INTRODUCTION TO STATISTICS

Jaya Lakshman

STATISTICS

Statistics is a Mathematical Science pertaining to **data** collection, analysis, interpretation and presentation.

it should be clear that statistics is much more than just the tabulation of numbers and the graphical presentation of these tabulated numbers. Statistics is the science of gaining information from numerical and categorical data. Statistical methods can be used to find answers to the questions like:

• What kind and how much data needs to be collected?

• How should we organize and summarize the data?

• How can we analyse the data and draw conclusions from it?

• How can we assess the strength of the conclusions and evaluate their uncertainty?

Statistics for data Science

Statistics is used to process complex problems in the real world so that Data Scientists and Analysts can look for trends and changes in Data to derive meaningful insights from data by performing mathematical computations on it.

Several Statistical functions, principles, and algorithms are implemented to analyse raw data, build a Statistical Model and infer or predict the result.

Statistics Applications – Math And Statistics For Data Science

Terminologies in Statistics

DATA

Data is a collection of facts that has been recorded about a number of people, things or events. The pieces of information are called variables and each of people, things or events are called cases, experimental units or observations.

- A **case** or **observation** is an object about which we collect data.
- A **variable** is a characteristic or property of an individual case.
- In its simplest form, data is a table in which:

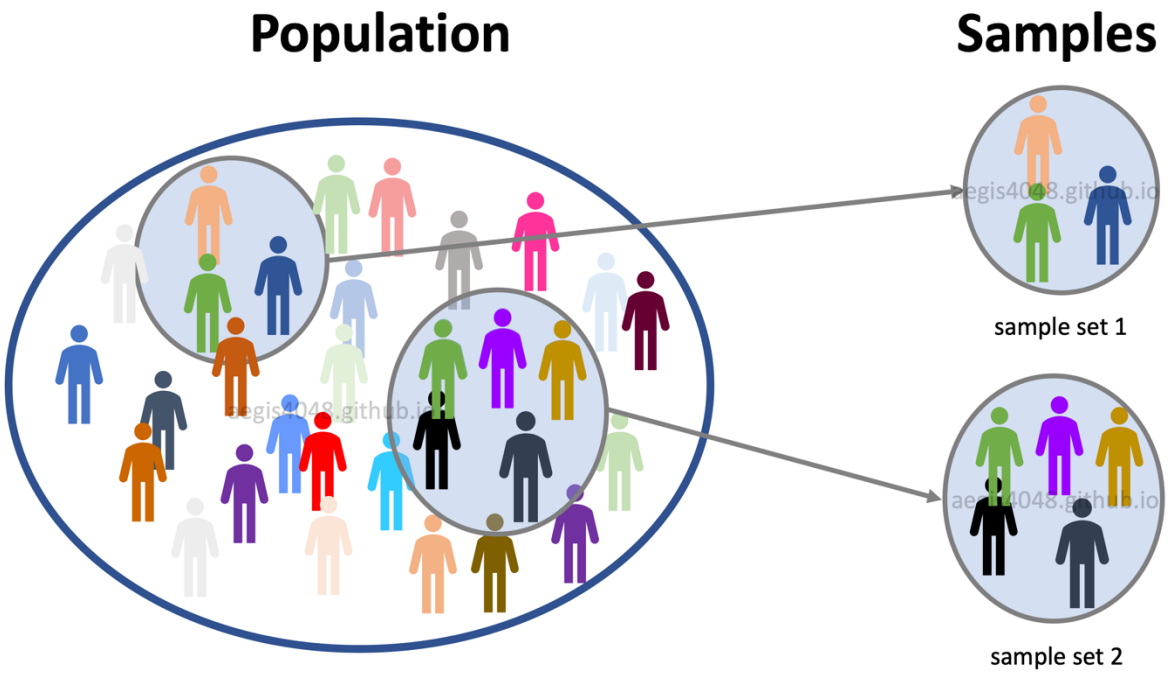
  Columns represent variables.

  Rows represent cases (observations).

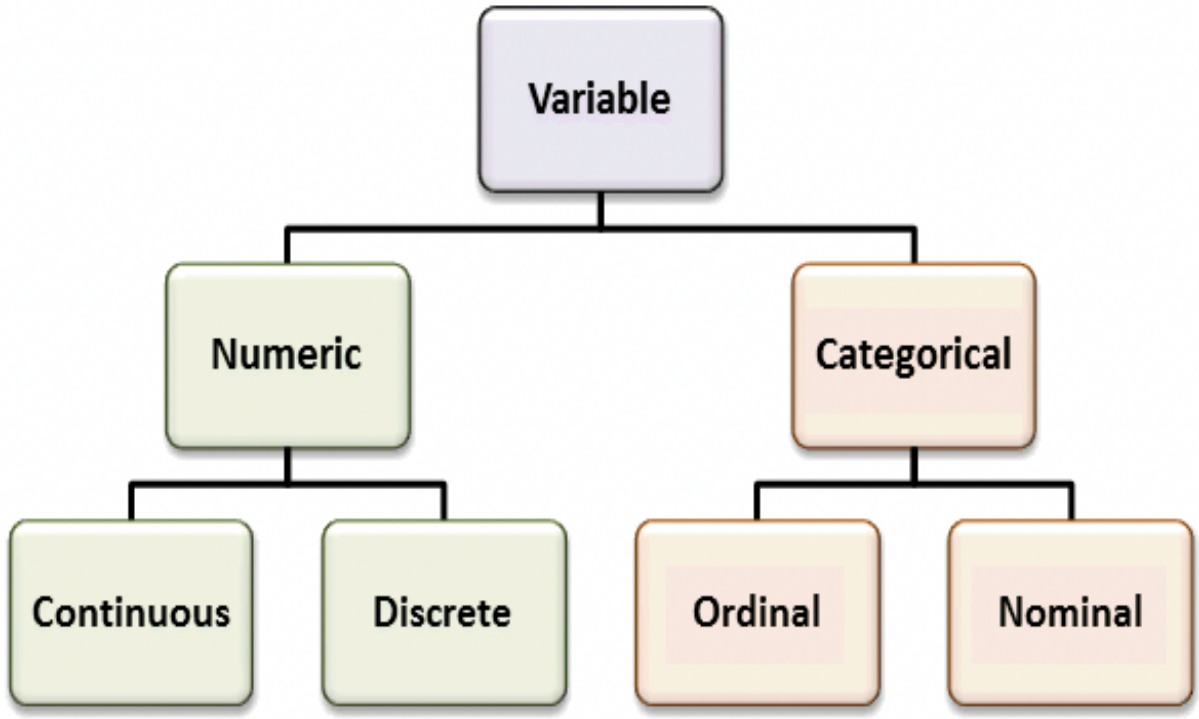
Population and sample

Population and sample are two basic concepts of statistics. Population can be characterized as the set of individual persons or objects in which an investigator is primarily interested during his or her research problem.

- **Population** is the collection of all individuals or items under consideration in a statistical study.
- A **Sample** is a subset of the Population
- A **Variable** is any characteristics, number, or quantity that can be measured or counted. A variable may also be called a data item.
- Any unknown quantity estimated from a population is known as a **parameter** and any quantity computed from a sample is called a **statistic.**



Types of variables



Quantitative (Numeric) – variable that takes numerical values representing different magnitudes. Examples: Age, Time, Price, Number of Days.

Discrete

Value that is countable.

If the possible values form a set of integers e.g. 0, 1, 2, 3, ... 9, 10 (e.g. number of siblings, loans in a bank, payments per day).

Continuous

Value that is measurable.

If a variable takes an infinite number of possible values. In other words, any value is possible for a variable. Example: time, height.

Qualitative (Categorical) – variable that belongs to a set of categories which describes the data. Examples: Gender, Race, Level of Education.

Nominal

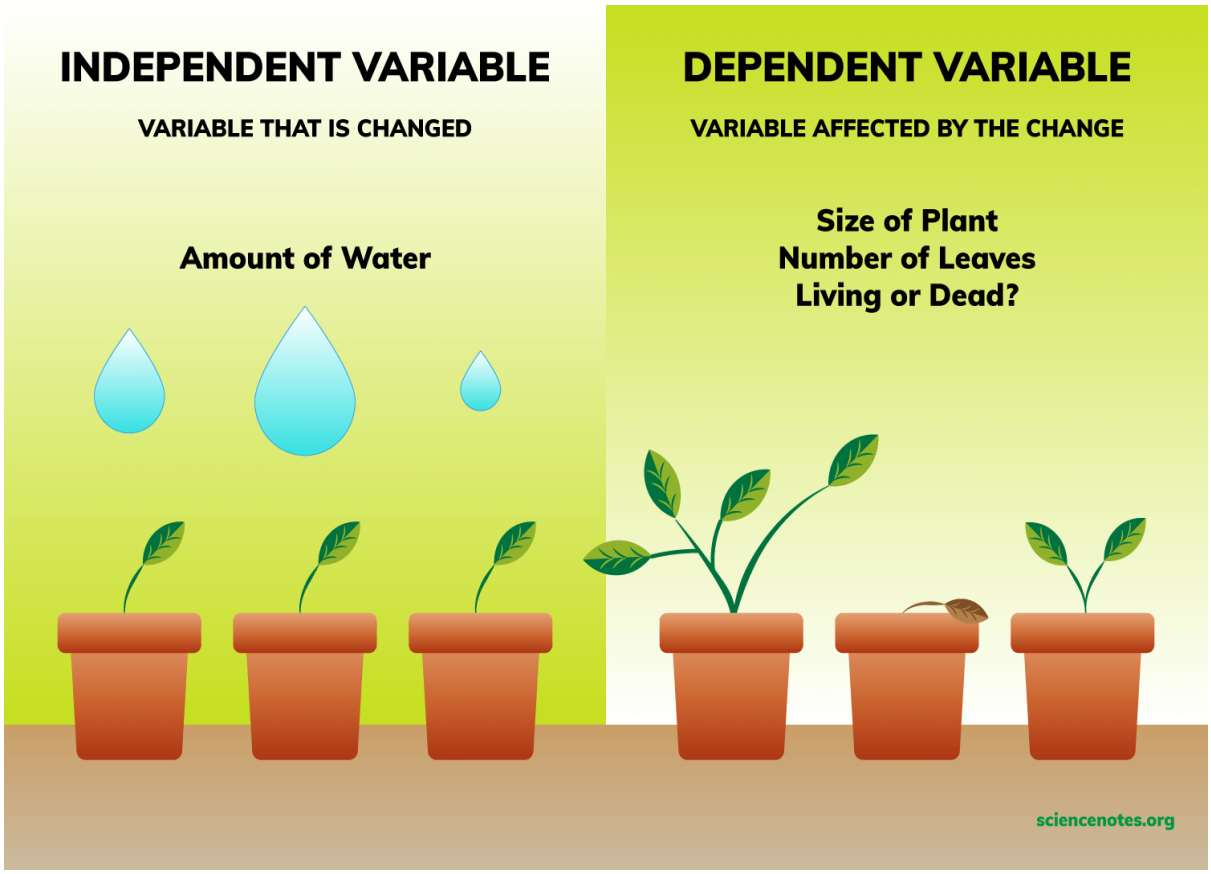
If a variable has two or more categories with no ordering. Example: political party affiliation, gender (male/female, even coded as numbers e.g. male = 0, female = 1).

Ordinal

If there is a hierarchy or order of preference. Example: economic status (low/medium/ high).

Dependent and Independent Variables

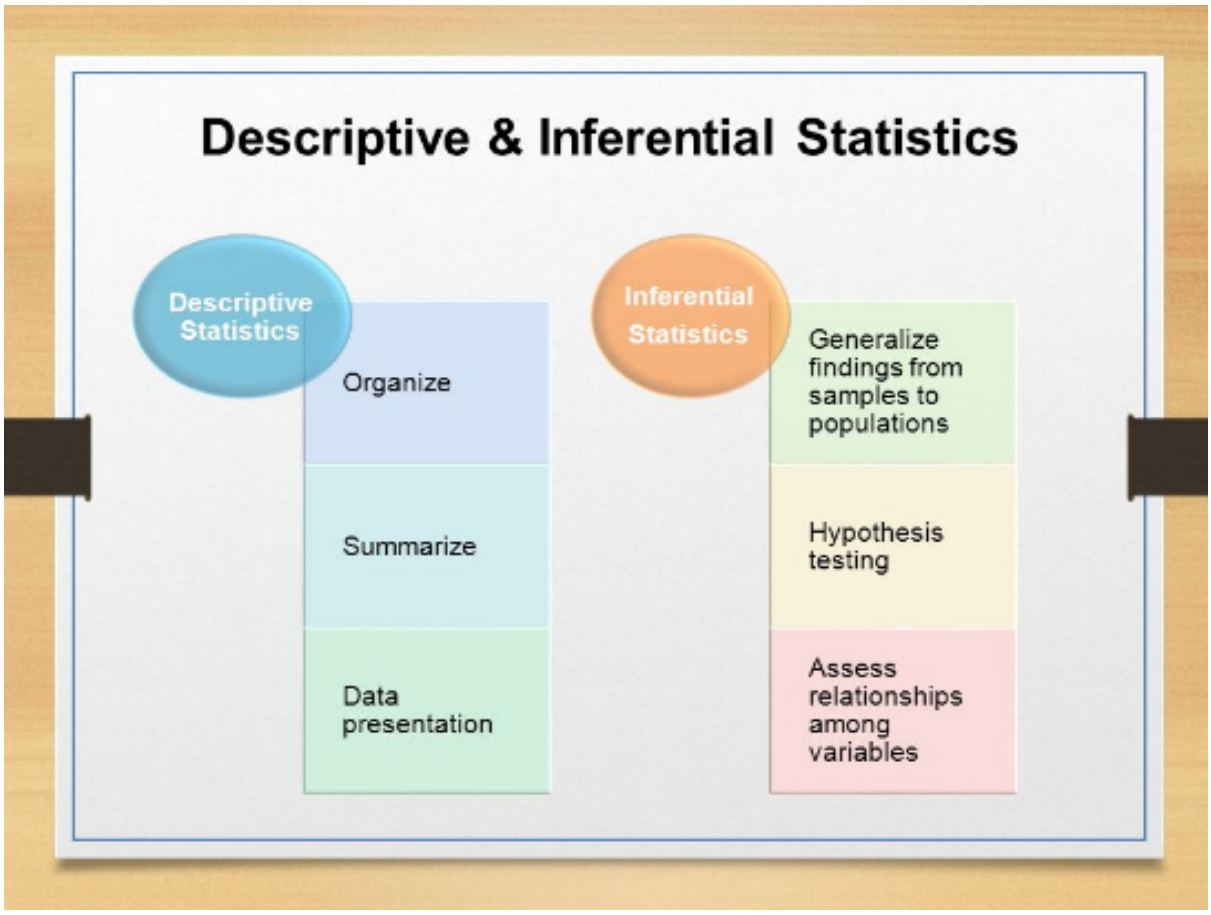
An **independent variable**, sometimes called an experimental or predictor **variable**, is a **variable** that is being manipulated in an experiment in order to observe the effect on a **dependent variable**, sometimes called an target **variable** .



Categorizing Statistics

1. Descriptive Statistics: In Descriptive Statistics your are describing, presenting, summarizing and organizing your data (population), either through numerical calculations or graphs or tables.

2. Inferential statistics: Inferential Statistics are produced by more complex mathematical calculations, and allow us to infer trends and make assumptions and predictions about a population based on a study of a sample taken from it.



Descriptive Statistics

CHARACTERISTICS OF FREQUENCY DISTRIBUTION

There are four important characteristics of frequency distribution. They are as follows:

- **Central tendency** :Measures of central tendency and location (mean, median, mode)
- **Modality** - Measures of dispersion (range, variance, standard deviation)
- **Skewness-** The extent of symmetry/asymmetry (skewness)
- **kurtosis :** The flatness or peakiness

Measures of Central Tendency

In statistics we have to deal with the mean, mode and the median. These are called the Central Tendency

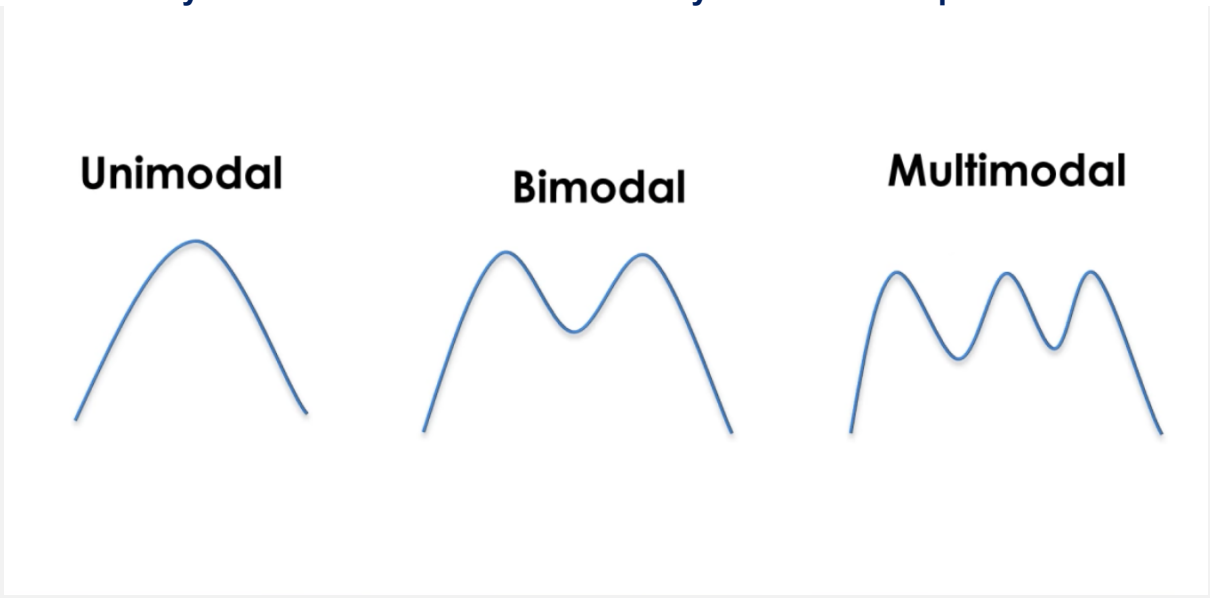
1. Mean	average of the data
2. Median	middle vale of the ordered data
3. Mode	value that occurs most often in the data

Example:
6, 1, 13, 1, 4
Mean=5; Median=4;Mode=1

An important shortcoming of the mean is that they are easily affected by extreme values.

Modality

The modality of a distribution is determined by the number of peaks it contains.



Unimodal means that the distribution has only one peak, which means it has only one frequently occurring score, clustered at the top. A bimodal distribution has two values that occur frequently (two peaks) and a multimodal has two or several frequently occurring values.

Skewness

Skewness is a measurement of the symmetry of a distribution.

it describes how much a distribution differs from a normal distribution, either to the left or to the right. The skewness value can be either positive, negative or zero

1.Symmetric

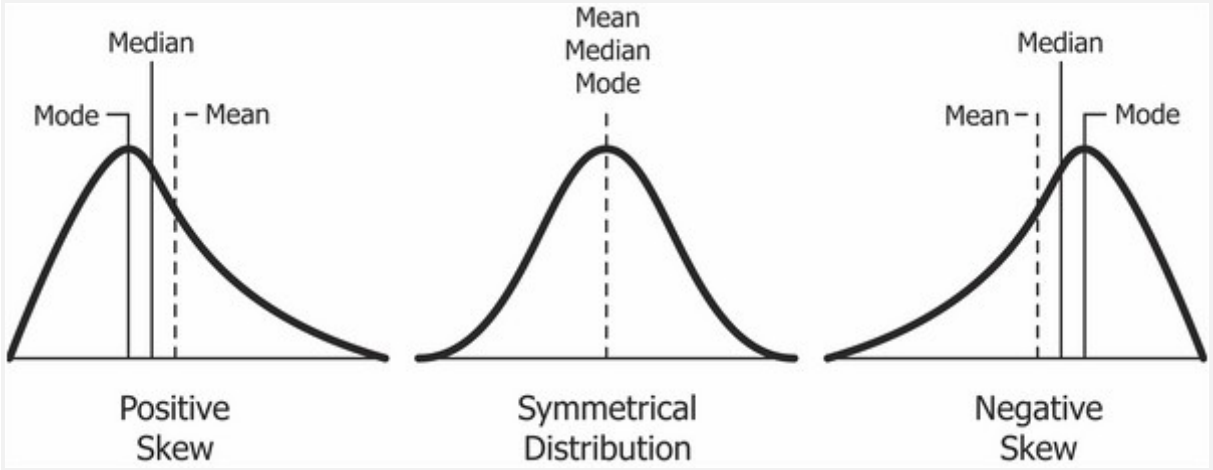
Mean, median, and mode are all the same. The distribution is bell shaped with no skewness. The distribution is described as symmetric.

2. Negatively or Left Skewed

Mean is smaller than median. Mean to the left of the median, long tail on the left.

3. Positively or Right Skewed

Mean is greater than median. Mean to the right of the median, long tail on the right.



Measuring Variability and Spread

1. Range = largest value – smallest value

Range is easy to calculate but it is very sensitive to extreme values.

Example: 5.9, 7.3, 2.5, 4.3, 8.2, 1.4
Range = 8.2 – 1.4 = 6.8

2. Interquartile Range

Quartiles divide the distribution of dataset into four equal parts.

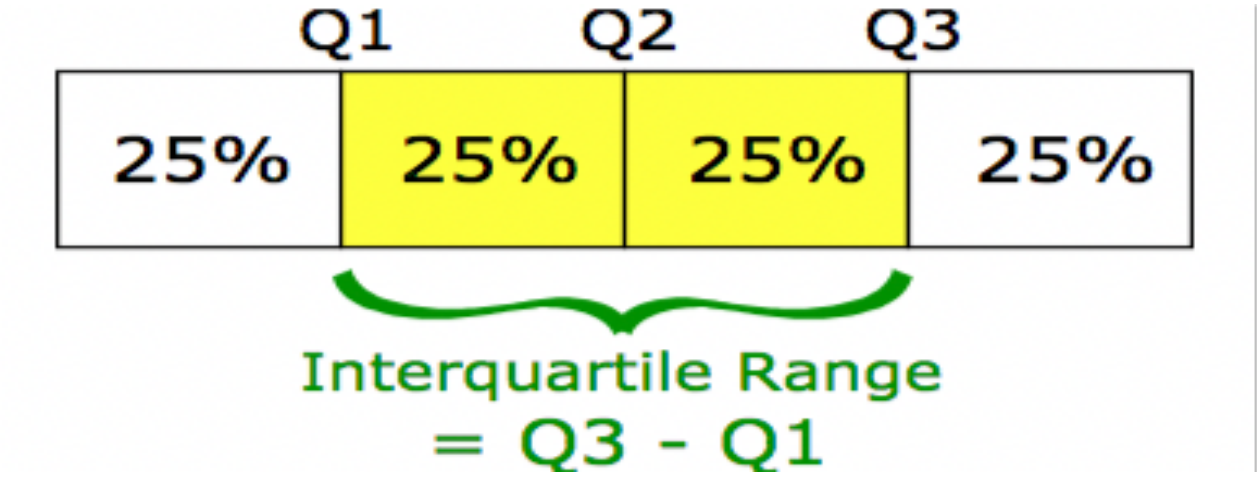
The second quartile Q2 of the data is its median (50% of data is less than Q2 and 50% is greater than Q2).

The first quartile Q1 of the dataset is the median of the lower set. 25% of data is less than Q1 and 75% of data (1 – Q1) is greater than Q1.

The third quartile Q3 of the dataset is the median of the upper set. 75% of data is less than Q3 and 25% of data (1 – Q3) is greater than Q3.

First quartile = lower quartile = 25th percentile Third quartile = upper quartile = 75th percentile

The interquartile range is the difference between the upper and the lower quartiles.
IQR = Q3 – Q1



IQR is a robust measure of variability not sensitive to outliers.

Example:
Ordered data is: 113, 124, 124, 132, 146, 151, 170

Median = Q2 = 132
Q1 = 124
Q3 = 151
IQR=Q3–Q1=151–124=27 => 27 is the range of the middle 50% of the data

3. Variance and Standard Deviation

Variance is the average squared distance from the mean.

Population Variance

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

Sample Variance

$$S^2 = \frac{\sum (X - \overline{X})^2}{n - 1}$$

μ - population mean, N - population size, X - sample mean, n - sample size

The most common measure of the data dispersion is **Standard Deviation**. It is the deviation from the mean has the same units as the original data.

Sample Standard Deviation

$$S = \sqrt{S^2}$$

Example. Calculate variance and standard deviation for the following 6 values:

1, 2, 3, 3, 4, 5

Variance:

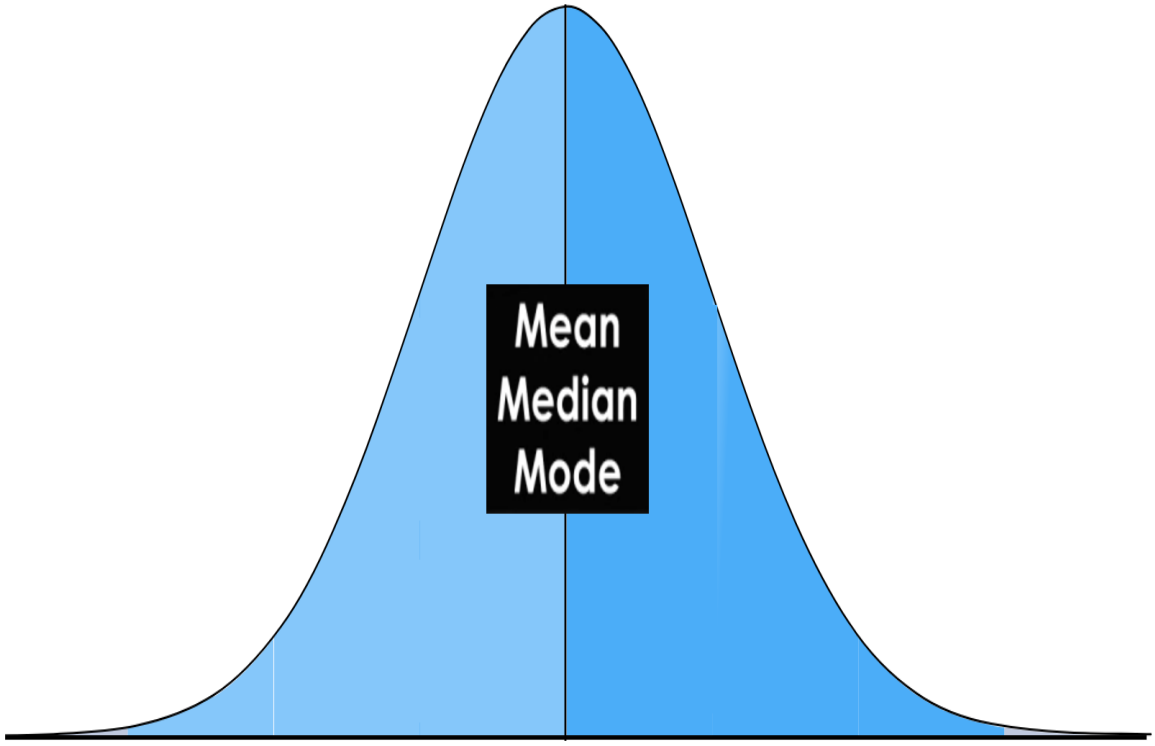
Mean = (1 + 2 + 3 + 3 + 4 + 5) / 6 = 3
s2 = [(1 – 3)2 + (2 – 3)2 + (3 – 3)2 + (3 – 3)2 + (4 – 3)2 + (5 – 3)2] / (6 – 1) = 2

Standard deviation:
s = sqrt(s2) = sqrt(2) = 1.414

Normal Distribution

A normal Distribution is given if the data is symmetrical, bell-shaped, centred and unimodal.

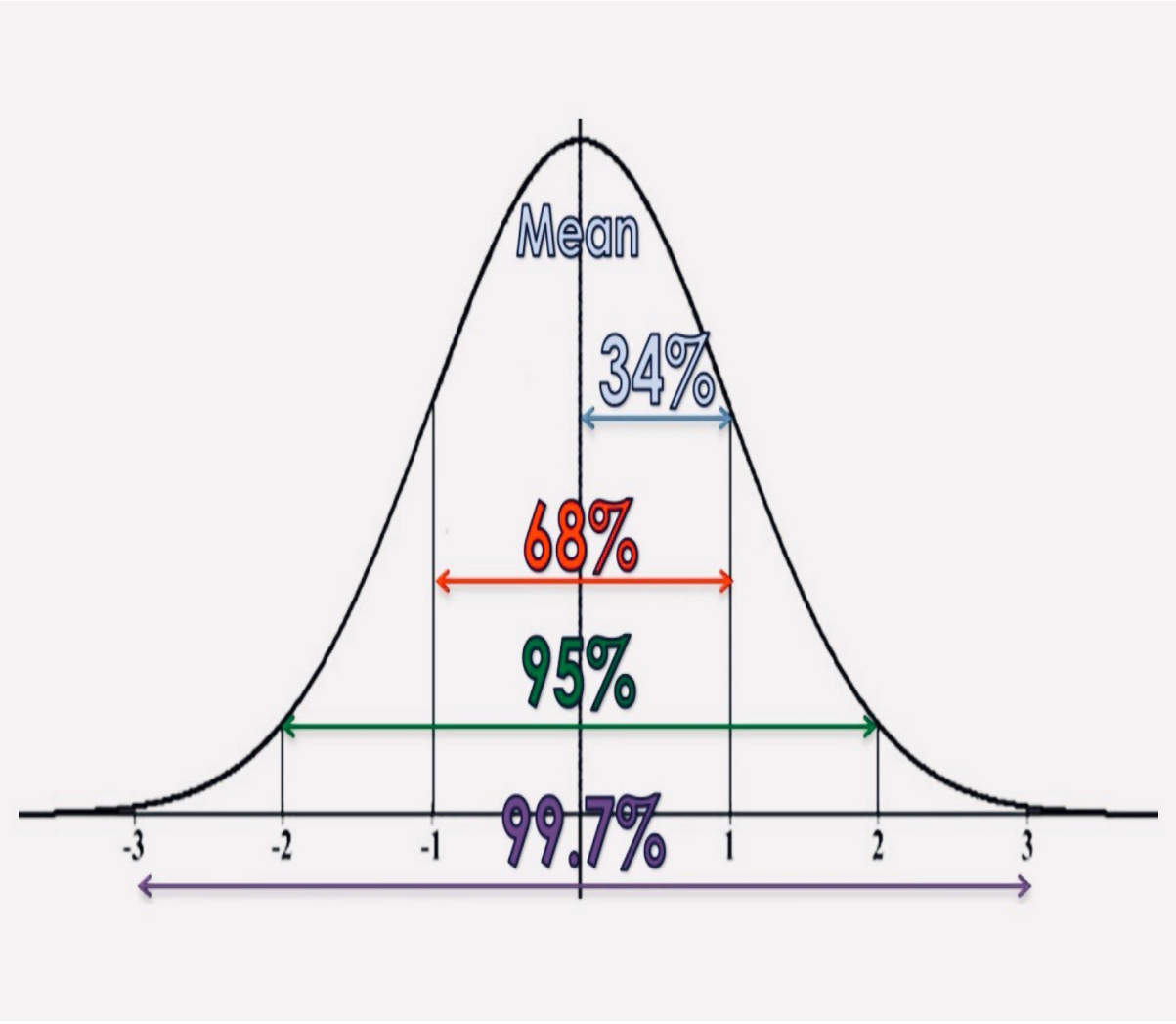
It is sometimes called the “bell curve“ or the “Gaussian curve“.In a perfect normal distribution, each side is an exact mirror of the other. It should look like the distribution on the picture below:



Interpreting standard deviation

Empirical rule – if the values in the dataset follow a bell shaped (normal) distribution then:

- Approximately 68% of observations lies within 1 standard deviation of the mean.
- Approximately 95% of observations lies within 2 standard deviations of the mean.
- Approximately 99.7% of observations lies within 3 standard deviations of the mean.



Boxplot

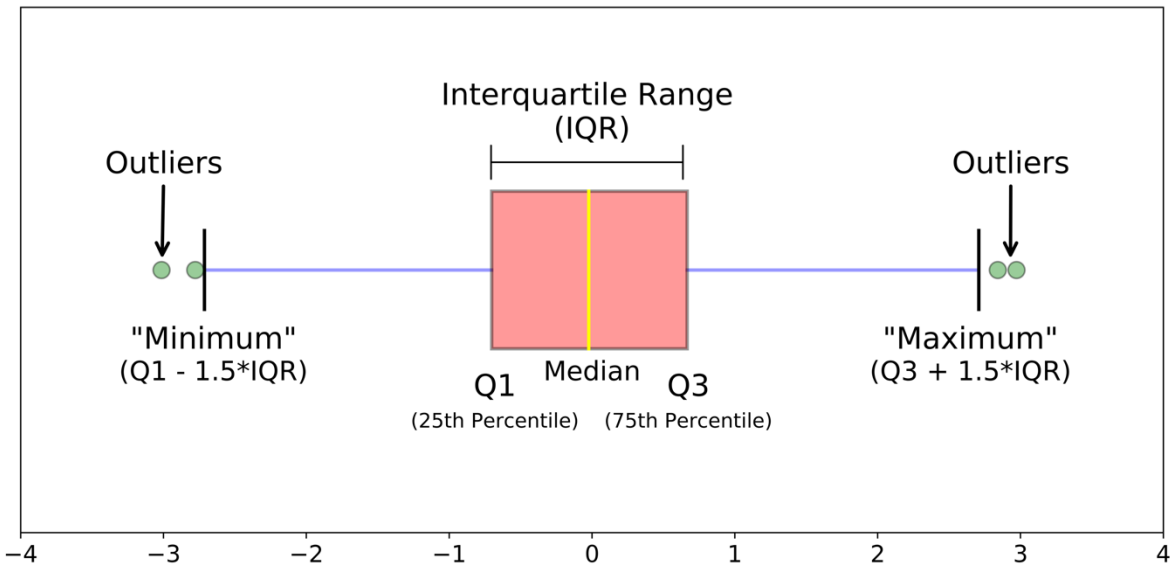
A boxplot is a graph that gives you a good indication of how the values in the data are spread out
The box plot is a standardized way of displaying the distribution of data based on the five number summary:
1.median (Q2/50th Percentile): the middle value of the dataset.

2.First quartile (Q1/25th Percentile): the middle number between the smallest number (not the “minimum”) and the median of the dataset.

3.third quartile (Q3/75th Percentile): the middle value between the median and the highest value (not the “maximum”) of the dataset.

4.Maximum value - Q3 + 1.5*IQR

5.Minimum Value - Q1 -1.5*IQR

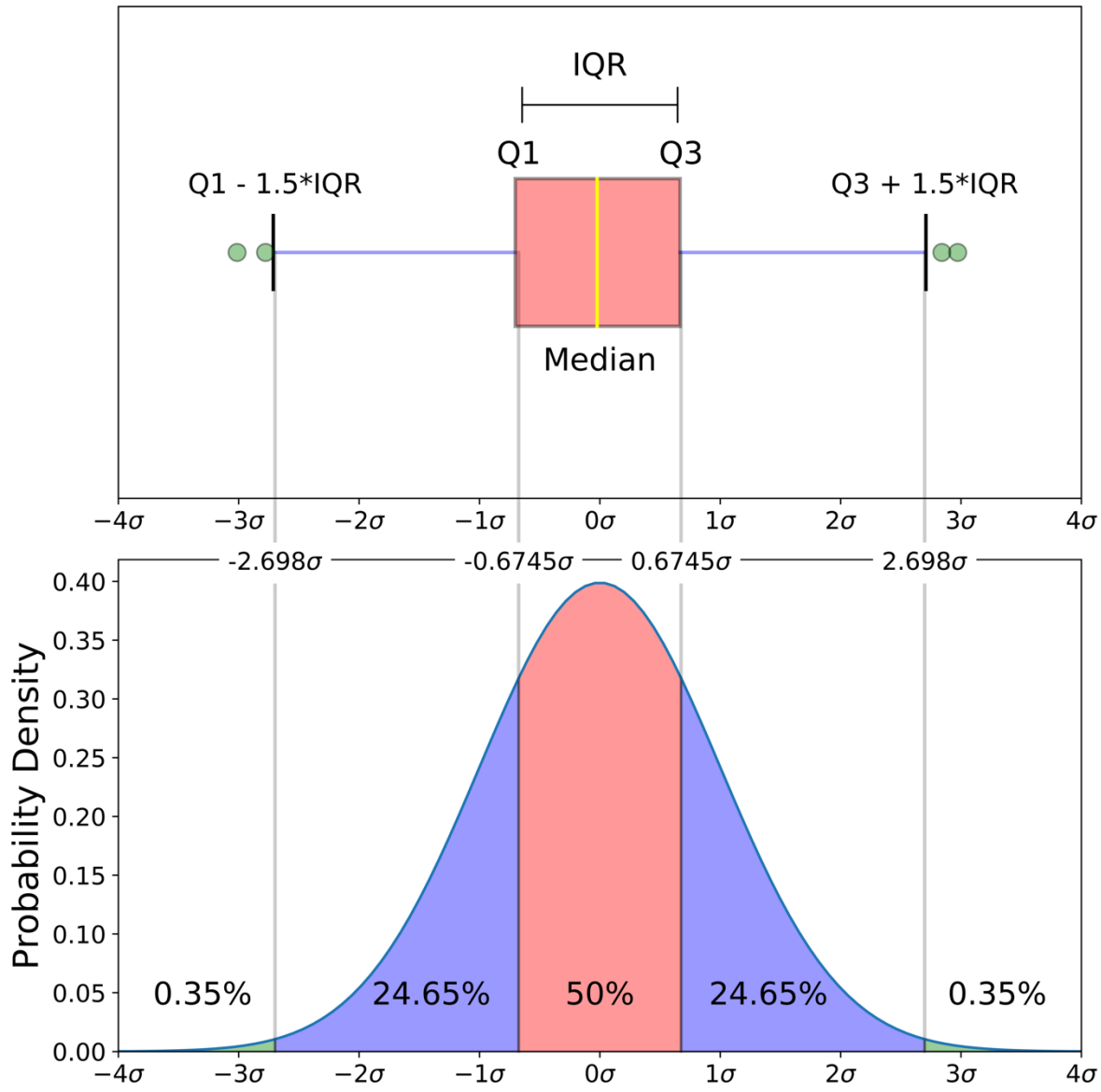


interquartile range (IQR): 25th to the 75th percentile.

whiskers (shown in blue)

outliers (shown as green circles)

Boxplot on Normal Distribution



Coefficient of Variation

The coefficient of variation (relative standard deviation) is a statistical **measure of the dispersion of data points around the mean**

Coefficient of Variation = (Standard Deviation / Mean) * 100

Covariance

Covariance is a measure of how much two random variables vary together. It's similar to variance, but where variance tells you how a *single* variable varies, **co** variance tells you how **two** variables vary together.

Higher the value stronger the relationship between them

Covariance Formula

For Population

$$Cov(x,y) = \frac{\sum (x_i - \bar{x}) * (y_i - \bar{y})}{N}$$

For Sample

$$Cov(x,y) = \frac{\sum (x_i - \bar{x}) * (y_i - \bar{y})}{(N - 1)}$$

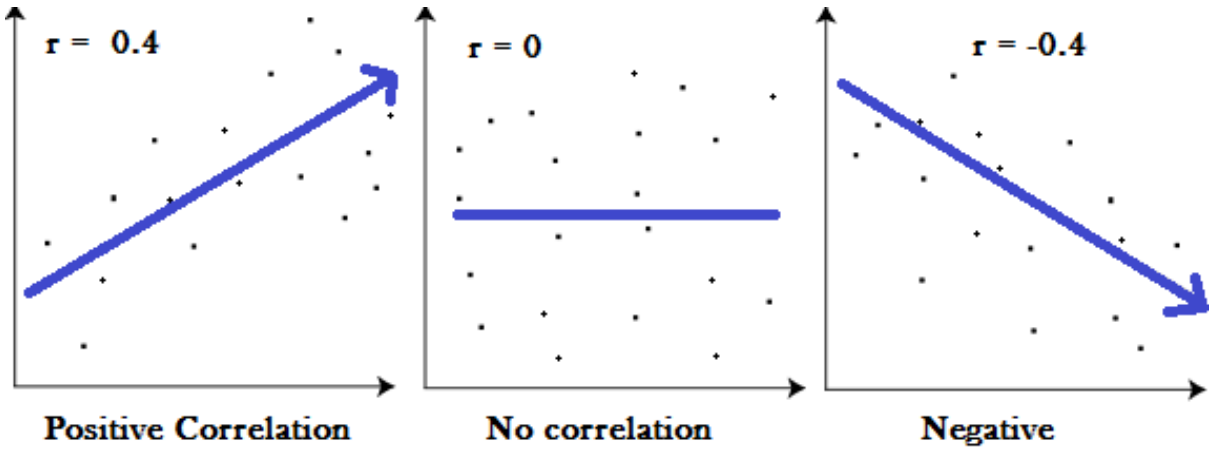
Correlation Coefficient

A correlation coefficient is a numerical measure of some type of correlation, meaning a statistical **relationship between two variables**.
The variables may be two columns of a given data set of observations, often called a sample, or two components of a multivariate random variable with a known distribution.

$$r_{xy} = \frac{Cov(x,y)}{S_x S_y}$$

Types of Correlation

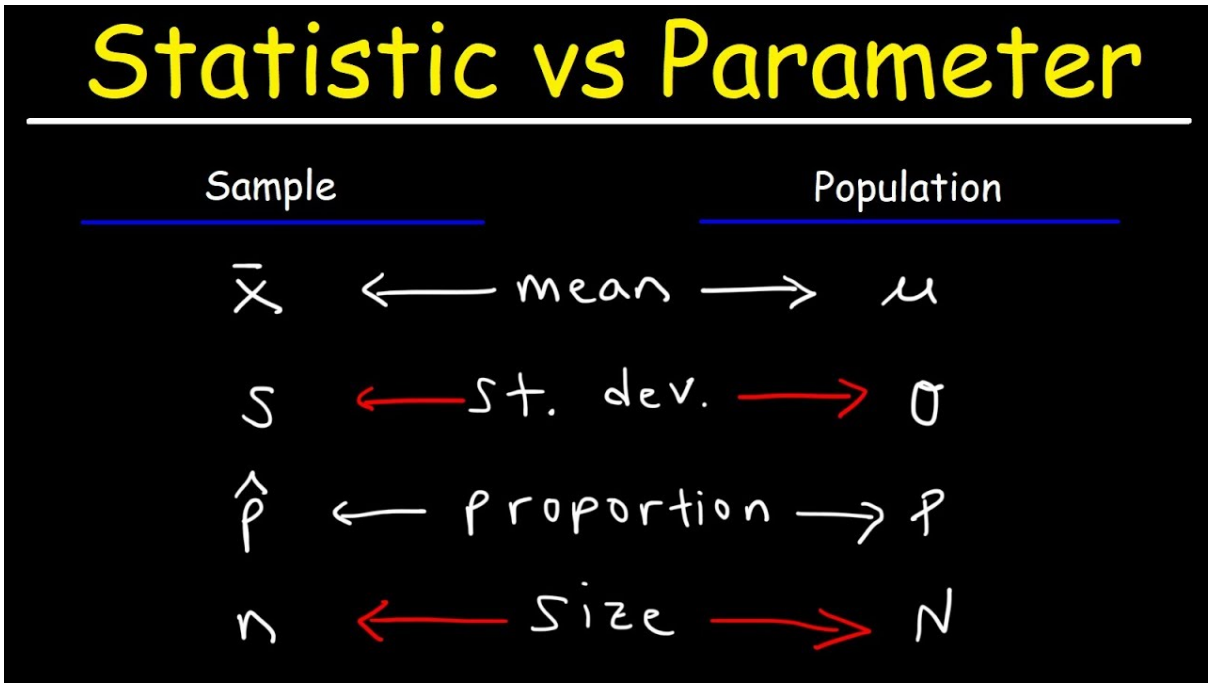
Correlation coefficient formulas are used to find how strong a relationship is between data. The formulas return a value between -1 and 1, where:
1 indicates a strong positive relationship.
-1 indicates a strong negative relationship.
A result of zero indicates no relationship at all.



- A correlation coefficient of 1 means that for every positive increase in one variable, there is a positive increase of a fixed proportion in the other. For example, shoe sizes go up in (almost) perfect correlation with foot length.

- A correlation coefficient of -1 means that for every positive increase in one variable, there is a negative decrease of a fixed proportion in the other. For example, the amount of gas in a tank decreases in (almost) perfect correlation with speed.
- Zero means that for every increase, there isn't a positive or negative increase. The two just aren't related.

Statistics and Parameter



Degrees of Freedom

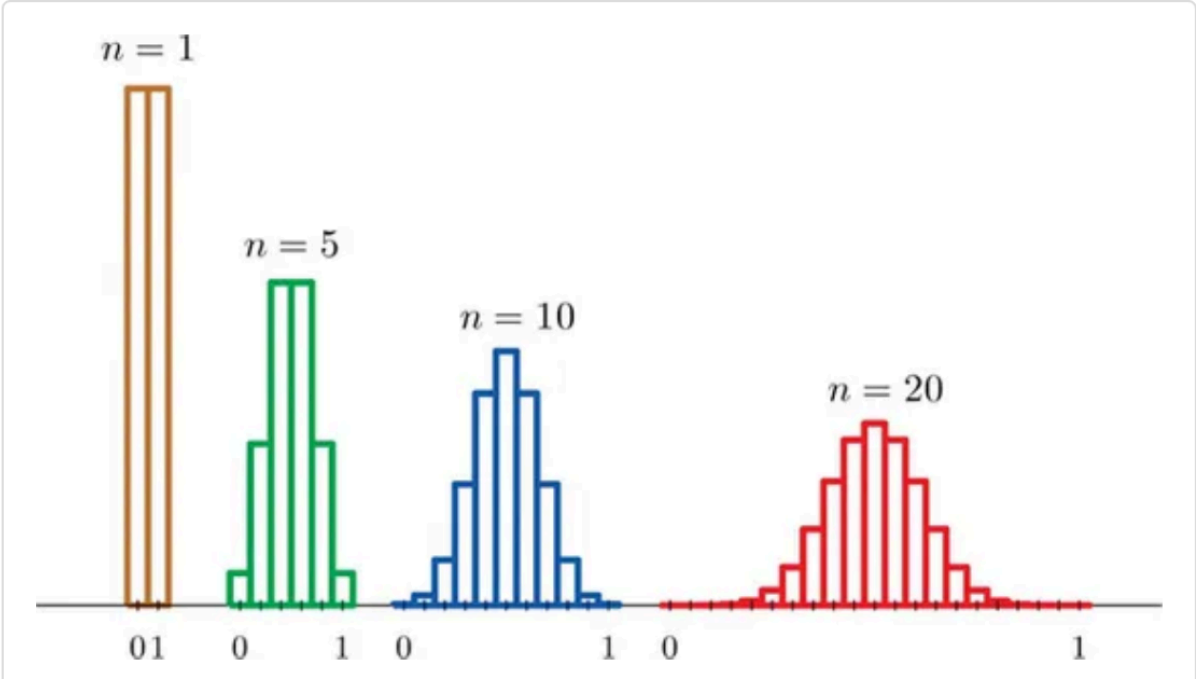
In statistics, the degrees of freedom (DF) indicate the number of independent values that can vary in an analysis without breaking any constraints.

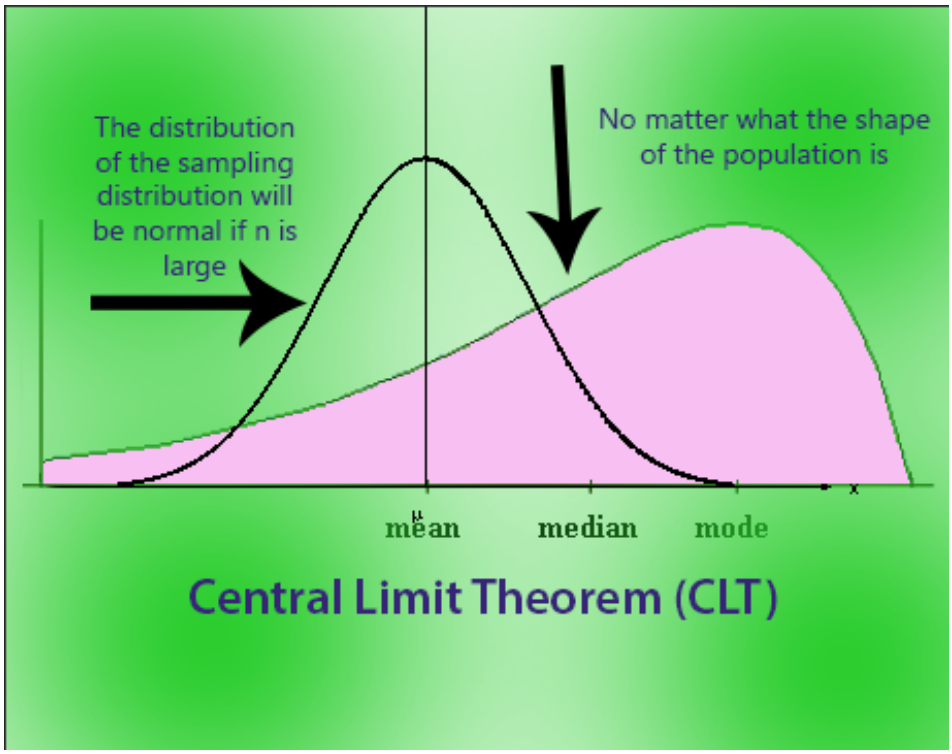
The degrees of freedom equal your sample size minus the number of parameters you need to calculate during an analysis.

Degrees of freedom is a combination of how much data you have and how many parameters you need to estimate

Central limit theorem

The central limit theorem states that the sampling distribution of the mean approaches a normal distribution, as the sample size increases.





• As the sample size increases, the distribution of frequencies approximates a bell-shaped curved (i.e. normal distribution curve).

• Sample size equal to or greater than 30 are required for the central limit theorem to hold true.

• A sufficiently large sample can predict the parameters of a population such as the mean and standard deviation.

Z-Score

Z-value (z-score or standard score) represents the number of standard deviations an observation is from the mean.

Z = (observed value – mean) / SD