**Statistics**

Table of Contents

[Curriculum for studying Statistics 6](#_Toc144466429)

[Things to know before studying Statistics 6](#_Toc144466430)

[Basic/Easy Level Topics 6](#_Toc144466431)

[Moderate Level Topics 6](#_Toc144466432)

[Expert Level Topics 7](#_Toc144466433)

[1. Type of statistics 8](#_Toc144466434)

[1.1. Descriptive statistics 8](#_Toc144466435)

[1.2. Inferential statistics 9](#_Toc144466436)

[2. Population and sample 10](#_Toc144466437)

[2.1. Why do we work on sample data and not on population 11](#_Toc144466438)

[2.2. Types of sampling 11](#_Toc144466439)

[2.2.1. Probability sampling 12](#_Toc144466440)

[2.2.2. Non-probability sampling 13](#_Toc144466441)

[3. Types of data 14](#_Toc144466442)

[3.1. Numerical data 14](#_Toc144466443)

[3.2. Categorical data 14](#_Toc144466444)

[3.3. Timeseries data 14](#_Toc144466445)

[3.4. Text data 14](#_Toc144466446)

[4. Level of data measurement 15](#_Toc144466447)

[4.1. Nominal scale 15](#_Toc144466448)

[4.2. Ordinal scale 15](#_Toc144466449)

[4.3. Interval scale 15](#_Toc144466450)

[4.4. Ratio scale 15](#_Toc144466451)

[5. Measure of central tendency 15](#_Toc144466452)

[5.1. Mean | Average 15](#_Toc144466453)

[5.2. Median 16](#_Toc144466454)

[5.3. Mode 16](#_Toc144466455)

[5.4. Mean, Mode and Median in different distributions 17](#_Toc144466456)

[6. Measure of spread/dispersion/variation 17](#_Toc144466457)

[6.1. Range 17](#_Toc144466458)

[6.2. Variance & standard deviation 17](#_Toc144466459)

[6.2.1. Variance 17](#_Toc144466460)

[6.2.2. Standard Deviation 18](#_Toc144466461)

[6.2.3. Mean Absolute Deviation (MAD) 18](#_Toc144466462)

[6.3. Interquartile distance | Box plot 19](#_Toc144466463)

[6.4. Percentiles – WIP 19](#_Toc144466464)

[7. Effect of transformation on central tendency and spread – WIP 20](#_Toc144466465)

[8. Measure of symmetricity 20](#_Toc144466466)

[8.1. Skewness 20](#_Toc144466467)

[8.2. Kurtosis 21](#_Toc144466468)

[8.3. Types of kurtosis 22](#_Toc144466469)

[8.3.1. Mesokurtic distribution 22](#_Toc144466470)

[8.3.2. Platokurtic distribution 23](#_Toc144466471)

[8.3.3. Leptokurtic distribution 24](#_Toc144466472)

[Note 25](#_Toc144466473)

[9. Measure of relationship b/w two variables 25](#_Toc144466474)

[9.1. Covariance 25](#_Toc144466475)

[9.2. Correlation Coefficient | Pearson’s R 25](#_Toc144466476)

[9.3. Spearman rank correlation coefficient 26](#_Toc144466477)

[9.4. Kendall rank correlation coefficient 27](#_Toc144466478)

[9.5. Point-biserial correlation coefficient 27](#_Toc144466479)

[9.6. Phi coefficient 27](#_Toc144466480)

[9.7. Cramer's V 27](#_Toc144466481)

[Note 27](#_Toc144466482)

[10. Measure of relationship b/w more than 2 variables [WIP] 27](#_Toc144466483)

[10.1. Multiple Linear Regression (MLR) 27](#_Toc144466484)

[10.2. Multivariate Linear Regression (MVLR) 27](#_Toc144466485)

[10.3. Canonical Correlation Analysis (CCA) 27](#_Toc144466486)

[10.4. Partial Least Squares Regression (PLSR) 28](#_Toc144466487)

[10.5. Principal Component Regression (PCR) 28](#_Toc144466488)

[Note 28](#_Toc144466489)

[11. Outlier 29](#_Toc144466490)

[11.1. Causes of outlier 29](#_Toc144466491)

[11.2. Methods to identify outlier 29](#_Toc144466492)

[Note 30](#_Toc144466493)

[12. Z – score 30](#_Toc144466494)

[13. Linear and non-linear variables or function or equation 31](#_Toc144466495)

[13.1. Linear variable or function 31](#_Toc144466496)

[13.2. Non-linear variable or function 31](#_Toc144466497)

[Note 31](#_Toc144466498)

[14. Parametric and non-parametric [WIP] 31](#_Toc144466499)

[15. Hypothesis Testing 32](#_Toc144466500)

[15.1. Step-by-step guide to learn Hypothesis Testing 33](#_Toc144466501)

**STATISTICS**

# Curriculum for studying Statistics

## Things to know before studying Statistics

1. Probability Fundamentals: Review the basic concepts of Probability, including sample spaces, events, and the axioms of Probability.
2. Mathematical Notation: Familiarize yourself with common Statistical notation, including notation for random variables, parameters, and estimators.

## Basic/Easy Level Topics

1. Data Types and Data Collection: Understand different types of data (nominal, ordinal, interval, ratio) and methods of data collection.
2. Exploratory Data Analysis (EDA): Learn techniques for summarizing and visualizing data using histograms, box plots, scatter plots, and summary statistics.
3. Probability Distributions: Review common Probability distributions, such as the binomial, geometric, Poisson, uniform, and normal distributions, and their properties.
4. Sampling Techniques: Explore various sampling methods, including simple random sampling, stratified sampling, and cluster sampling.
5. Confidence Intervals: Understand how to construct confidence intervals to estimate population parameters based on sample data.
6. Hypothesis Testing: Learn the principles and techniques of hypothesis testing, including null and alternative hypotheses, significance levels, and p-values.

## Moderate Level Topics

1. Regression Analysis: Study simple linear regression, multiple regression, and their applications in modeling relationships between variables.
2. Analysis of Variance (ANOVA): Understand the principles and techniques for comparing means across multiple groups or treatments.
3. Nonparametric Methods: Explore nonparametric statistical tests and techniques that do not rely on specific distributional assumptions.
4. Experimental Design: Learn about principles and techniques for designing experiments and controlling for confounding variables.
5. Sampling Distributions: Understand the concept of sampling distributions and their role in statistical inference.
6. Bayesian Statistics: Familiarize yourself with the basics of Bayesian inference, including prior and posterior distributions, Bayesian estimation, and hypothesis testing.

## Expert Level Topics

1. Time Series Analysis: Study methods for analyzing and forecasting time series data, including autoregressive integrated moving average (ARIMA) models.
2. Multivariate Analysis: Explore techniques for analyzing and modeling relationships between multiple variables, such as principal component analysis (PCA) and factor analysis.
3. Advanced Regression Techniques: Dive deeper into regression analysis, including logistic regression, generalized linear models (GLMs), and regularization techniques.
4. Experimental Design: Learn advanced topics in experimental design, including factorial designs, fractional factorial designs, and response surface methodology.
5. Statistical Modeling: Study advanced statistical models, such as hierarchical models, generalized estimating equations (GEE), and mixed-effects models.
6. Statistical Software: Gain proficiency in using statistical software packages like R or Python libraries (e.g., NumPy, pandas, statsmodels) for data analysis and modeling.

Remember, this curriculum is designed to provide a comprehensive understanding of Statistics, with practical relevance for data analysis and modeling. Probability and Statistics are essential components in the field of machine learning, enabling you to understand and apply various algorithms and assess model performance.

# Type of statistics

## Descriptive statistics

Descriptive statistics are used to summarize, describe and present the main features of a dataset, and to provide a general understanding of the data.

Some of the topics that are commonly studied under descriptive statistics include:

* **Data summarization**: This includes techniques for summarizing data such as measures of central tendency (mean, median, mode), measures of dispersion (range, variance, standard deviation) and frequency distributions.
* **Data visualization**: This includes techniques for visualizing data such as histograms, bar charts, scatter plots, line plots, and box plots.
* **Exploratory data analysis**: This includes techniques for exploring and understanding data, such as identifying patterns, outliers, and trends.
* **Measures of association**: This includes techniques for measuring the relationship between two or more variables, such as correlation and covariance.
* **Probability distributions**: This includes the study of probability distributions such as the normal distribution, binomial distribution, Poisson distribution, etc.

Overall, descriptive statistics provide a summary of the data and it's important to understand the data properties before applying inferential statistics or making predictions using the data.

## Inferential statistics

Inferential statistics is a branch of statistics that deals with making inferences or predictions about a population based on a sample of data. It allows researchers to make generalizations about a population from a sample and to estimate population parameters from sample statistics.

Most of the time, you can only acquire data from samples, because it is too difficult or expensive to collect data from the whole population that you’re interested in.

While descriptive statistics can only summarize a sample’s characteristics, inferential statistics use your sample to make reasonable guesses about the larger population.

With inferential statistics, it’s important to use random and unbiased sampling methods. If your sample isn’t representative of your population, then you can’t make valid statistical inferences or generalize.

Inferential statistics have two main uses:

* Making estimates about populations (for example, the mean SAT score of all 11th graders in the US).
* Testing hypotheses to draw conclusions about populations (for example, the relationship between SAT scores and family income).

~~Some of the topics studied under inferential statistics include:~~

* ~~Estimation of population parameters (e.g. mean, standard deviation)~~
* ~~Hypothesis testing~~
* ~~Confidence intervals~~
* ~~Correlation and regression analysis~~
* ~~Analysis of variance (ANOVA)~~
* ~~Parametric statistics~~

~~Parametric statistics assume specific distributions for the data, often assuming a normal distribution, and involve estimating population parameters (e.g., means, variances) and performing hypothesis tests using those assumptions.~~

* ~~Non-parametric statistics~~

~~Non-parametric statistics, also known as distribution-free statistics, are a set of statistical methods that do not make strong assumptions about the underlying distribution of the data. In contrast to parametric statistics, which assume specific distribution shapes (like normal distribution), non-parametric methods are more flexible and can be applied to a wider range of data types and distributions.~~

~~Non-parametric statistics are often used in situations where the data do not meet the assumptions of parametric methods, or when the sample size is small. They are also useful when dealing with ordinal or categorical data, where the concept of means and variances may not be meaningful.~~

* ~~Bayesian statistics~~

~~Bayesian statistics is a branch of statistics that is based on the principles of Bayesian probability theory. It provides a framework for updating our beliefs about uncertain events or hypotheses as new evidence becomes available.~~

* ~~Power and sample size calculations~~

# Population and sample

In statistics, a **population** is defined as the set of all individuals or objects that possess a certain characteristic or set of characteristics and is denoted by the symbol N. A population can be finite or infinite, but it is usually not possible to measure or observe every member of the population.

A **sample**, on the other hand, is a smaller group of individuals or objects selected from the population to represent it. The sample is denoted by the symbol n. The sample size can vary depending on the research question and the specific requirements of the study, but it is typically much smaller than the population.

It's important to ensure that a sample is true representative of the population.

Statistical inferences or conclusions about the population are made based on the sample data. The sample mean, sample variance, and sample standard deviation, for example, are used as estimates of the population mean, population variance, and population standard deviation, respectively.

It's also important to note that sample statistics are subject to random variation, so the calculated statistics such as sample mean, variance, etc. may differ from the population value.

The **characteristics** of **samples** are called **statistics** and of ***populations,*** it is called a **parameter**:

* A **statistic** is a measure that describes the **sample** (e.g., sample mean).
* A **parameter** is a measure that describes the whole **population** (e.g., the population mean).
* **Sampling error** is the difference between a parameter and a corresponding statistic. Since in most cases you don’t know the real population parameter, you can use inferential statistics to estimate these parameters in a way that takes sampling error into account.

There are two important types of estimates you can make about the population: point estimates and interval estimates.

* A **point estimate** is a single value estimate of a parameter. For instance, a sample mean is a point estimate of a population mean.
* An **interval estimate** gives you a range of values where the parameter is expected to lie. A confidence interval is the most common type of interval estimate.

Both types of estimates are important for gathering a clear idea of where a parameter is likely to lie.

## Why do we work on sample data and not on population

There are several reasons why researchers often work with sample data instead of the entire population:

* **Cost**: Collecting data from the entire population can be expensive, time-consuming, and resource-intensive. Sampling allows researchers to collect data from a smaller, more manageable subset of the population.
* **Feasibility**: In some cases, it may not be possible to access or collect data from the entire population. For example, a study on endangered species would not be possible if the population is small and hard to find.
* **Efficiency**: Analysing data from a sample is typically faster and more efficient than analysing data from the entire population.
* **Representativeness**: A well-chosen sample can be representative of the population, making it possible to generalize findings from the sample to the population.
* **Ethical concerns**: In some cases, it may not be ethical or practical to collect data from the entire population. For example, a study on sensitive personal information would not be possible if the population is not willing to share such information.

It's important to note that the accuracy and generalizability of the findings are dependent on how well the sample represents the population and how the sample is selected. A good sample design, proper execution of the sampling process and appropriate analysis of the sample data can ensure that the results are valid and reliable.

## Types of sampling

Probability sampling and non-probability sampling are two broad categories of sampling methods.

### Probability sampling

**Probability sampling** refers to sampling methods where every member of the population has a **non-zero chance of being selected** for the sample. In probability sampling, the sample is selected **randomly** and the process is based on mathematical probability. Examples of probability sampling methods include simple random sampling, stratified sampling, and cluster sampling.

**Probabilistic sampling** is considered to be a more rigorous and objective method of sampling because the sample is selected randomly, and the probability of selection is known for every member of the population. As a result, it's possible to calculate sampling error and the results are generalizable to the population. Probabilistic sampling is more restrictive and may be more difficult to implement in certain situations, but it can provide more accurate and generalizable results.

#### Simple random sampling

In simple random sampling, each member of the population has an equal chance of being selected. For example, a researcher may use a random number generator to select a certain number of individuals from a phone book.

#### Stratified sampling

Stratified sampling involves dividing the population into different groups or **strata** based on certain characteristics, such as age, gender, or income level. A random sample is then selected from each stratum. The main purpose of stratified sampling is to ensure that the sample is representative of the population with respect to the characteristic that is used to create the strata.

For example, if a researcher wants to conduct a survey on the population that is divided by age, he/she can use stratified sampling to ensure that the sample has a proportionate representation of different age groups.

#### Cluster sampling

In cluster sampling, the population is divided into groups (clusters), and a random sample of clusters is selected. All members of the selected clusters are included in the sample. For example, a researcher may conduct a survey on a population that is divided into different geographic regions, the researcher may select a random sample of regions and conduct the survey on all the individuals living in those regions.

In cluster sampling, the criteria used to create clusters can vary depending on the research question and the population being studied. Some common criteria that are used to create clusters include:

* **Geography**: Clusters can be created based on geographic regions, such as neighbourhoods, cities, or states. This is useful when the population is spread out over a large area, and it is not practical or cost-effective to sample every individual in the population.
* **Demographics**: Clusters can be created based on demographic characteristics, such as age, gender, or income. This is useful when the population has homogeneous subgroups and the goal is to make sure that each subgroup is represented in the sample.
* **Organizational Structure**: Clusters can be created based on organizational structure, such as schools, hospitals, or workplaces. This is useful when the population is composed of individuals who are part of a larger organization or group.
* **Random**: Clusters can be created randomly, by selecting individuals from the population at random and grouping them together. This is useful when the population is homogeneous and there are no obvious criteria for creating clusters.

It's worth noting that researchers can use a combination of criteria to create clusters, depending on the study goals and population.

### Non-probability sampling

**Non-probability sampling** refers to sampling methods where the probability of selection is not known and can be **zero** for some members of the population. **In non-probability sampling, the sample is not selected randomly, instead, the sample is selected based on the availability, convenience, or judgment of the researcher.** Examples of non-probability sampling methods include convenience sampling, quota sampling, and snowball sampling.

**Non-probability sampling** is considered to be a less rigorous and objective method of sampling because the sample is not selected randomly, and the probability of selection is not known for every member of the population. As a result, it's difficult to calculate sampling error and the results may not be generalizable to the population. However, non-probability sampling is more flexible and may be useful when the research question requires a specific type of sample, or when the population is hard to access.

#### Convenience sampling

In convenience sampling, participants are selected based on their convenience or accessibility to the researcher. For example, a researcher may conduct a survey at a mall and only survey individuals who are willing to participate.

#### Quota sampling

In quota sampling, the sample is selected based on pre-determined quotas for different subgroups in the population. It is based on pre-set condition.

For example:

* A manager has decided to choose a 10 people for promotion so he set the condition that whoever will wish him on his birthday will get the promotion. Then all the first 10 people who wished him will get selected.
* A researcher may conduct a survey and try to get a certain number of participants from different age groups, gender, or ethnic background.

#### Snowball sampling

In snowball sampling, participants recruit others from their own network to participate in the study. For example, a researcher may conduct a study on a specific group of people and ask the participants to recruit their friends or family members to participate in the study.

#### Intentional sampling

In intentional sampling, participants are selected based on the specific characteristic or criteria specified by the researcher. For example, a researcher may conduct a study on people who have a specific medical condition.

# Types of data

## Numerical data

Numerical data, also known as quantitative data, is data that can be measured and expressed as numbers.

It can be further divided into two types: discrete and continuous.

* **Discrete data** can only take on specific, separate values (such as whole numbers).
* **Continuous data** can take on any value within a certain range (such as weight or temperature).

## Categorical data

Categorical data, also known as qualitative data, is data that can be divided into categories or groups. It cannot be measured or expressed as numbers, but it can be described using words or labels.

Categorical data can be further divided into two types: nominal and ordinal.

* **Nominal data** has no inherent order or ranking.
* **Ordinal data** has a specific order or ranking.

## Timeseries data

Time series data is a set of data points collected at specific times. It is often used in fields such as economics, finance, and weather forecasting. It can be either numerical or categorical data.

## Text data

Text data is a set of unstructured or semi-structured data, often in the form of text, images, videos, etc. and it's mainly used in Natural Language Processing and Computer Vision.

It's important to note that these categories are not mutually exclusive, and a data set may contain elements of multiple types. For example, a survey asking for both age and gender would have both numerical (age) and categorical (gender) data.

# Level of data measurement

## Nominal scale

It is the lowest level of measurement, nominal data consists of categorical data that can be grouped into categories or labels, but the categories have no inherent order or ranking. Examples of nominal data include:

* Gender
* Race
* Religious affiliation

## Ordinal scale

Ordinal data also consists of categorical data, but the categories have a specific order or ranking.

Examples of ordinal data include:

* Education level (high school, college, graduate school).
* Social class (lower, middle, upper).
* Satisfaction rating (very satisfied, satisfied, neutral, dissatisfied, very dissatisfied).

## Interval scale

This level of measurement is for numerical data, where the differences between values are meaningful, but there is no true zero point.

Examples of interval data include:

* Temperature measured in degrees Celsius or Fahrenheit.
* Latitude and Longitude measurement in degrees.

## Ratio scale

The highest level of measurement, ratio data is also numerical, but it has a true zero point and meaningful ratios between values.

Examples of ratio data include:

* Weight measured in kilograms or pounds.
* Length measured in meters or feet.
* Income measured in any currency

# Measure of central tendency

## Mean | Average

The mean (also known as the average) is a mathematical concept used to describe a set of numerical data.

It is the sum of the data set divided by the number of data points.

The mean is often used to describe the central tendency of a set of data, providing a single value that represents the "typical" or "average" value within the set.

Note: The presence of an outlier can change the mean value significantly.

## Median

The median is a statistical measure that is used to describe the central tendency of a set of numerical data.

It is the middle value of a dataset when the values are arranged in ascending or descending order.

Properties:

* It is considered to be a more **robust** measure than the mean because it is not affected by outliers or extreme values.
* Median is more **stable** than mean because adding a new value may not change the median significantly.
* It is **not** **calculated** **using** **the** **entire data**, we are simply looking for the middle value instead of using the actual values of data.

To calculate the median of a dataset:

* First arrange the data in numerical order.
* If the dataset has an **odd number** of observations, the median is the middle value. Simply find the value that separates the dataset in two equal halves.
* If the dataset has an **even number** of observations, the median is the average of the two middle values. Find the middle two values and add them together, then divide by two.

## Mode

In statistics, the mode is a measure of central tendency that describes the most frequently occurring value in a dataset. A dataset can have one mode, more than one mode, or no mode at all.

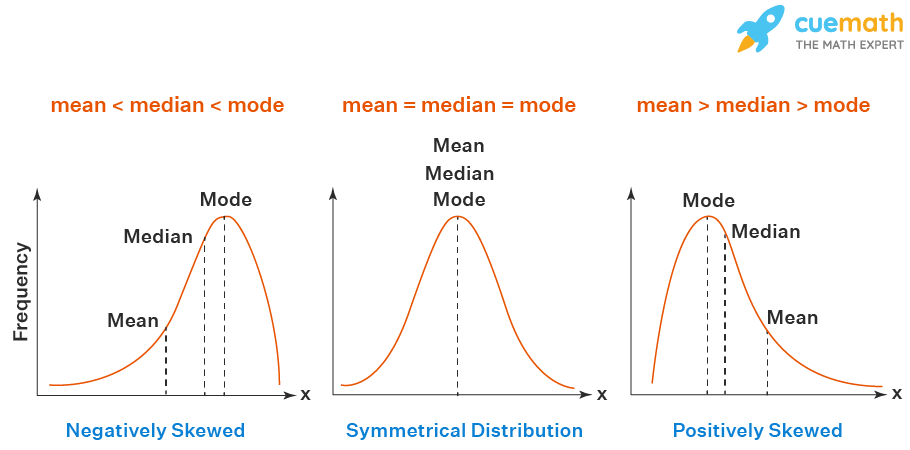
A dataset that has only one mode is called **unimodal**, a dataset that has two modes is called **bimodal**, and a dataset that has more than two modes is called **multimodal**.

To find the mode of a dataset, you can simply count the number of occurrences of each value and identify the value(s) that occurs most frequently.

If each value in the dataset appears only once, then there is no mode in the dataset.

It is worth noting that mode is not always defined for continuous data, because in continuous data it is hard to identify the most frequent value, and also mode is not useful in datasets with a large number of unique values.

## Mean, Mode and Median in different distributions



# Measure of spread/dispersion/variation

## Range

The range is a simple but crude measure of the spread of a dataset.

**Range** = (Maximum value – Minimum value) in the dataset

It is **sensitive to outliers** and does not take into account the distribution of the values within the dataset.

It is important to note that range is not always the best measure of spread and there are other measures such as standard deviation or interquartile range that may be more appropriate.

## Variance & standard deviation

Variance, standard deviation, mean standard deviation and somewhat box plot (5 point summary) are some methods that measures of the spread of a dataset. They are used to describe how much the values in a dataset deviate from the mean (average) of the dataset.

Note: Always try to interpret variance with the mean.

### Variance

**It is a measure of the average of the squared differences from the mean.**

**It is calculated by taking the average of the squared differences of each value in the dataset from the mean.**

### **Standard Deviation**

**It is the square root of the variance, it also measures the dispersion around the mean but in the same units as the values (instead of square units with variance).**

**Standard deviation is a more interpretable measure than variance, as it is expressed in the same units as the original data. It is a commonly used measure of spread and can provide a sense of how spread out the data is.**

* **Low SD: A low SD indicates that the data points are clustered closely around the mean. This means that the data is relatively consistent, and there are not many extreme values. For example, if a set of test scores has a low SD, it means that most of the students scored close to the average.**
* **High SD: A high SD indicates that the data points are spread out over a wider range of values. This means that the data is not as consistent, and there are several extreme values. For example, if a set of test scores has a high SD, it means that some students scored very well, while others scored very poorly.**

**It is important to note that the SD is a relative measure. What is considered a low SD for one set of data may be considered a high SD for another set of data. For example, a set of test scores with an SD of 10 may be considered to have a low SD, while a set of income levels with an SD of 10 may be considered to have a high SD.**

**Also, it is important to note that Standard deviation and Variance are both sensitive to outliers, and may not be appropriate for datasets with extreme values.**

**Other measures of spread such as the Mean Absolute Deviation and Interquartile Range may be more appropriate.**

### **Mean Absolute Deviation (MAD)**

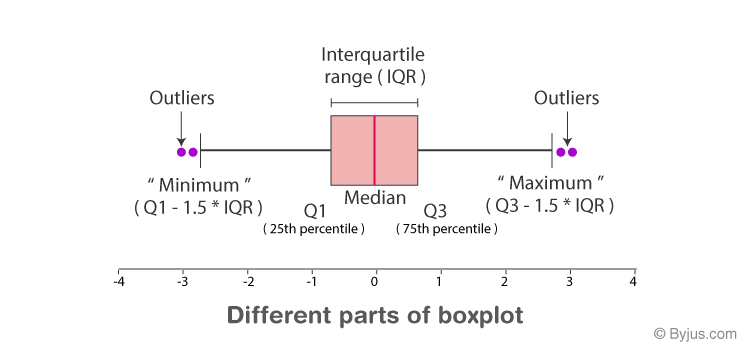
The MAD is another measure of dispersion that is less influenced by outliers than the standard deviation. It calculates the average absolute difference between each data point and the mean of the dataset.

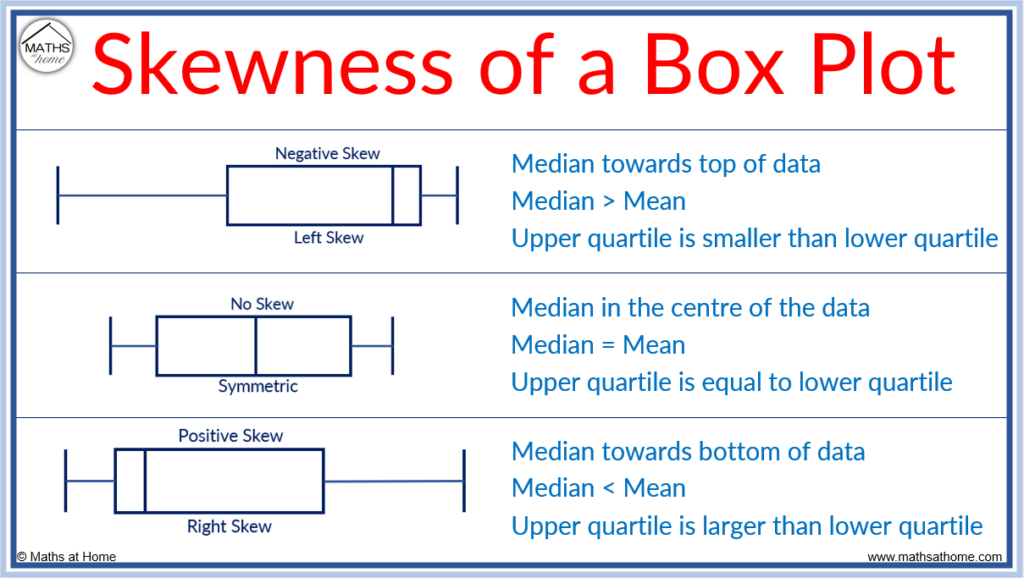
The MAD provides a robust measure of dispersion by considering the magnitude of the differences between data points and the mean, rather than squaring them as in the standard deviation.

## Interquartile distance | Box plot

The interquartile range (IQR) measures the spread of the middle half of your data.

Inter Quartile Range is the measure of distance between 1st and 3rd Quartiles i.e., P25 and P75.





Note: Values of data below Q1 – 1.5 \* IQR and above Q3 + 1.5 \* IQR are classified as **outliers**.

Box plot tells us the following about data:

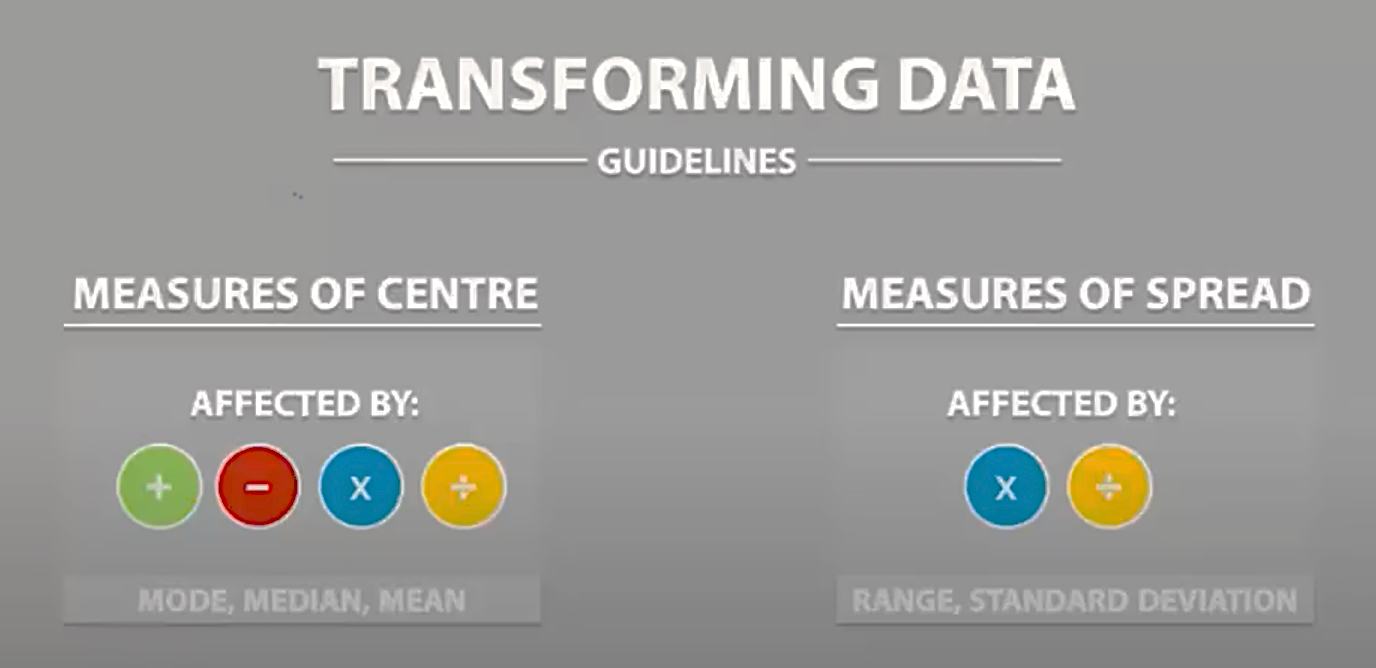
1. Central tendency (median)
2. Spread (IQR3 – IQR1)
3. Symmetricity (Skewness)

## Percentiles – WIP

Work in progress…

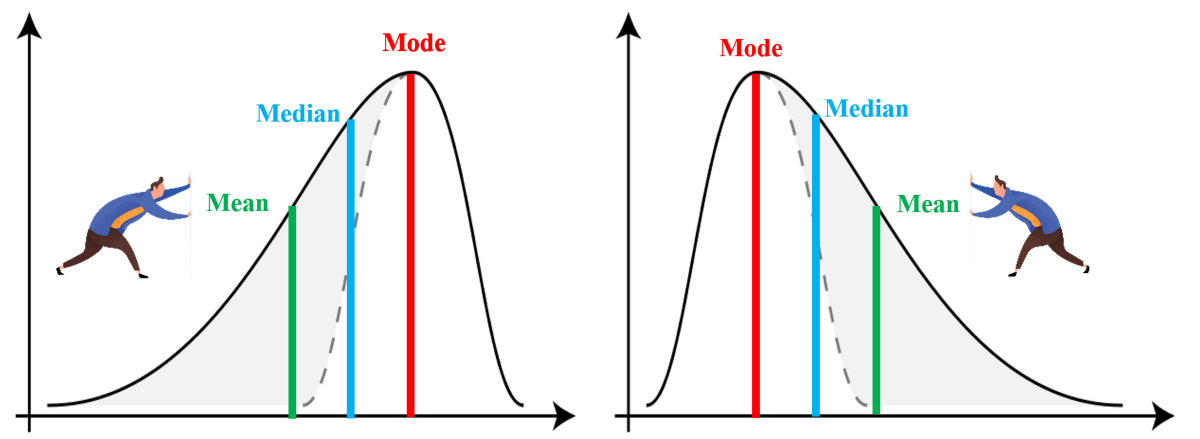
# Effect of transformation on central tendency and spread – WIP

Work in progress…



# Measure of symmetricity

## Skewness

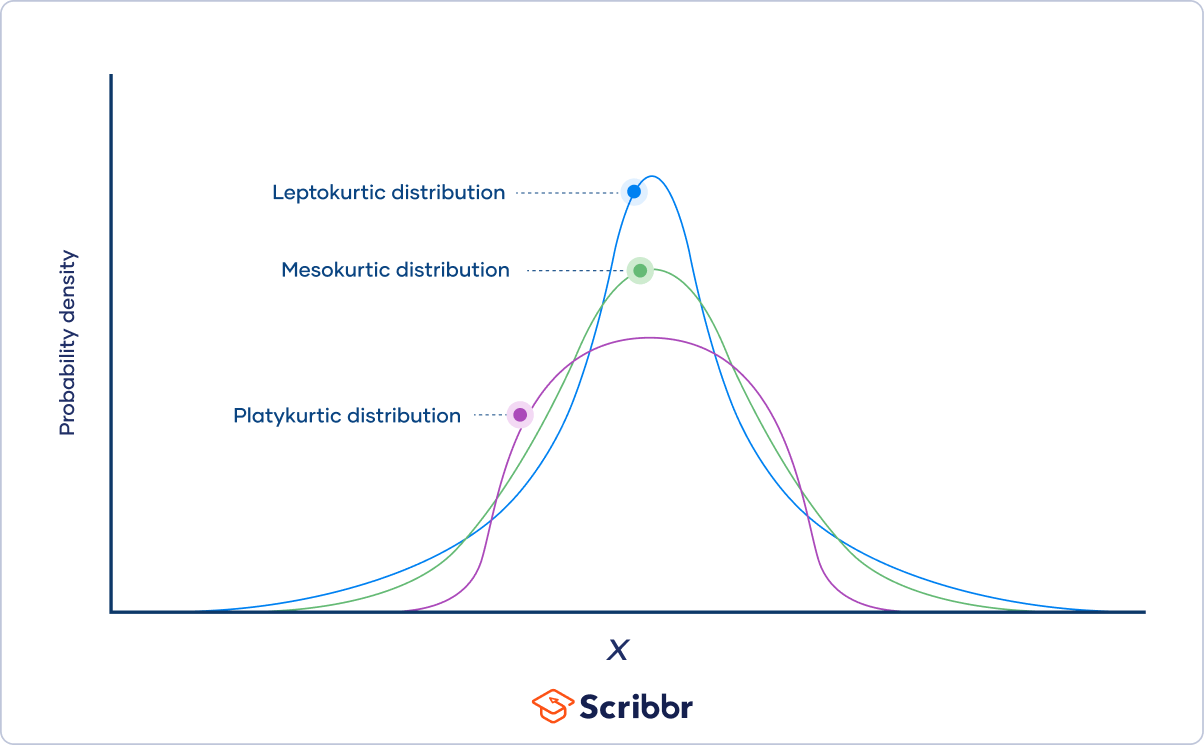


**Skewness is the measure of symmetry or lack of symmetry. Measure of skewness can be used to identify whether the distribution is left skewed (long tail on left side of distribution) or right skewed (long tail on right side of distribution).**

**Pearson’s moment is calculated as below:**

**So, when is the skewness too much? The rule of thumb seems to be:**

## Kurtosis



**Kurtosis is another measure of shape, aimed at shape of the tail, that is, whether the tail of the distribution is heavy or light, what it means is that how much of the data is in the tails and the center.**

**In simple words, kurtosis tells us how much of the distribution is in the tails and how much is in the center.**

**Kurtosis is calculated as below:**

**The range of kurtosis with this formula lies between but it is more common to see kurtosis values between .**

**Often, kurtosis is described in terms of excess kurtosis, which is:**

**Since normal distributions have a kurtosis of 3, excess kurtosis makes comparing a distribution’s kurtosis to a normal distribution even easier.**

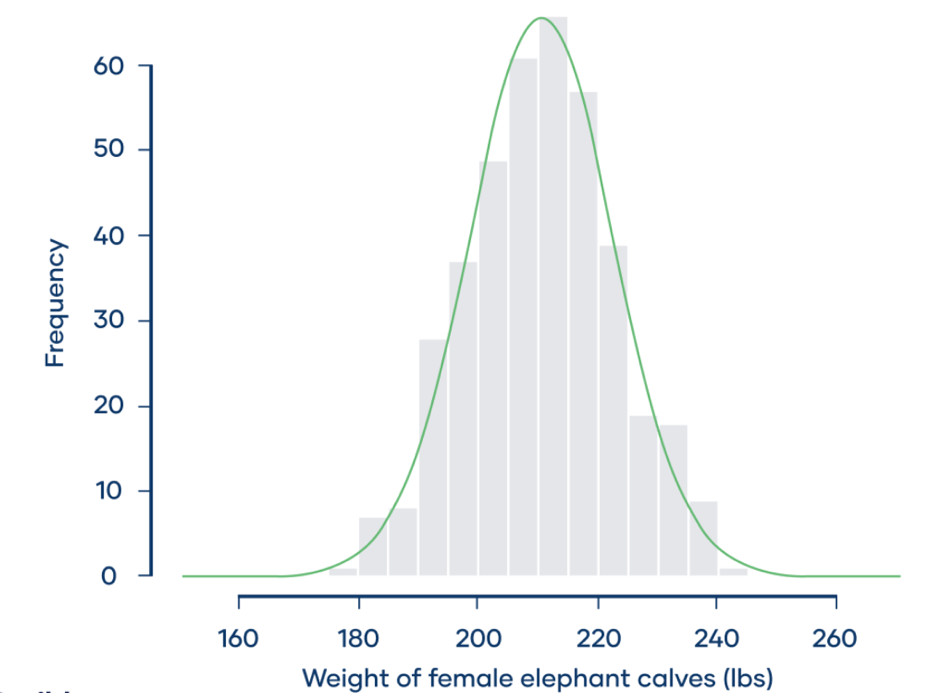
**Interpretation:**

## **Types of kurtosis**

|  |  |  |  |
| --- | --- | --- | --- |
| **Properties** | **Types of Kurtosis** | | |
| **Mesokurtic** | **Platokurtic** | **Lepokurtic** |
| **Tailedness** | Medium-tailed | Thin-tailed | Fat-tailed |
| **Outlier frequency** | Medium | Low | High |
| **Kurtosis** | Moderate (=3) | Low (<3) | High (>3) |
| **Excess kurtosis** | 0 | Negative (-ve) | Positive (+ve) |
| **Example distribution** | Normal | Uniform | Laplace |

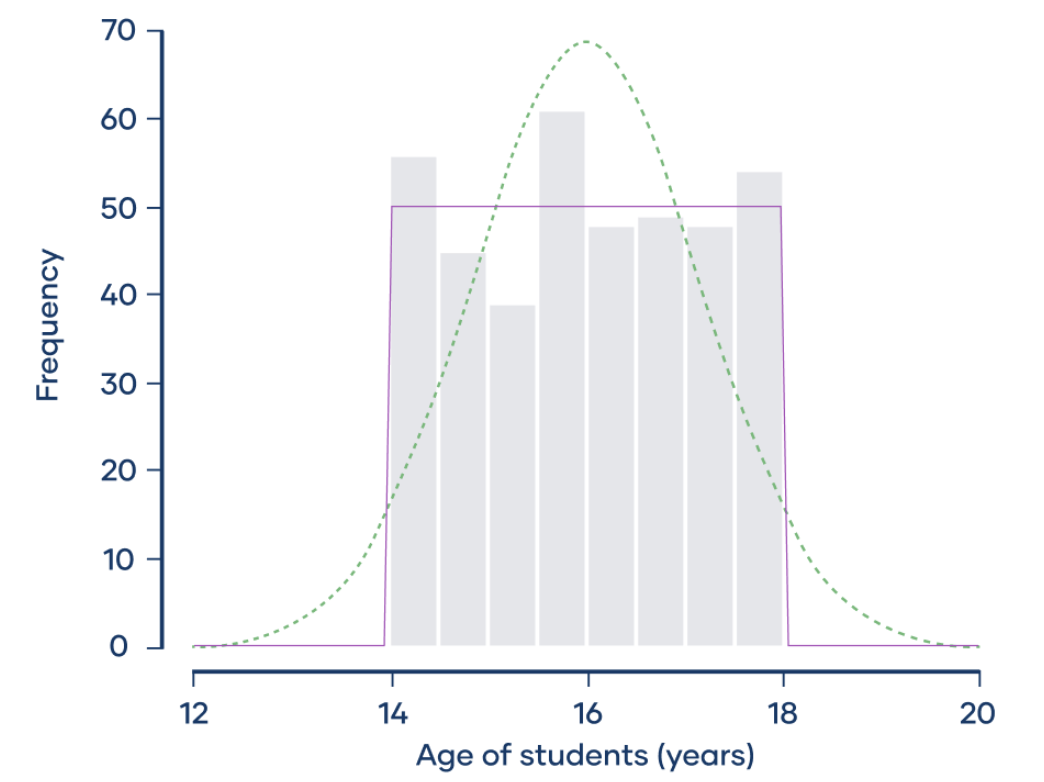
**Kurtosis is measured in comparison to normal distributions.**

## **Mesokurtic distribution**



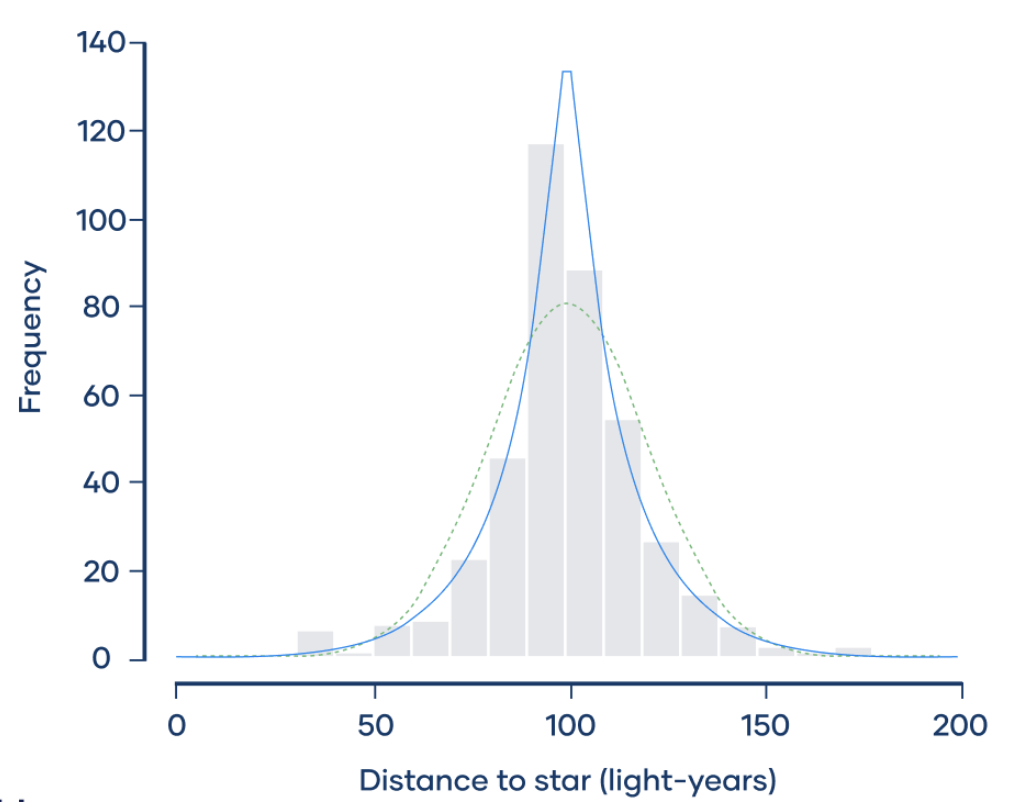
* A mesokurtic distribution is medium-tailed, so outliers are neither highly frequent, nor highly infrequent.
* Normal distributions have a kurtosis of 3, so any distribution with a kurtosis of approximately 3 is mesokurtic.
* Normal distributions have an excess kurtosis of 0, so any distribution with an excess kurtosis of approximately 0 is mesokurtic.
* **Example**: On average, a female baby elephant weighs an impressive 210 lbs at birth. Suppose that a zoologist is interested in the distribution of elephant birth weights, so she contacts zoos and sanctuaries around the world and asks them to share their data. She collects birth weight data for 400 female baby elephants. From the graph, we can see that the frequency distribution (shown by the grey bars) approximately follows a normal distribution (shown by the green curve). Normal distributions are mesokurtic. The zoologist calculates the kurtosis of the sample. She finds that the kurtosis is 3.09 and the excess kurtosis is 0.09, and she concludes that the distribution is mesokurtic. Mesokurtic distributions have outliers that are neither highly frequent, nor highly infrequent, and this is true of the elephant birth weights. Occasionally, a female baby elephant will be born weighing less than 180 or more than 240 lbs.

## **Platokurtic distribution**



* A platokurtic distribution is thin-tailed, meaning that outliers are infrequent.
* Platokurtic distributions have less kurtosis than normal distribution. In other words, platokurtic distributions have:
  + A kurtosis of less than 3
  + An excess kurtosis of less than 0
* Platokurtosis is sometimes called negative kurtosis, since the excess kurtosis is negative.
* Statisticians now understand that kurtosis is a measure of tailedness, not “peakedness.”
* **Example**: A sociologist is studying the social media use of students at a small high school. There are 400 students at the school, ranging in age from 14 to 18 years old. The frequency distribution (shown by the gray bars) doesn’t follow a normal distribution (shown by the dotted green curve). Instead, it approximately follows a uniform distribution (shown by the purple curve). Uniform distributions are platokurtic. The sociologist calculates that the kurtosis of the sample is 1.78 and its excess kurtosis is −1.22. He concludes that the distribution is platokurtic. Platokurtic distributions have a low frequency of outliers. Uniform distributions, like the distribution of students’ ages, are the extreme cases of platokurtic distributions because outliers are so rare that they’re completely absent. There are no students younger than 14 or older than 18 years.

## **Leptokurtic distribution**



* A leptokurtic distribution is fat-tailed, meaning that there are a lot of outliers.
* Leptokurtic distributions are more kurtotic than a normal distribution. They have:
  + A kurtosis of more than 3
  + An excess kurtosis of more than 0
* Leptokurtosis is sometimes called positive kurtosis, since the excess kurtosis is positive.
* **Example**: Imagine that four astronomers are all trying to measure the distance between the Earth and Nu2 Draconis A, a blue star that’s part of the Draco constellation. Each of the four astronomers measures the distance 100 times, and they put their data together in the same dataset. The frequency distribution (shown by the gray bars) doesn’t follow a normal distribution (shown by the dotted green curve). Instead, it approximately follows a Laplace distribution (shown by the blue curve). Laplace distributions are leptokurtic. The astronomers calculate that the kurtosis of the sample is 6.54 and its excess kurtosis is 3.54. They conclude that the distribution is leptokurtic. Leptokurtic distributions have frequent outliers. The distribution of the astronomers’ measurements has more outliers than you would expect if the distribution were normal, with several extreme observations that are less than 50 or more than 150 light-years.

## **Note**

**It's worth noting that kurtosis is affected by outliers, so it is important to have a good understanding of the data before interpreting kurtosis values.**

# Measure of relationship b/w two variables

If we wish to measure both direction and strength of the relationship between X and Y, there are two measures developed for the same.

## Covariance

The covariance of two variables tells us the strength and direction of the relation between those two variables.

Unfortunately, does not tell us much about the strength of such a relationship because it is affected by changes in the units of measurement.

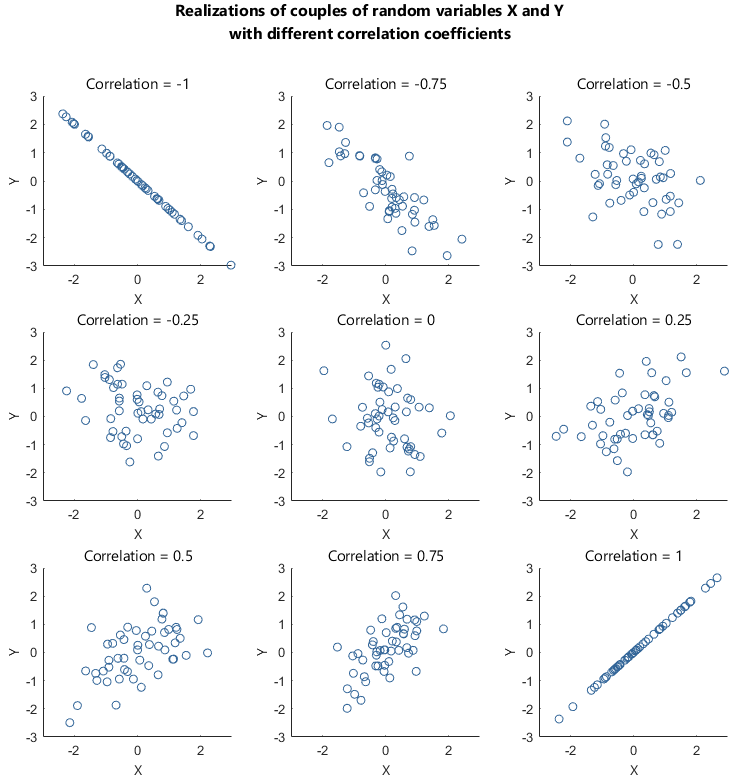
To avoid this disadvantage of the covariance, we *standardize* the data before computing the covariance.

## Correlation Coefficient | Pearson’s R

Pearson’s r measures the linear relationship between two variables. It is commonly used for continuous variables.

The covariance between **standardized** X and Y data is known as the **correlation coefficient** or **Pearson’s R**.

It ranges from -1 to 1 and is given by:



**Note:**  does not necessarily mean that Y and X are not related. It only implies that they are not linearly related because *the correlation coefficient measures only linear relationships.*

## Spearman rank correlation coefficient

Spearman’s rank correlation measures the strength and direction of association between two ranked variables. It basically gives the measure of monotonicity of the relation between two variables i.e. how well the relationship between two variables could be represented using a monotonic function.

Spearman correlation is used for **ordinal or continuous and non-linear variables** and is a non-parametric test, so it does not require the data to be normally distributed, it's also **robust to outliers**.

## Kendall rank correlation coefficient

Kendall rank correlation coefficient measures the ordinal association between two variables. It is commonly used for ordinal or categorical variables.

## Point-biserial correlation coefficient

Point-biserial correlation coefficient measures the correlation between a binary variable and a continuous variable.

## Phi coefficient

Phi coefficient measures the correlation between two binary variables.

## Cramer's V

Cramer's V measures the correlation between two categorical variables.

## Note

It's important to choose the appropriate correlation coefficient based on the type of data you are working with. For example, if you have two continuous variables, you would likely use Pearson correlation, while if you have two ordinal variables, you might use Kendall or Spearman rank correlation.

# Measure of relationship b/w more than 2 variables [WIP]

## Multiple Linear Regression (MLR)

It is a statistical method used to model the relationship between a single response variable and multiple predictor variables. It assumes that the relationship between the predictor and response variables is linear. It uses a linear equation to predict the value of the response variable based on the values of the predictor variables. The coefficients of the predictor variables in the equation indicate the strength and direction of the linear relationship between the predictor and response variables.

## Multivariate Linear Regression (MVLR)

It's similar to MLR but it has multiple response variables. It's used to model the relationship between multiple response variables and multiple predictor variables. The coefficients of the predictor variables in the equation indicate the strength and direction of the linear relationship between the predictor and response variables.

## Canonical Correlation Analysis (CCA)

It's a statistical technique used to explore the relationship between two sets of variables. It finds the linear combination of the variables in each set that have the highest correlation with each other. CCA is useful when we want to find the relationship between two sets of variables and the individual variables in each set may not be directly related. It's a multivariate technique that can be used when there are more than two variables in each set.

## Partial Least Squares Regression (PLSR)

It is a multivariate statistical method that extracts the latent variables that maximize the covariance between the predictors and the response. PLSR is useful when the number of predictors is large and there is multicollinearity among the predictors. It is a linear method that can handle nonlinear relationships between the predictors and response by transforming the data.

## Principal Component Regression (PCR)

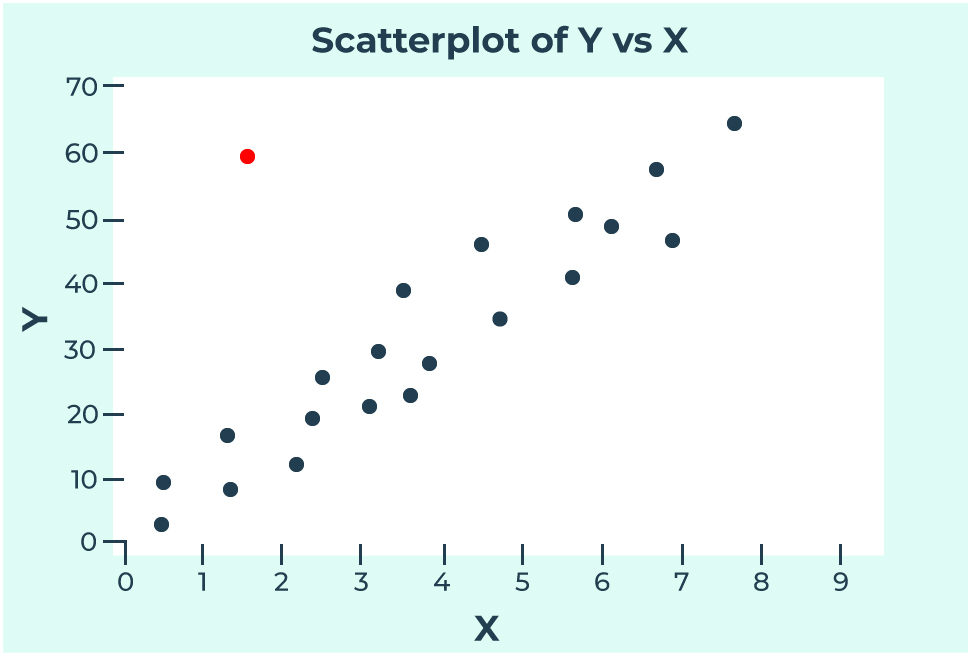
It's a multivariate statistical method that first performs a principal component analysis (PCA) on the predictors and then regresses the response variable on the principal components. PCR is useful when the number of predictors is large and there is multicollinearity among the predictors.

## Note

It's important to note that these methods are used to model the relationships between variables, but they can't prove causality or establish the direction of the relationships. It's also important to keep in mind that these methods assume that the relationships between the variables are linear, but in reality, the relationships between the variables might be non-linear. So, it's important to use the appropriate method and to interpret the result with caution.

# Outlier

An outlier is an observation that is significantly different from the other observations in a dataset.



## Causes of outlier

An outlier can be caused by a variety of factors:

* Measurement error
* Data entry error
* A genuinely unusual or unexpected event

## Methods to identify outlier

Outliers can be identified visually and statistically and are mentioned below:

* Statistical method
  + **z-score**: This method calculates the number of standard deviations a data point is from the mean. Data points with a large positive or negative z-score are considered outliers.
  + **IQR**: This method calculates the range between the 75th and 25th percentiles of the data. Data points that fall outside of the range of 1.5 times the IQR are considered outliers.
  + **Algorithms**: DBSCAN, Local outlier factor (LOF), Isolation Forest, One-class SVM etc.
* Graphical methods
  + **Box plot**: Also known as a box-and-whisker plot, this method shows the distribution of the data by plotting the median, quartiles, and minimum and maximum values. Data points outside of the "whiskers" of the plot are considered outliers.
  + **Scatter plot**: This method plots the data points on a graph with two variables on the x and y axes. Outliers are typically data points that fall far away from the main cluster of points.
  + **Histogram**: This method shows the distribution of the data by plotting the frequency of the data on the y-axis and the data values on the x-axis. Outliers are typically data points that fall outside of the main range of the histogram.
  + **Dot plot**: Similar to a scatter plot, this method plots each data point of one variable on the graph. Outliers are typically data points that fall far away from the main cluster of points.
  + **Violin plot**: This plot combines the box plot and the kernel density estimation, showing the distribution of the data. Outliers are typically data points that fall outside of the main range of the plot.
  + **Heatmap**: A graphical representation of data where individual values are represented as colours. Outliers can be identified by looking for cells with unusual or extreme colour values.
  + **Time-series plot**: This plot shows how a variable changes over time. Outliers can be identified by looking for data points that fall far away from the main trend or pattern of the data.
  + **Bee Swarm plot**: This plot is a variation of a scatter plot, where the observations are spread along the vertical axis, and the jittering is used to avoid overlapping of the observations. Outliers can be identified by looking for data points that fall far away from the main cluster of points.

## Note

Outliers can significantly impact the results of statistical analyses, such as affecting the mean, median, and standard deviation of a dataset. Therefore, it is important to identify and handle outliers appropriately in order to obtain accurate and reliable results.

# Z – score

A Z-score, also known as a standard score, is a measure of how many **standard deviations** a data point is **away** from the **mean** of a dataset. It is a way to standardize a variable and express its value in terms of standard deviations from the mean.

The formula for calculating a Z-score is:

where X is the value of the data point, μ is the mean of the dataset, and σ is the standard deviation of the dataset.

* **Positive Z-score** indicates that the data point is above the mean.
* **Negative Z-score** indicates that the data point is below the mean.
* **0 Z-score** means that the data point is exactly at the mean.

Z-scores can also be used to compare values from different datasets or distributions.

Z-scores are commonly used in statistics and data analysis to identify outliers and to standardize variables for comparison or analysis.

For example, if a Z-score is greater than 3 or less than -3, it indicates that the data point is an outlier, and it is more than 3 standard deviations away from the mean.

# Type of distributions

## Normal Distribution

# Linear and non-linear variables or function or equation

## Linear variable or function

A linear variable is one in which the relationship between the variable and the predictor variable(s) is linear, meaning it can be represented by a **straight line**.

For example, the relationship between the variable "price" and the predictor variable "square footage" in a real estate dataset is a linear relationship, because as the square footage of a property increases, the price of the property will also increase at a **constant rate**.

## Non-linear variable or function

A non-linear variable is one in which the relationship between the variable and the predictor variable(s) is not linear. It can be represented by a **curved or non-linear line**.

Examples:

* The relationship between height and age is non-linear because as age increases, height increases at a decreasing rate. This means that as children get older, they grow taller, but the rate at which they grow taller decreases as they get closer to reaching their full height.
* The relationship between sales and advertising budget is non-linear because as the advertising budget increases, sales may increase at a decreasing rate, reaching a point of diminishing returns where additional increases in advertising budget result in little or no additional sales. This means that as the advertising budget increases, the increase in sales will also increase, but the rate at which the sales increase will become less and less, and after a certain point, increasing the budget will not increase the sales significantly.

## Note

Generally speaking, the rate of increase and decrease is constant in linear functions and it is not constant in non-linear functions.

The correlation coefficient and regression analysis are good for linear relationships, but not for non-linear relationships. In those cases, other methods such as polynomial regression, decision trees, random forest, and neural networks should be used.

# Parametric and non-parametric [WIP]

Parametric methods are based on the assumption that the data follows a specific probability distribution, such as the normal distribution. These methods make use of the parameters of the distribution (such as the mean and standard deviation) to make inferences about the population from which the sample was drawn. Examples of parametric methods include t-tests, ANOVA, and linear regression.

Non-parametric methods, on the other hand, do not make assumptions about the underlying probability distribution of the data. These methods are based on the ranks or the ordering of the data rather than the actual values. Examples of non-parametric methods include the Wilcoxon rank-sum test, the Kruskal-Wallis test, and the Spearman rank correlation coefficient.

One of the main advantages of non-parametric methods is that they are more robust to violations of assumptions about the underlying distribution of the data. They are also less sensitive to outliers and can be used with ordinal, categorical, and continuous data. However, non-parametric methods have lower power and may not be as efficient as parametric methods when the assumptions of the parametric methods are met.

It's important to note that the choice of parametric or non-parametric methods depends on the assumptions made about the underlying probability distribution of the data and the research question. It's always a good idea to check the assumptions before selecting a method and interpreting the results.

<https://www.statisticshowto.com/probability-and-statistics/statistics-definitions/parametric-and-non-parametric-data/>

<https://www.scribbr.com/statistics/statistical-tests/>

# Hypothesis Testing

Hypothesis testing is a statistical method used to make inferences about a population based on a sample of data from that population. The objective of hypothesis testing is to determine whether a certain claim about the population (also known as the null hypothesis) can be rejected or not based on the sample data.

In hypothesis testing, one starts with two opposing hypotheses: the null hypothesis, which states that there is no significant difference or relationship between variables, and the alternative hypothesis, which states that there is a significant difference or relationship between variables. The sample data is then collected and analysed to determine whether the null hypothesis can be rejected or not.

The most common approach to hypothesis testing involves calculating a test statistic based on the sample data and comparing it to a critical value determined by a statistical distribution (such as the normal or t-distribution). If the test statistic falls outside of the critical value, the null hypothesis is rejected, and the alternative hypothesis is accepted.

Hypothesis testing is widely used in various fields, such as science, medicine, psychology, economics, and engineering, to make informed decisions based on data.

Hypothesis testing is a way to test if a claim about a group of things is true or not. The claim is called the "null hypothesis" and it says that there is no difference or relationship between things being studied. To test this claim, we collect some information or data and analyse it. Based on this analysis, we decide if we have enough evidence to reject the null hypothesis and say that the claim is not true.

For example, let's say a doctor claims that a new medicine can cure a certain illness. To test this claim, they might conduct a study with a group of patients and compare the results of those taking the medicine to those who don't. If the results show that the patients taking the medicine are getting better at a higher rate than those who don't, the doctor might reject the null hypothesis and say that the medicine is effective in curing the illness.

In simple terms, hypothesis testing helps us make decisions about a claim based on data.

## Step-by-step guide to learn Hypothesis Testing

1. Start with the basics: Understand the concept of hypothesis testing, the null and alternative hypotheses, and the idea of accepting or rejecting the null hypothesis based on the data.
2. Types of hypothesis tests: Familiarize yourself with different types of hypothesis tests, such as one-sample, two-sample, and chi-squared tests.
3. Test statistic and critical values: Learn about the calculation of test statistics and critical values, including the z-test, t-test, and F-test, and how to use them to make decisions about the null hypothesis.
4. Significance level: Understand the concept of significance level (α) and its role in hypothesis testing, including how it is used to determine the critical value and the consequences of changing the significance level.
5. P-value: Learn about the p-value and how it is used to evaluate the evidence against the null hypothesis.
6. Confidence intervals: Study the relationship between hypothesis testing and confidence intervals and how they can be used to make inferences about population parameters.
7. Power of a test: Learn about the power of a test, which is the probability of rejecting the null hypothesis when the alternative hypothesis is true.
8. Type I and Type II errors: Study the concepts of Type I and Type II errors and how they can impact the results of a hypothesis test.
9. Practice: Practice applying hypothesis tests to real-world examples and data sets to gain hands-on experience and solidify your understanding of the concepts.

It's also important to keep in mind that hypothesis testing is just one aspect of statistical inference and that it should be used in conjunction with other techniques, such as regression analysis, to make informed decisions about populations based on data.