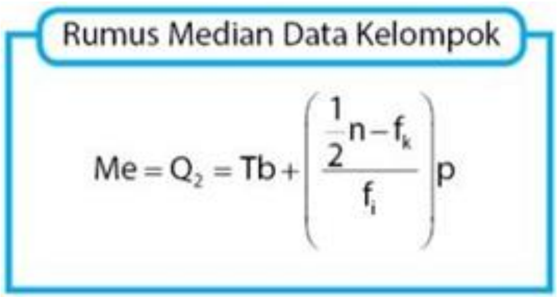
$\bar{x}$ = average calculated from group data

di = i-class frequency

xi = middle value of i-class

Median

Median is the point which divides the entire data into two equal halves.



Rumus Median Data Kelompok

$$Me = Q_2 = Tb + \left(\frac{\frac{1}{2}n - f_k}{f_i}\right)p$$

Tb = the lower edge of the median class

n = total frequency

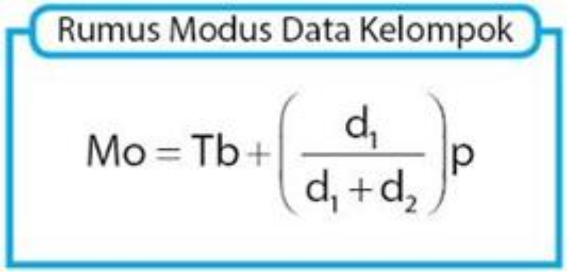
fk = number of frequencies before the medianclass

fi = median class frequency

p = length of the interval class

Mode

Mode is the number which has the maximum frequency in the entire data set.



Rumus Modus Data Kelompok

$$Mo = Tb + \left(\frac{d_1}{d_1 + d_2}\right)p$$

Tb = the bottom edge of the mode class

d1 = difference in mode class frequency with frequency before mode class

d2 = difference in mode class frequency with frequency after mode class

p = length of class interval

Descriptive Statistics

2. Measures of Dispersion (or Variability)

Measures of Dispersion describes the spread of the data around the central value (or the Measures of Central Tendency).

1 Absolute Deviation from Mean — also called Mean Absolute Deviation (MAD), describe the variation in the data set, in sense that it tells the average absolute distance of each data point in the set. It is calculated as

$$\text{Mean Absolute Deviation} = \frac{1}{N}\sum_{i=1}^{N}\left|X_i - \overline{X}\right|$$

2 Variance — Variance measures how far are data points spread out from the mean. It is calculated as

$$\text{Variance} = \frac{1}{N}\sum_{i=1}^{N}\left(X_i - \overline{X}\right)^2$$

3 Standard Deviation — The square root of Variance is called the Standard Deviation. It is calculated as

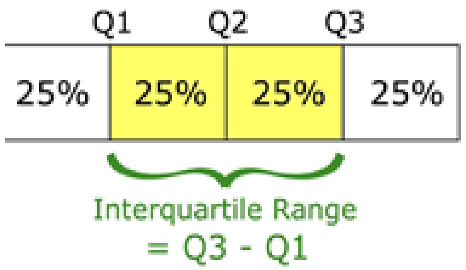
$$\text{Std Deviation} = \sqrt{\text{Variance}} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left(X_i - \overline{X}\right)^2}$$

Descriptive Statistics

4 Range — Range is the difference between the Maximum value and the Minimum value in the data set. It is given as

$$\text{Range} = \text{Maximum} - \text{minimum}$$

5 Quartiles — Quartiles are the points in the data set that divides the data set into four equal parts

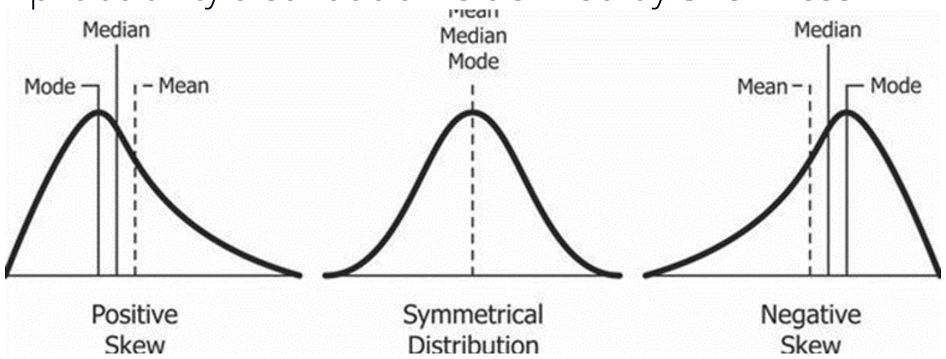


Q1 Q2 Q3

25%	25%	25%	25%

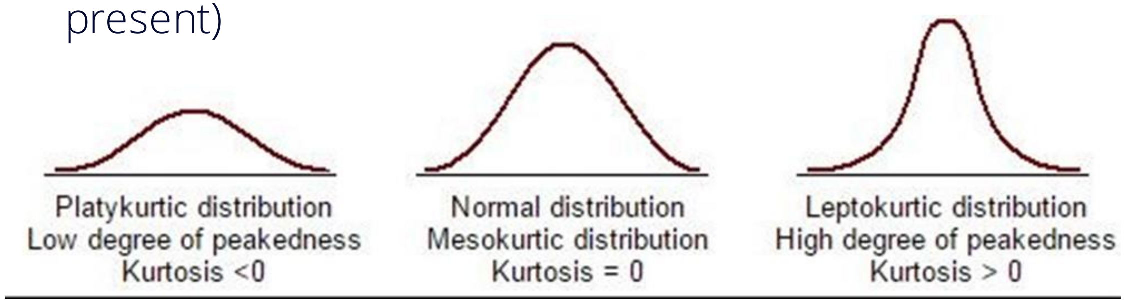
Interquartile Range
= Q3 - Q1

6 Skewness — The measure of asymmetry in a probability distribution is defined by Skewness



Source :

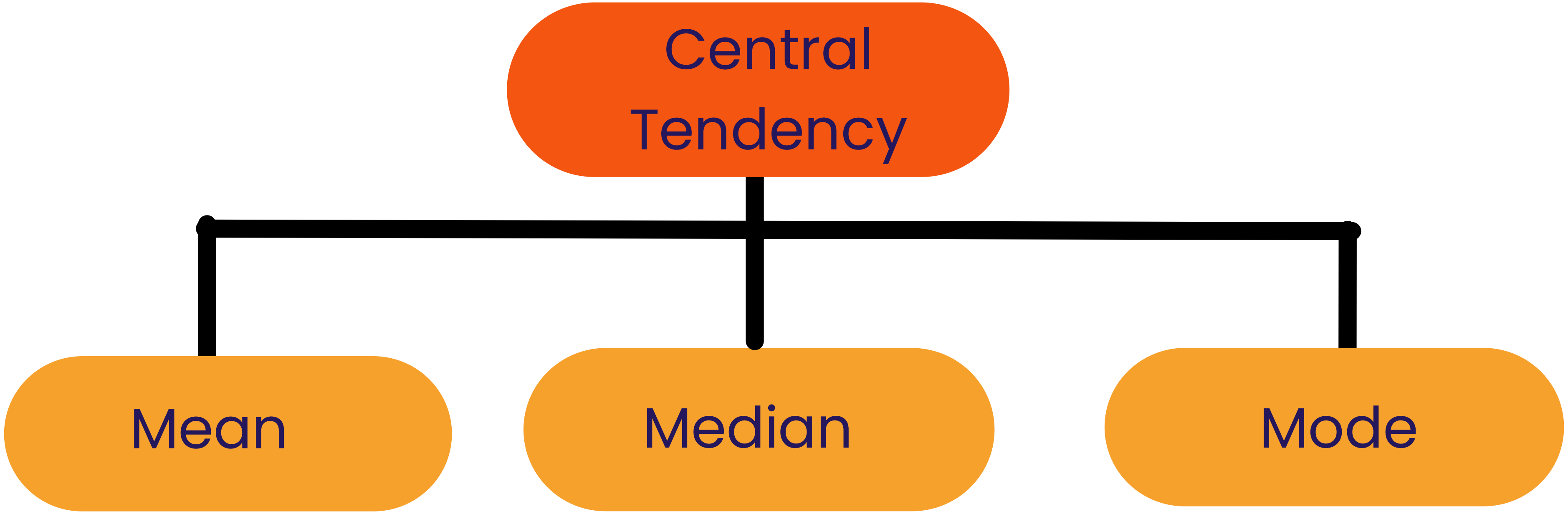
7 Kurtosis — Kurtosis describes the whether the data is light tailed (lack of outliers) or heavy tailed (outliers present)



Source :
https://towardsdatascience.com/

Central Tendency

A measure of central tendency is a single value that attempts to describe a set of data by identifying the central position within that set of data. As such, measures of central tendency are sometimes called measures of central location.



Mean

Mean is equal to the sum of all the values in the data set divided by the number of values in the data set.

Consider the wages of staff at a factory below

Staff	1	2	3	4	5	6	7	8	9	10
Salary	15k	18k	16k	14k	15k	15k	12k	17k	90k	95k

Median

Median is the middle score for a set of data that has been arranged in order of magnitude. The median is less affected by outliers and skewed data. In order to calculate the median, suppose we have the data below:

Consider the wages of staff at a factory below

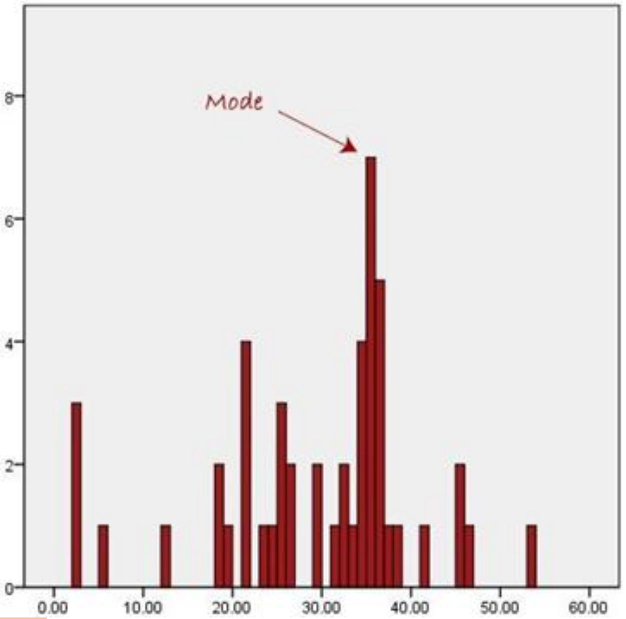
65	55	89	56	35	14	56	55	87	45	92

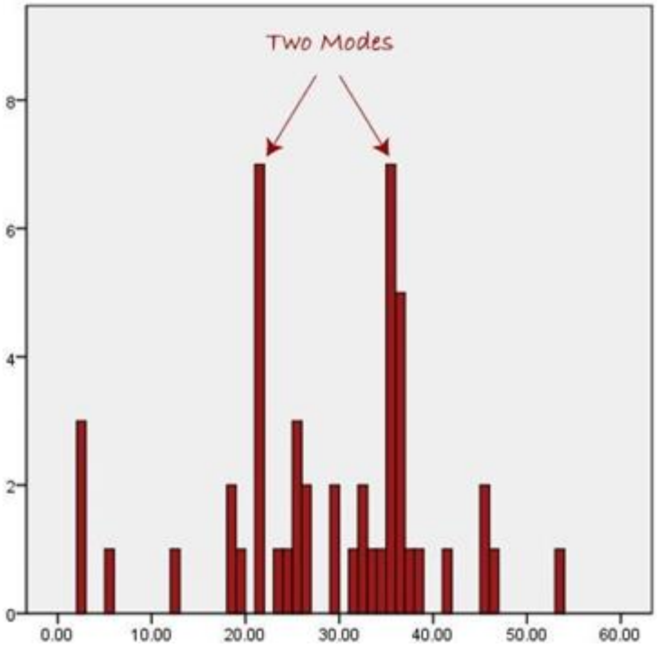
So, if we look at the example below:

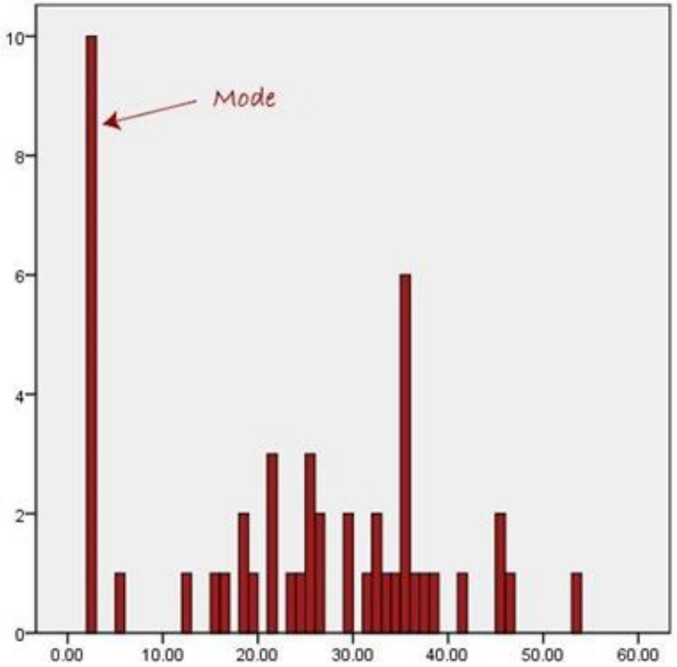
65	55	89	56	35	14	56	55	87	45

Mode

Mode is the most frequent score in our data set. On a histogram it represents the highest bar in a bar chart or histogram.







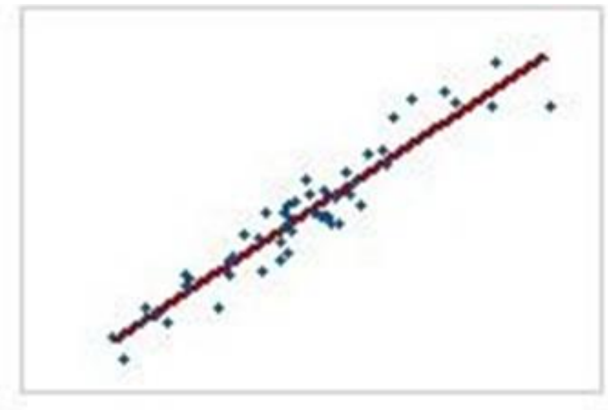
Source :
https://statistics.laerd.com/

When to use the mean, median and mode

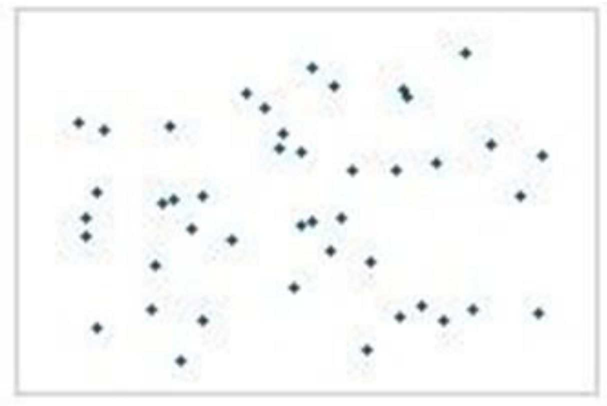
Type of Variable	Best measure of central tendency
Nominal	Mode
Ordinal	Median
Interval/Ratio (not skewed)	Mean
Interval/Ratio (skewed)	Median

Correlation

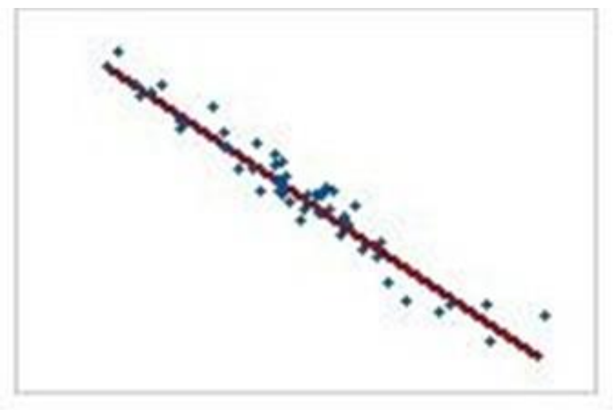
Correlation is a statistical measure that describes how two variables are related and indicates that as one variable changes in value, the other variable tends to change in a specific direction.



Positive correlation



no correlation



Negative correlation

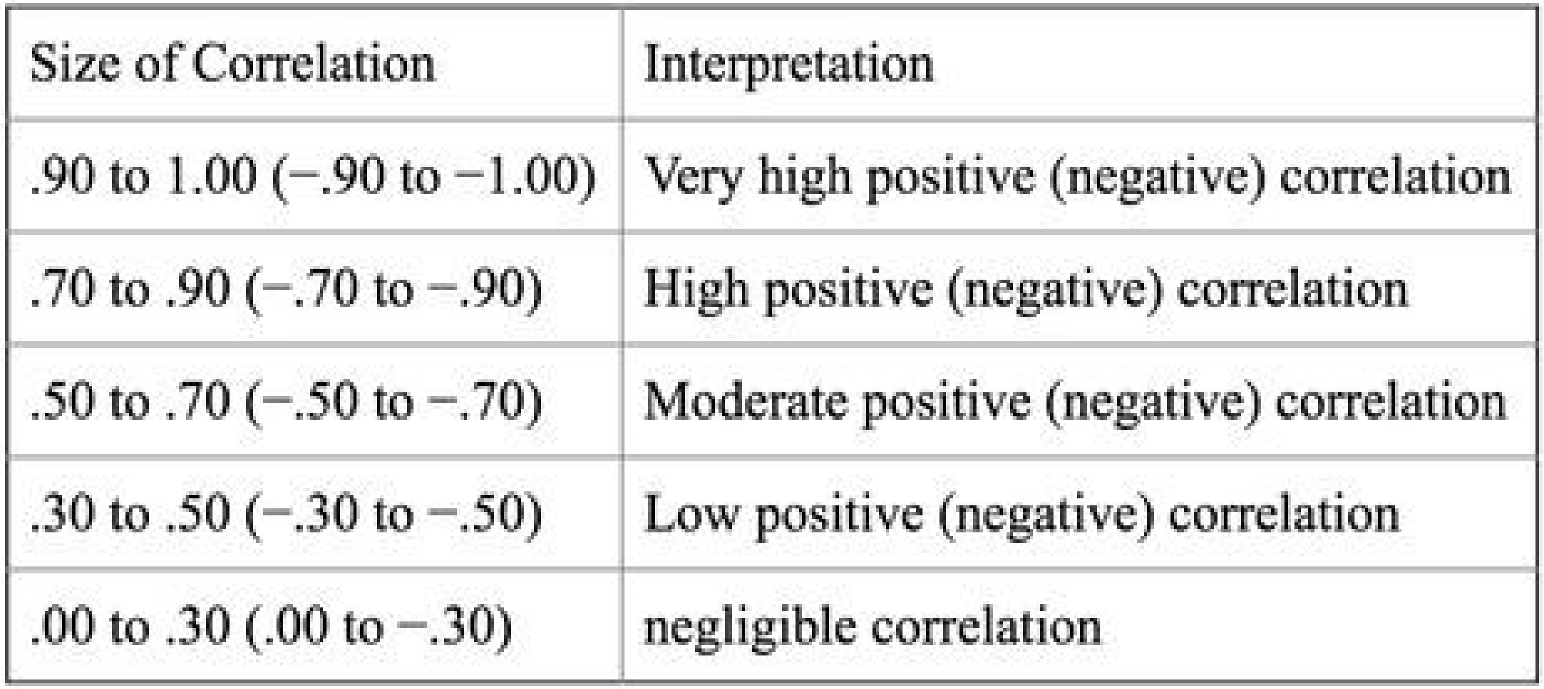
Source :
https://towardsdatascience.com/

Correlation value:

CHARACTERISTIC	INTERPRETATION
Positive correlation value	Positive direction / association; as one variable increases, so does the other
Negative correlation value	Negative direction / association; as one variable increases, the other decreases
|r| is closer to 1	The closer |r| is to one, the stronger the association
|r| closer to 0	The closer |r| is to zero, the weaker the association

Size of Correlation

The table below demonstrates how to interpret the size (strength) of a correlation coefficient

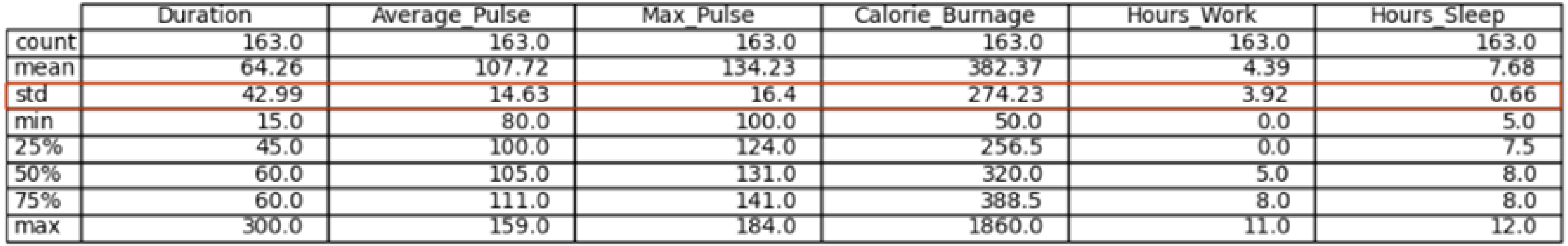


Size of Correlation	Interpretation
.90 to 1.00 (−.90 to −1.00)	Very high positive (negative) correlation
.70 to .90 (−.70 to −.90)	High positive (negative) correlation
.50 to .70 (−.50 to −.70)	Moderate positive (negative) correlation
.30 to .50 (−.30 to −.50)	Low positive (negative) correlation
.00 to .30 (.00 to −.30)	negligible correlation

Standard Deviation

Standard deviation is a number that describes how spread out the observations are.

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



	Duration	Average_Pulse	Max_Pulse	Calorie_Burnage	Hours_Work	Hours_Sleep
count	163.0	163.0	163.0	163.0	163.0	163.0
mean	64.26	107.72	134.23	382.37	4.39	7.68
std	42.99	14.63	16.4	274.23	3.92	0.66
min	15.0	80.0	100.0	50.0	0.0	5.0
25%	45.0	100.0	124.0	256.5	0.0	7.5
50%	60.0	105.0	131.0	320.0	5.0	8.0
75%	60.0	111.0	141.0	388.5	8.0	8.0
max	300.0	159.0	184.0	1860.0	11.0	12.0

A mathematical function will have difficulties in predicting precise values, if the observations are "spread". Standard deviation is a measure of uncertainty.

A low standard deviation means that most of the numbers are close to the mean (average) value.

A high standard deviation means that the values are spread out over a wider range.

Variance

Variance measures how far a set of data is spread out.

$$\text{Population Variance} = (\sigma x)^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

$$\text{Sample Variance} = (Sx)^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

$\sum$ = "the sum of ..."

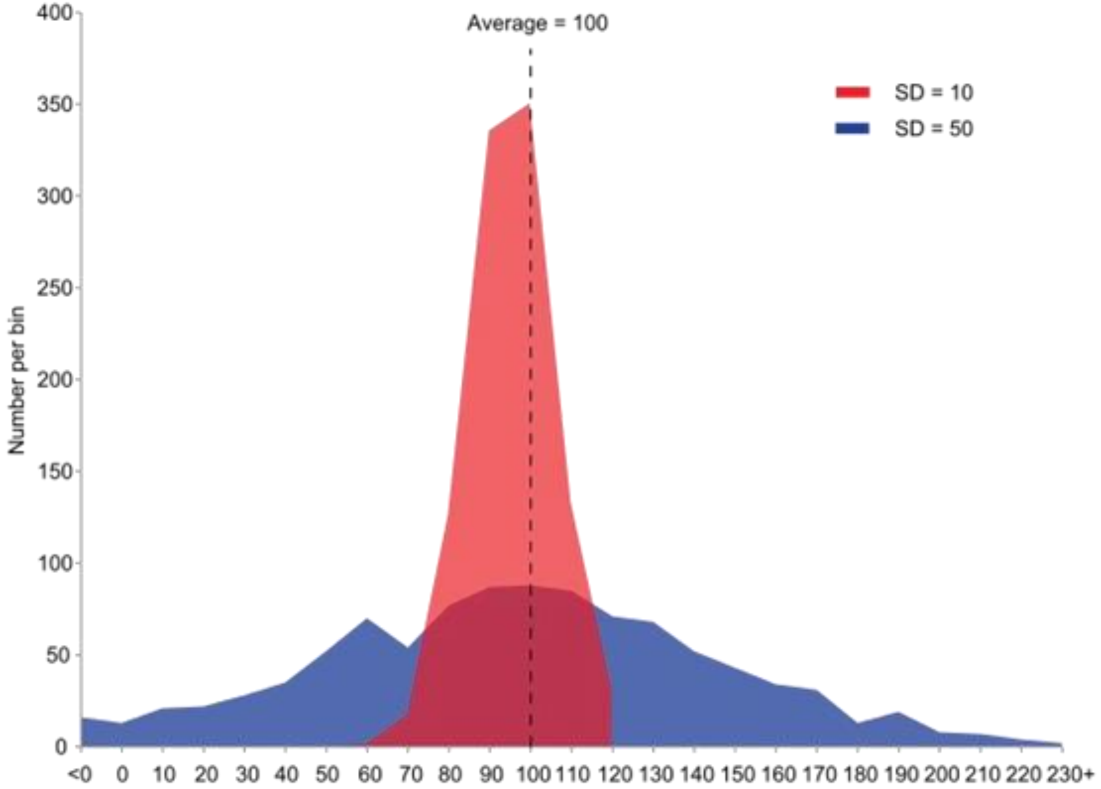
n = number of pieces of data (population)

$n-1$ = number of pieces of data (sample)

$\bar{x}$ = mean (average) of data

x_i = each of the values in the data

$x_1, x_2, x_3, x_4, \ldots x_n$ (as i goes from 1 to n)



Covariance

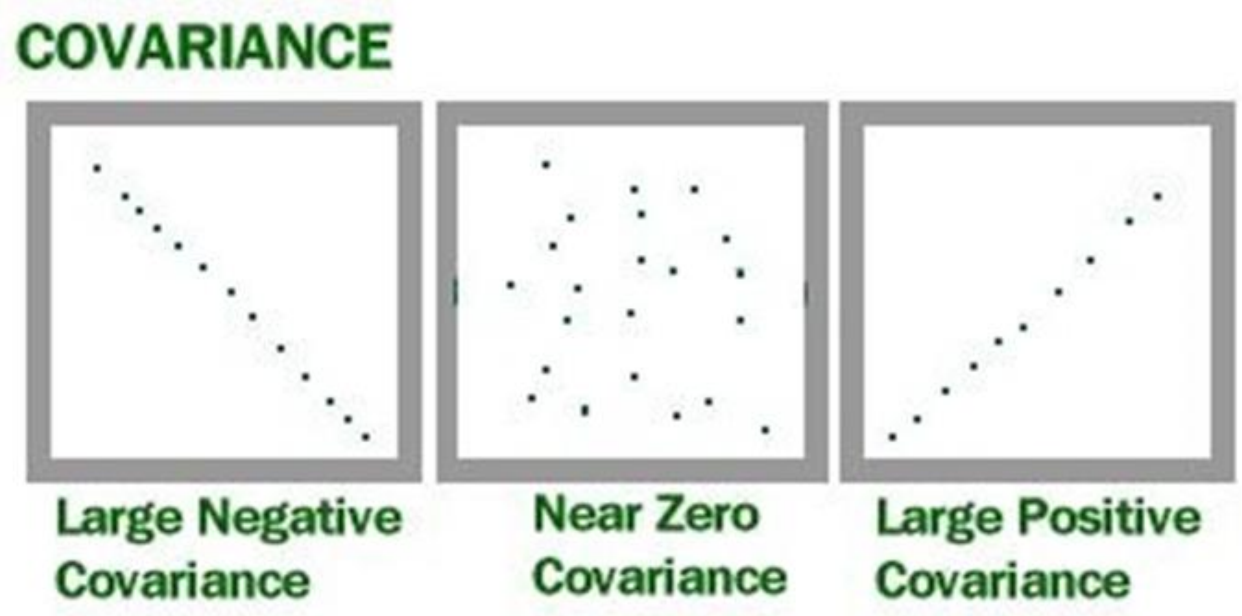
Covariance is a measure of how much two random variables vary together

Population $$\text{Cov}(X,Y) = \frac{\sum (X_i - \overline{X})(Y_j - \overline{Y})}{n}$$

Sample $$\text{Cov}(X,Y) = \frac{\sum (X_i - \overline{X})(Y_j - \overline{Y})}{n - 1}$$

Where:

- X_i = the values of the X-variable
- Y_j = the values of the Y-variable
- $\overline{X}$= the mean (average) of the X-variable
- Ȳ = the mean (average) of the Y-variable
- n – the number of data points



Source : https://www.statisticshowto.com/