# Unit -2
# PROBABILITY AND STATISTICS

Introduction to probability and statistics, Population and sample, Normal and Gaussian distributions, Probability Density Function, Descriptive statistics, notion of probability, distributions, mean, variance, covariance, covariance matrix, understanding univariate and multivariate normal distributions, introduction to hypothesis testing, confidence interval for estimates.

## 1.1.What is Data?

Data is the information collected through different sources which can be qualitative or quantitative in nature. Mostly, the data collected is used to analyse and draw insights on a particular topic.

**Types of Data.**

Numerical Data

Numerical data is the information in numbers i.e., numeric which poses as a quantitative measurement of things.

**For example:**

1. Heights and weights of people
2. Stock Prices

**a) Discrete Data**

Discrete data is the information that often counts of some event i.e., can only take specific values. These are often integer-based, but not necessarily.

**For example:**

1. Number of times a coin was flipped
2. Shoe sizes of people

**b) Continuous Data**

Continuous Data is the information that has the possibility of having infinite values i.e., can take any value within a range.

**For example:**

How many centimeters of rain fell on a given day?

Categorical Data

This type of data is qualitative in nature which has no inherent mathematical significance. It is sort of a fixed value under which a unit of observation is assigned or “categorized”.

**For example:**

1. Gender

2. Binary Data (Yes/No)
3. Attributes of a vehicle like color, mileage, number of doors, etc.

## 1.2.What are Statistics?

The field of **Statistics** deals with the collection, presentation, analysis, and use of data to make decisions, solve problems, and design products and processes. **Statistics** is the science of learning from data, and of measuring, controlling, and communicating uncertainty; and it thereby provides the navigation essential for controlling the course of scientific and societal advances. In simple terms, statistics is the science of data. **Statistics is defined as collection, compilation, analysis and interpretation of numerical data.**

### 1.2.1. Statistics is the science of data

The most important aspect of any Data Science approach is how the information is processed. When we talk about developing insights out of data it is basically digging out the possibilities. Those possibilities in Data Science are known as **Statistical Analysis**. Most of us wonder how can data in the form of text, images, videos, and other highly unstructured formats get easily processed by Machine Learning models. But the truth is we actually convert that data into a numerical form which is not exactly our data but the numerical equivalent of it. So, this brings us to the very important aspect of Data Science. With data in numerical format, it provides us with infinite possibilities to understand the information out it. Statistics acts as a pathway to understand your data and process that for successful results. Not only the power of statistics is limited to understanding the data it also provides methods to measure the success of our insights, getting different approaches for the same problem, getting the right mathematical approach for your data.

- In an agricultural study, researchers want to know which of four fertilizers (which vary in their nitrogen contents) produces the highest corn yield. In a clinical trial, physicians want to determine which of two drugs is more effective for treating HIV in the early stages of the disease. In a public health study, epidemiologists want to know whether smoking is linked to a particular demographic class in high school students.
- To develop an appreciation for variability and how it effects product, process and system.
- It is estimating the present; predicting the future
- Study methods that can be used to solve problems, build knowledge.
- Statistics make data into information
- Develop an understanding of some basic ideas of statistical reliability, stochastic process (probability concepts).
- Statistics is very important in every aspect of society (Govt., People or Business)

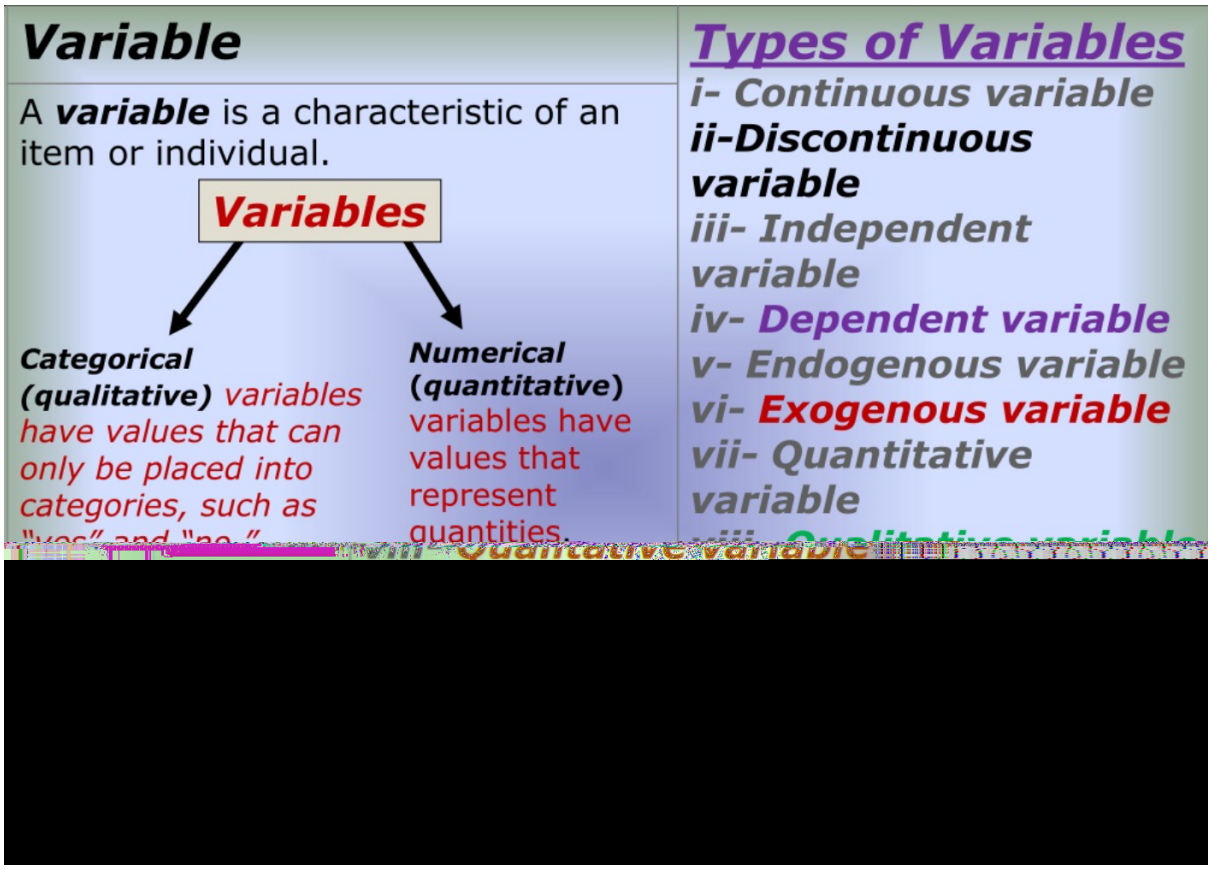
### 1.2.2. Basic terms

**Variable:** Property with respect to which data from a sample differ in some measurable way
**Measurement:** assignment of numbers to something
**Data:** collection of measurements
**Population:** all possible data

**Sample:** collected data

## 1. Variable

A **variable** is a characteristic or condition that can change or take on different values. Most research begins with a general question about the relationship between two variables for a specific group of individuals.




**Types of Variables**

Variables can be classified as discrete or continuous. **Discrete variables** (such as class size) consist of indivisible categories, and **continuous variables** (such as time or weight) are infinitely divisible into whatever units a researcher may choose. For example, time can be measured to the nearest minute, second, half-second, etc.

## 2. Measuring Variables

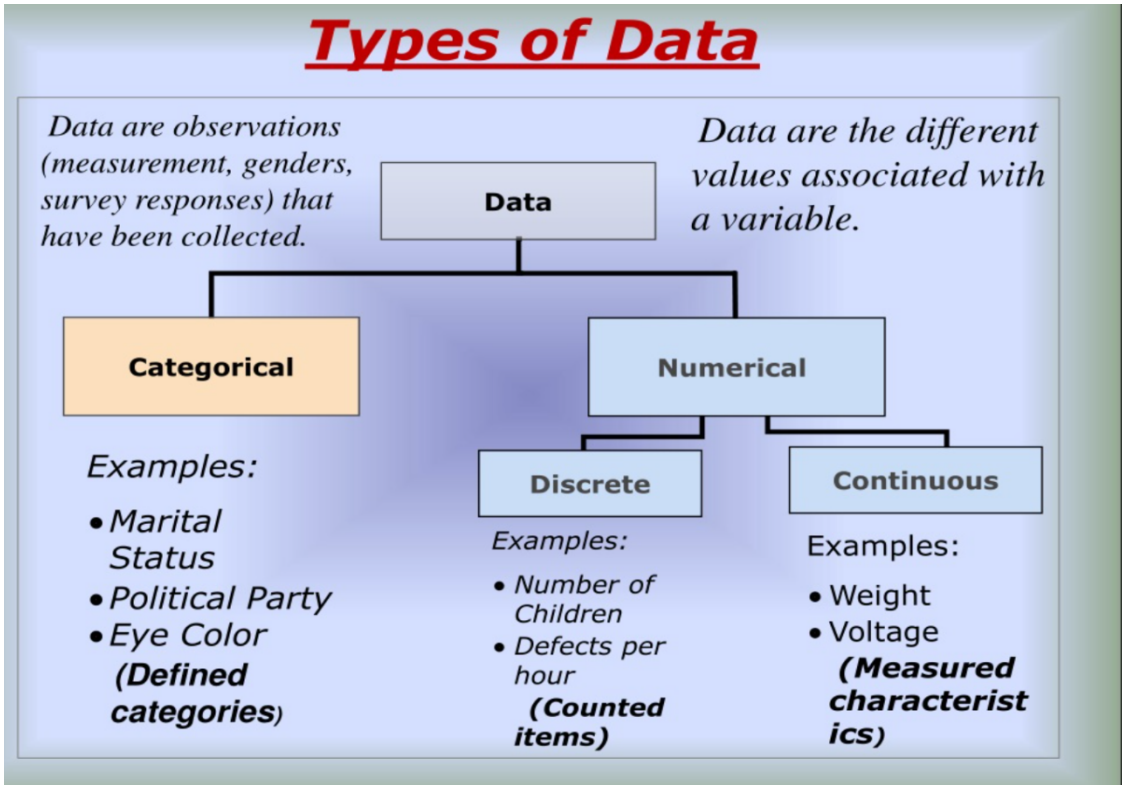
To establish relationships between variables, researchers must observe the variables and record their observations. This requires that the variables be **measured**. The process of measuring a variable requires a set of categories called a **scale of measurement** and a process that classifies each individual into one category.

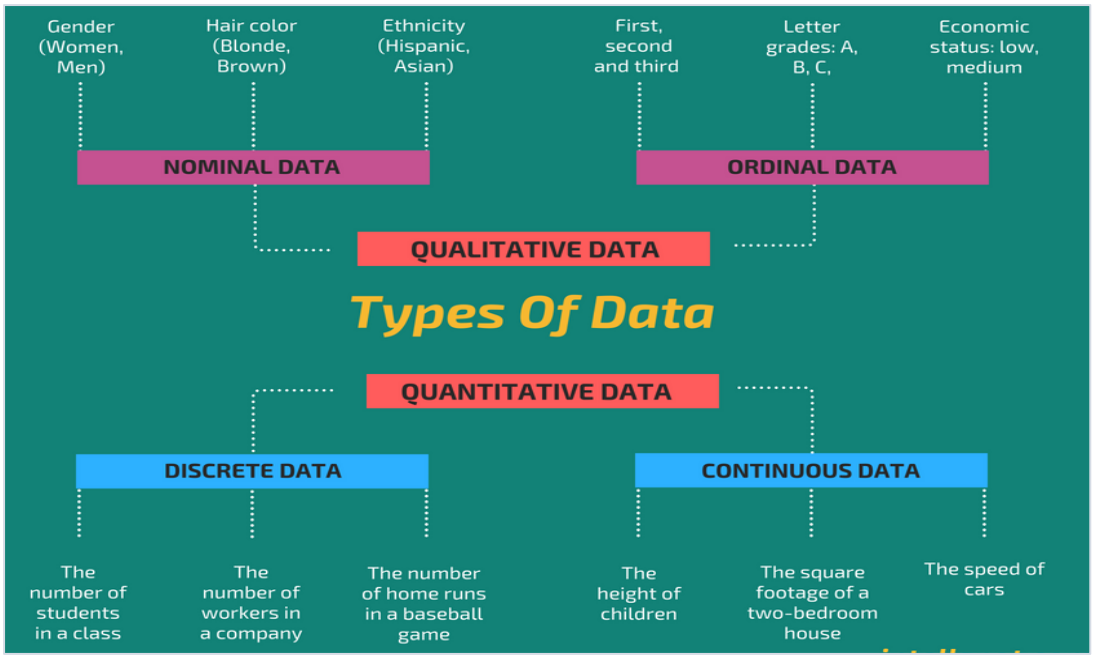
**4 Types of Measurement Scales**

A **nominal scale** is an unordered set of categories identified only by name. Nominal measurements only permit you to determine whether two individuals are the same or different.
An **ordinal scale** is an ordered set of categories. Ordinal measurements tell you the direction of difference between two individuals.
An **interval scale** is an ordered series of equal-sized categories. Interval measurements identify the direction and magnitude of a difference. The zero point is located arbitrarily on an interval scale.
A **ratio scale** is an interval scale where a value of zero indicates none of the variable. Ratio measurements identify the direction and magnitude of differences and allow ratio comparisons of measurements.

**3. Data**

The measurements obtained in a research study are called the **data**. The goal of statistics is to help researchers organize and interpret the data.






**Quantitative**

The data which are statistical or numerical are known as Quantitative data. Quantitative data is generated through. Quantitative data is also known as Structured data. Experiments, Tests, Surveys, Market Report. Quantitative data is again divided into **Continuous data and Discrete data.**

**Continuous Data**

Continuous data is the data which can have any value. That means Continuous data can give infinite outcomes so it should be grouped before representing on a graph.

**Examples**

- The speed of a vehicle as it passes a checkpoint
- The mass of a cooking apple
- The time taken by a volunteer to perform a task

**Discrete Data**

- Discrete data can have certain values. That means only a finite number can be categorized as discrete data.
- Numbers of cars sold at a dealership during a given month
- Number of houses in certain block
- Number of fish caught on a fishing trip
- Number of complaints received at the office of airline on a given day
- Number of customers who visit at bank during any given hour
- Number of heads obtained in three tosses of a coin

**Differences between Discrete and Continuous data**

- Numerical data could be either discrete or continuous

- Continuous data can take any numerical value (within a range); For example, weight, height, etc.
- There can be an infinite number of possible values in continuous data
- Discrete data can take only certain values by finite 'jumps', i.e., it 'jumps' from one value to another but does not take any intermediate value between them (For example, number of students in the class

**Qualitative**

Data that deals with description or quality instead of numbers are known as Quantitative data. Qualitative data is also known as **unstructured data.** Because this type of data is loosely compact and can't be analyzed conventionally.

4. **Population**

The entire group of individuals is called the **population**. For example, a researcher may be interested in the relation between class size (variable 1) and academic performance (variable 2) for the population of third-grade children.

5. **Sample**

Usually, **populations** are so large that a researcher cannot examine the entire group. Therefore, a **sample** is selected to represent the population in a research study. The goal is to use the results obtained from the sample to help answer questions about the population.

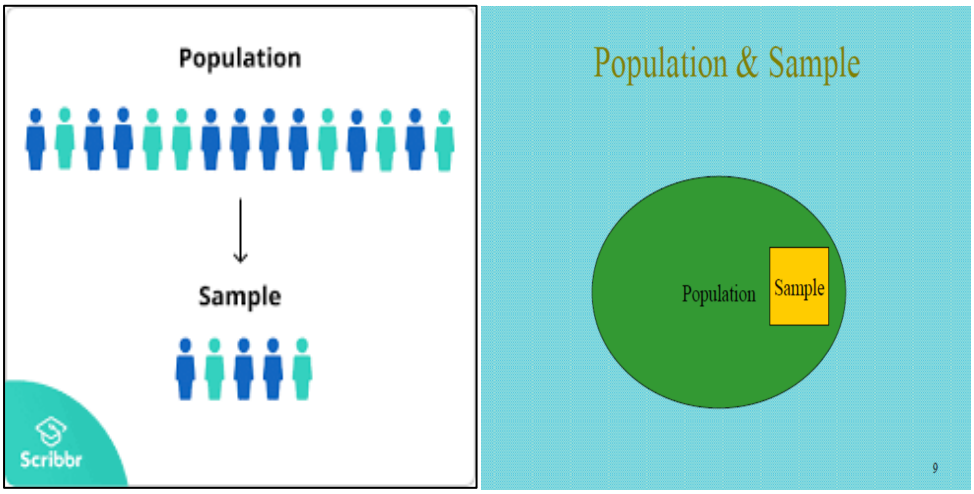
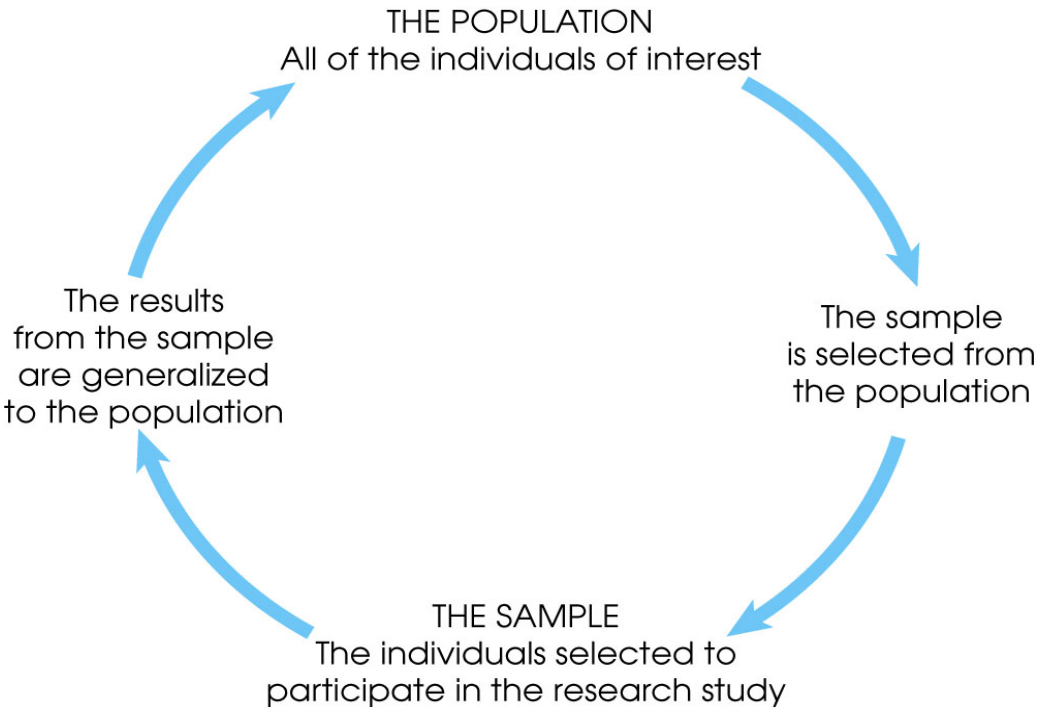



Figure 2.1: Population and Sample

| POPULATION | SAMPLE |
| --- | --- |
| ▪ The measurable quality is called a parameter. | ▪ The measurable quality is called a statistic. |
| ▪ The population is a complete set. | ▪ The sample is a subset of the population. |
| ▪ Reports are a true representation of opinion. | ▪ Reports have a margin of error and confidence interval. |
| ▪ It contains all members of a specified group. | ▪ It is a subset that represents the entire population. |

**Sampling Error**

- The discrepancy between a sample statistic and its population parameter is called **sampling error**.
- Defining and measuring sampling error is a large part of inferential statistics.




## 1.3.Frequency Distribution

**Frequency Distribution (or Frequency Table)**

Shows how a data set is partitioned among all of several categories (or classes) by listing all of the categories along with the number (frequency) of data values in each of them

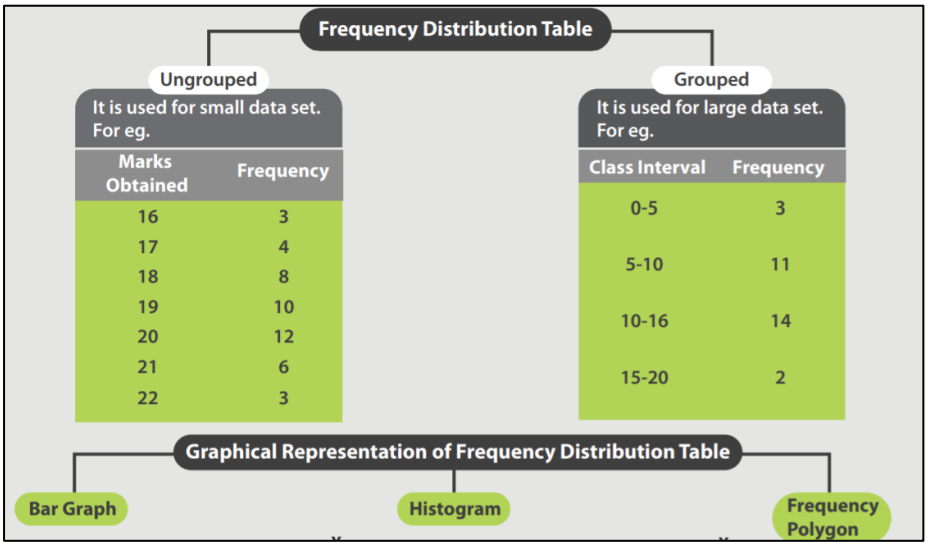
**Frequency Distribution**

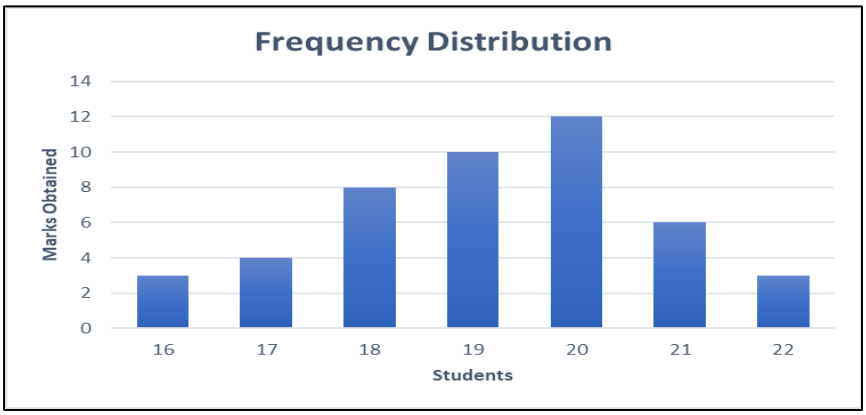
When data are in original form, they are called raw data

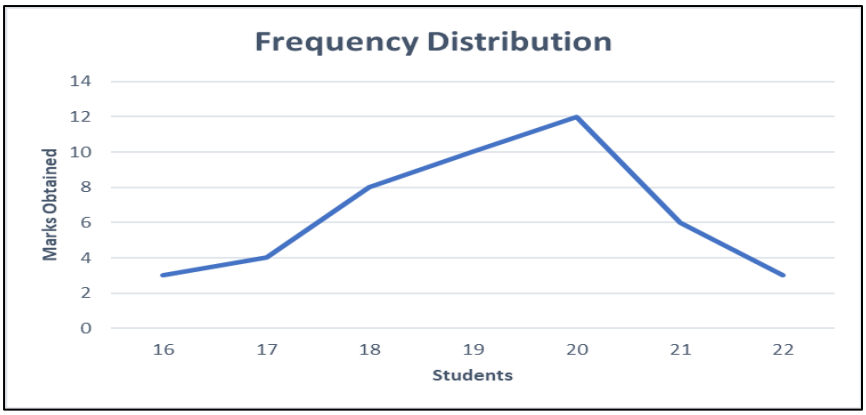
**Organizing Data:**

Categorical distribution

Grouped distribution
Ungrouped distribution



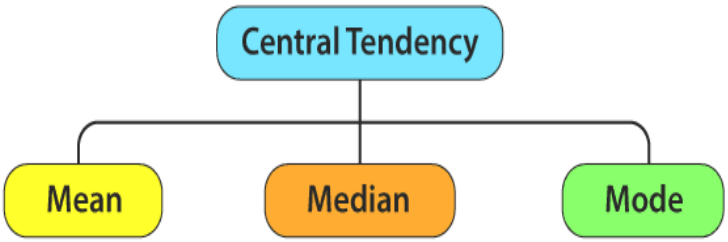




**Frequency distribution** refers to data classified on the basis of some variable that can be measured such as prices, weight, height, wages etc.

| Weights (in kg) | Number of students |
|---|---|
| 31 - 35 | 9 |
| 36 - 40 | 5 |
| 41 - 45 | 14 |
| 46 - 50 | 3 |
| 51 - 55 | 1 |
| 56 - 60 | 2 |
| 61 - 65 | 2 |
| 66 - 70 | 1 |
| 71 - 75 | 1 |
| **Total** | 38 |

**Measures of Centre Tendency**

- In statistics, the **central tendency** is the descriptive summary of a data set.
- Through the single value from the dataset, it reflects the centre of the data distribution.
- Moreover, it does not provide information regarding individual data from the dataset, where it gives a summary of the dataset. Generally, the central tendency of a dataset can be defined using some of the measures in statistics.




**Mean**

- The mean represents the average value of the dataset.
- It can be calculated as the sum of all the values in the dataset divided by the number of values. In general, it is considered as the **arithmetic mean**.
- Some other measures of mean used to find the central tendency are as follows:
- **Geometric Mean** (nth root of the product of n numbers)
- **Harmonic Mean** (the reciprocal of the average of the reciprocals)
- **Weighted Mean** (where some values contribute more than others)
- It is observed that if all the values in the dataset are the same, then all geometric, arithmetic and harmonic mean values are the same. If there is variability in the data, then the mean value differs.

**Calculating the Mean**

Calculate the mean of the following data:

1 5 4 3 2

**Sum the scores** (ΣX):

1 + 5 + 4 + 3 + 2 = 15

**Divide the sum** (ΣX = 15) by the number of scores (N = 5):

15 / 5 = 3

**Mean = X = 3**

**The Median**

- The *median* is simply another name for the 50th percentile
- Sort the data from highest to lowest
- Find the score in the middle
- If N, the number of scores, is even the median is the average of the middle two scores

Median Example

What is the median of the following scores:

10 8 14 15 7 3 3 8 12 10 9

Sort the scores:

15 14 12 10 10 9 8 8 7 3 3

Determine the middle score:

middle = (N + 1) / 2 = (11 + 1) / 2 = 6

Middle score = median = 9

**Median Example**

What is the median of the following scores:

24 18 19 42 16 12

- Sort the scores:

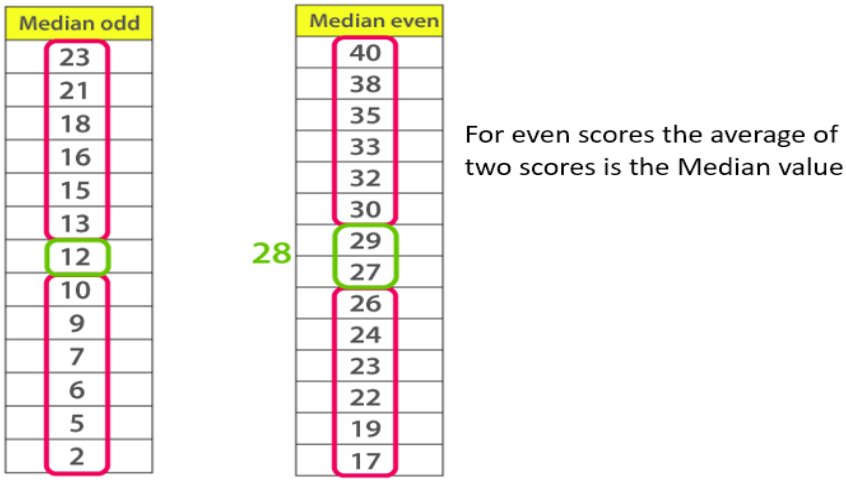
  42 24 19 18 16 12

- Determine the middle score:

  middle = (N + 1) / 2 = (6 + 1) / 2 = 3.5

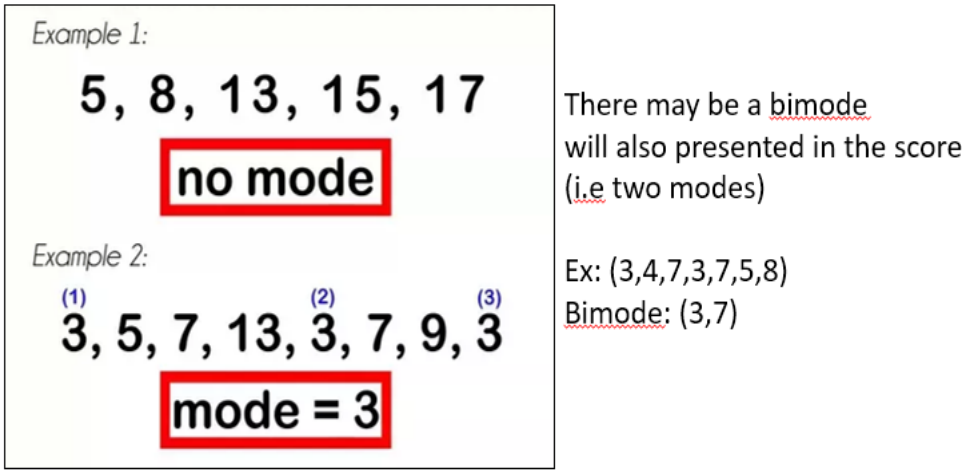
- Median = average of 3rd and 4th scores:

  (19 + 18) / 2 = 18.5




**Mode**

The *mode* is the score that occurs most frequently in a set of data.




**Variance**

- **Variance** is the average squared deviation from the mean of a set of data.
- It is used to find the Standard deviation.
- $\sigma^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$

**Variance**

- This is a good measure of how much variation exists in the sample, normalized by sample size.
- It has the nice property of being additive.
- The only problem is that the variance is measured in units squared

**How to find Variance**

- Find the **Mean** of the data.
- Subtract the mean from each value – the result is called the **deviation from the mean.**

- Square each deviation of the mean.
- Find the sum of the squares.
- Divide the total by the number of items.

**How to find Variance? - Example**

- Suppose you're given the data set 1, 2, 2, 4, 6. (X = 1,2,2,4,6) One Variable X
- Calculate the mean of your data set. The mean of the data is (1+2+2+4+6)/5
- Mean= 15/5 = 3.
- Subtract the mean from each of the data values and list the differences. Subtract 3 from each of the values 1, 2, 2, 4, 6
- 1-3 = -2  2-3 = -1  2-3 = -1  4-3 = 1  6-3 = 3
- Your list of differences is -2, -1, -1, 1, 3 (deviation)
- You need to square each of the numbers -2, -1, -1, 1, 3
  $(-2)^2 = 4$, $(-1)^2 = 1$, $(-1)^2 = 1$, $(1)^2 = 1$, $(3)^2 = 9$
- Your list of squares is 4, 1, 1, 1, 9, Add the squares 4+1+1+1+9 = 16
- Subtract one from the number of data values you started with. You began this process (it may seem like a while ago) with five data values. One less than this is 5-1 = 4.
- Divide the sum from step four by the number from step five. The sum was 16, and the number from the previous step was 4. You divide these two numbers 16/4 = 4.

**Variation in one variable**

- So, these four measures all describe aspects of the variation in a single variable:
- a. Sum of the squared deviations
- b. Variance
- c. Standard deviation
- d. Standard error
- Can we adapt them for thinking about the way in which two variables might vary together?

**Covariance**

- In mathematics and statistics, **covariance** is a measure of the relationship between two random variables. (X, Y)
- More precisely, covariance refers to **the measure of how two random variables in a data set will change together**.
- **Positive covariance**: Indicates that two variables tend to move in the same direction.
- **Negative covariance**: Reveals that two variables tend to move in inverse directions.

- The covariance between two random variables X and Y can be calculated using the following **formula** (for **population**):

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

- For a **sample** covariance, the formula is slightly adjusted:

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

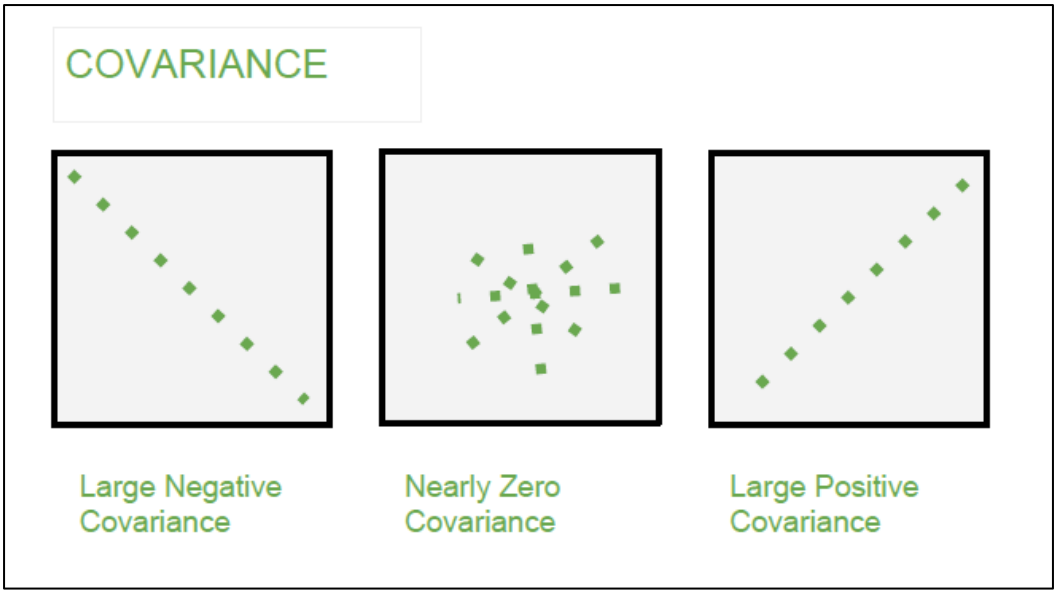
Where:
**$X_i$** – the values of the X-variable
**$Y_j$** – the values of the Y-variable
**$\bar{X}$** – the mean (average) of the X-variable
**$\bar{Y}$** – the mean (average) of the Y-variable
**n** – the number of data points

**Covariance Example**

**Example 1:** Find covariance for following data set (Two Variables X and Y)
X = {2,5,6,8,9}, Y = {4,3,7,5,6}

**Solution:**

- Given data sets X = {2,5,6,8,9}, Y = {4,3,7,5,6} and N = 5
- **Mean(X)** = (2 + 5 + 6 + 8 + 9) / 5 = 30 / 5 = **6**
- **Mean(Y)** = (4 + 3 +7 + 5 + 6) / 5 = 25 / 5 = **5**
- **Sample covariance Cov(X,Y) = $\sum(X_i - X) \times (Y_i - Y)/ (N - 1)$**
- = [(2 - 6)(4 - 5) + (5 - 6)(3 - 5) + (6 - 6)(7 - 5) + (8 - 6)(5 - 5) +
- (9 - 6)(6 - 5)] / 5 - 1
- = 4 + 2 + 0 + 0 + 3 / 4 = 9 / 4 = **2.25**
- **Population covariance Cov(X,Y) = $\sum(X_i - X) \times (Y_i - Y)/ (N)$**
- = [(2 - 6)(4 - 5) + (5 - 6)(3 - 5) + (6 - 6)(7 - 5) + (8 - 6)(5 - 5) + (9 - 6)(6 - 5)] / 5
- = 4 + 2 + 0 + 0 + 3 /5
- = 9 / 5
- = 1.8
- **Answer:** The **sample covariance** is 2.25 and the **population covariance** is 1.8
- **Positive and Negative Covariance**




Covariance Matrix

- The **covariance matrix** is a math concept that occurs in several areas of machine learning.If you have a set of n numeric data items, where each data item has d dimensions, then the covariance matrix is a d-by-d symmetric square matrix where there are variance values on the diagonal and covariance values off the diagonal.
- Suppose you have a set of n=5 data items, representing 5 people, where each data item has a Height (X), test Score (Y), and Age (Z) (therefore d = 3):
- Covariance Matrix

| | X | Y | Z |
|---|---|---|---|
| | Height | Score | Age |
| | 64.0 | 580.0 | 29.0 |
| | 66.0 | 570.0 | 33.0 |
| | 68.0 | 590.0 | 37.0 |
| | 69.0 | 660.0 | 46.0 |
| | 73.0 | 600.0 | 55.0 |
| mean = | 68.0 | 600.0 | 40.0 |
| | n=5 | | |

**Covariance Matrix**

- The covariance matrix for this data set is:

```
    X         Y         Z
X   11.50     50.00    34.75
Y   50.00   1250.00   205.00
Z   34.75    205.00   110.00
```

- The 11.50 is the variance of X, 1250.0 is the variance of Y, and 110.0 is the variance of Z. For variance, in words, subtract each value from the dimension mean. Square, add them up, and divide by n-1. For example, for X:
- Var(X) = [ (64–68.0)^2 + (66–68.0^2 + (68-68.0)^2 + (69-68.0)^2 +(73-68.0)^2 ] / (5-1) = (16.0 + 4.0 + 0.0 + 1.0 + 25.0) / 4 = 46.0 / 4 = 11.50.

**Covariance Matrix**

- Covar(XY) =
- [ (64-68.0)*(580-600.0) + (66-68.0)*(570-600.0) + (68-68.0)*(590-600.0) + (69-68.0)*(660-600.0) + (73-68.0)*(600-600.0) ] / (5-1) =
- [80.0 + 60.0 + 0 + 60.0 + 0] / 4 =
- 200 / 4 = 50.0
- If you examine the calculations carefully, you'll see the pattern to compute the covariance of the XZ and YZ columns. And you'll see that Covar(XY) = Covar(YX).

**Standard Deviation**

- Variability is a term that describes how spread out a distribution of scores (or darts) is.
- **Variance** and **standard deviation** are closely related ways of measuring, or quantifying, variability.
- Standard deviation is simply the square root of variance
- **Find the mean** (or arithmetic average) of the scores. To find the mean, add up the scores and divide by n where n is the number of scores.
- **Find the sum of squared deviations** (abbreviated SSD). To get the SSD, find the sum of the squares of the differences (or deviations) between the mean and all the individual scores.
- **Find the variance**. If you are told that the set of scores constitute a population, divide the SSD by n to find the variance. If instead you are told, or can infer, that the set of scores constitute a sample, divide the SSD by (n – 1) to get the variance.
- **Find the standard deviation**. To get the standard deviation, take the square root of the variance.

**How to find Standard Deviation – Example (in Population score)**

- **Example 1: Find the SSD, variance, and standard deviation** for the following **population of scores**: 1, 2, 3, 4, 5 using the list of steps given above.
- **Find the mean**. The mean of these five numbers (the population mean) is (1+2+3+4+5)/5 = 15/5 = 3.
- Let's use the definitional formula for SSD for its calculation: SSD is the sum of the squares of the differences (squared deviations) between the mean and the individual scores. The squared deviations are $(3-1)^2$, $(3-2)^2$, $(3-3)^2$, $(3-4)^2$, and $(3-5)^2$. That is, 4, 1, 0, 1, and 4. The SSD is then 4 + 1 + 0 + 1 + 4 = 10.
- Divide SSD by n, since this is a population of scores, to **get the variance**. So the variance is 10/5 = 2.
- The standard deviation is the square root of the variance. So the standard deviation is the square root of 2. = $\sqrt{2}$ =1.4142
- For practice, let's also compute the SSD using the computational formula, $\sum_i (x_i)^2 - (1/N)(\sum_i x_i)^2$. $\sum_i (x_i)^2 = 12 + 22 + 32 + 42 + 52 = 1 + 4 + 9 + 16 + 25 = 55$. $(1/N)(\sum_i x_i)^2 = (1/5)(1 + 2 + 3 + 4 + 5)^2 = (1/5)(15^2) = 45$. So SSD = 55 – 45 = 10, just like before.

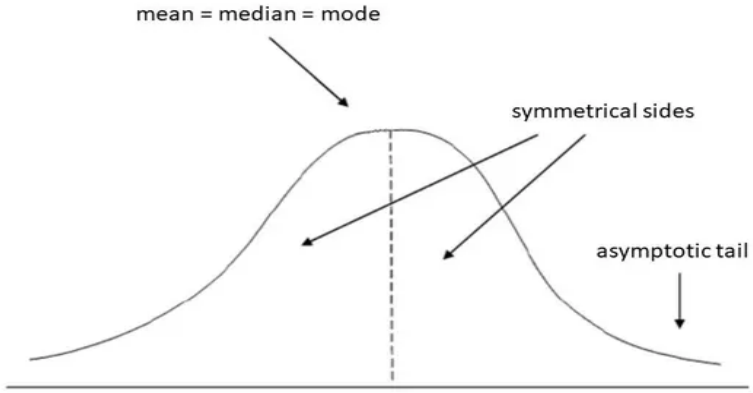
**How to find Standard Deviation – Example (in Sample score)**

- **Example 2:** Find the SSD, variance, and standard deviation for the following **sample of scores**: 1, 3, 3, 5.
- The average of these four numbers (the sample mean) is (1+3+3+5)/4 = 12/4 = 3.
- So, SSD = $(3-1)^2 + (3-3)^2 + (3-3)^2 + (3-5)^2 = 4 + 0 + 0 + 4 = 8$.
- Now, because we were told that these scores constitute a sample, we'll divide SSD by n-1 to get the sample variance.
- In our case we have four scores, so n = 4 so n-1 = 3. Therefore, our sample variance is 8/3.
- And the **sample standard deviation** is square root of 8/3 = $\sqrt{2.6}$ (SQRT 0F 2.6) = 1.6124

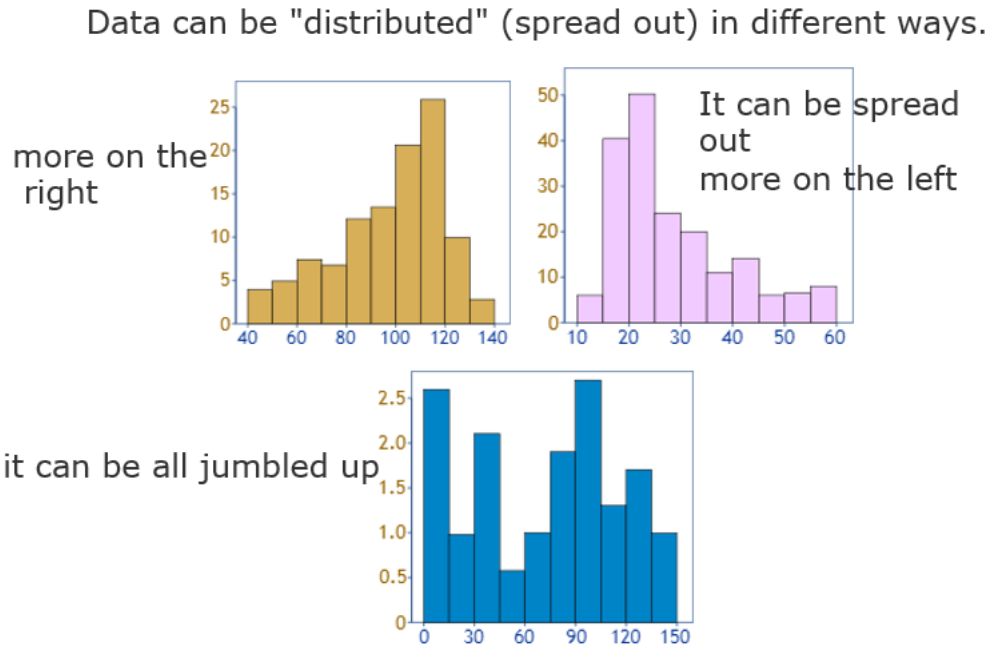
**1.4.What is Distribution in statistics?**

- A **distribution** is simply a collection of data, or scores, on a variable. Usually, these scores are arranged in order from smallest to largest and then they can be presented graphically.
- **A distribution** is an arrangement of values of a variable showing their observed or theoretical frequency of occurrence.
- A **bell curve** showing how the class did on our last exam would be an example of a distribution.

- All distributions can be characterized by the following two dimensions:
- **Central Tendency:** Mean, Median and Mode(s) of the distribution
- **Variability:** All distributions have a variance and standard deviation
- **Bell Curve**
- The term **bell curve** is used to describe the mathematical concept called normal distribution, sometimes referred to as Gaussian distribution.
- "Bell curve" refers to the bell shape that is created when a line is plotted using the data points for an item that meets the criteria of **normal distribution.**
- In a bell curve, the center contains the greatest number of a value and, therefore, it is the highest point on the arc of the line. This point is referred to the mean, but in simple terms, it is the highest number of occurrences of an element (in statistical terms, the mode).




**Distribution**



- But there are many cases where the data tends to be around a central value with no bias left or right, and it gets close to a **"Normal Distribution"** like this:

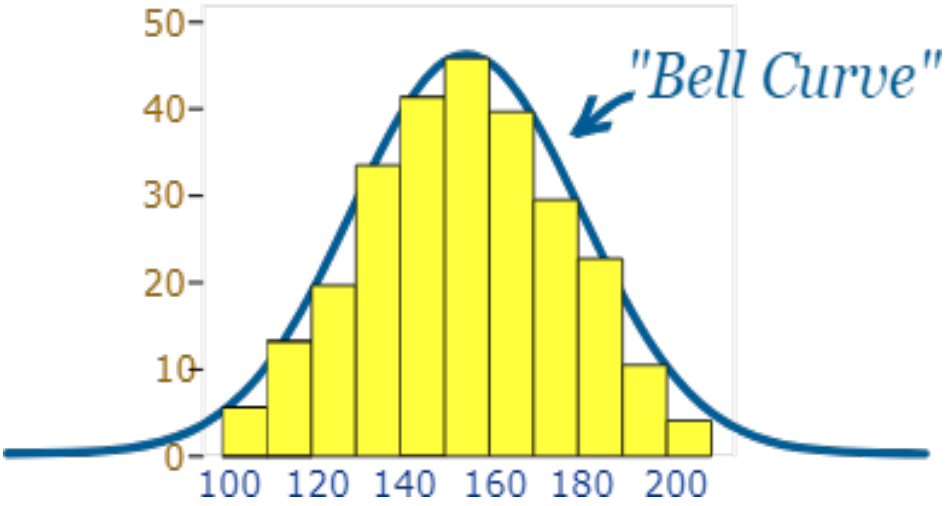



Figure 2.2: Bell cure

### 1.4.1. Normal Distribution

- **Normal distribution:** a bell-shaped, symmetrical distribution in which the mean, median and mode are all equal Z scores (also known as standard scores): the number of standard deviations that a given raw score falls above or below the mean Standard normal distribution: a normal distribution represented in z scores. The standard normal distribution always has a mean of zero and a standard deviation of one.
- The normal distribution is an important class of **Statistical Distribution** that has a wide range of applications. This distribution applies in most Machine Learning Algorithms and the concept of the Normal Distribution is a must for any **Statistician**, **Machine Learning Engineer**, and **Data Scientist.**

**Parameters of Normal Distribution**

- Mean
- Standard Deviation

**Properties of Normal Distribution**

- Symmetricity
- Measures of Central Tendencies are equal
- Empirical Rule
- Skewness and Kurtosis
- The area under the curve

**Properties of Normal Distribution**

- All forms of the normal distribution share the following characteristics:

**1. It is symmetric**

- The shape of the normal distribution is perfectly symmetrical.

- This means that the curve of the normal distribution can be divided from the middle and we can produce two equal halves. Moreover, the symmetric shape exists when an equal number of observations lie on each side of the curve.

**2. The mean, median, and mode are equal**

- The midpoint of normal distribution refers to the point with maximum frequency i.e., it consists of most observations of the variable.
- The midpoint is also the point where all three measures of central tendency fall. These measures are usually equal in a perfectly shaped normal distribution.

**3. Empirical rule**

- In normally distributed data, there is a constant proportion of data points lying under the curve between the mean and a specific number of standard deviations from the mean.
- Thus, for a normal distribution, almost all values lie within **3 standard deviations** of the mean.
- These check buttons of normal distribution will help you realize the appropriate percentages of the area under the curve.
- Remember that this empirical rule applies to all normal distributions. Also, note that these rules are applied only to the normal distributions.

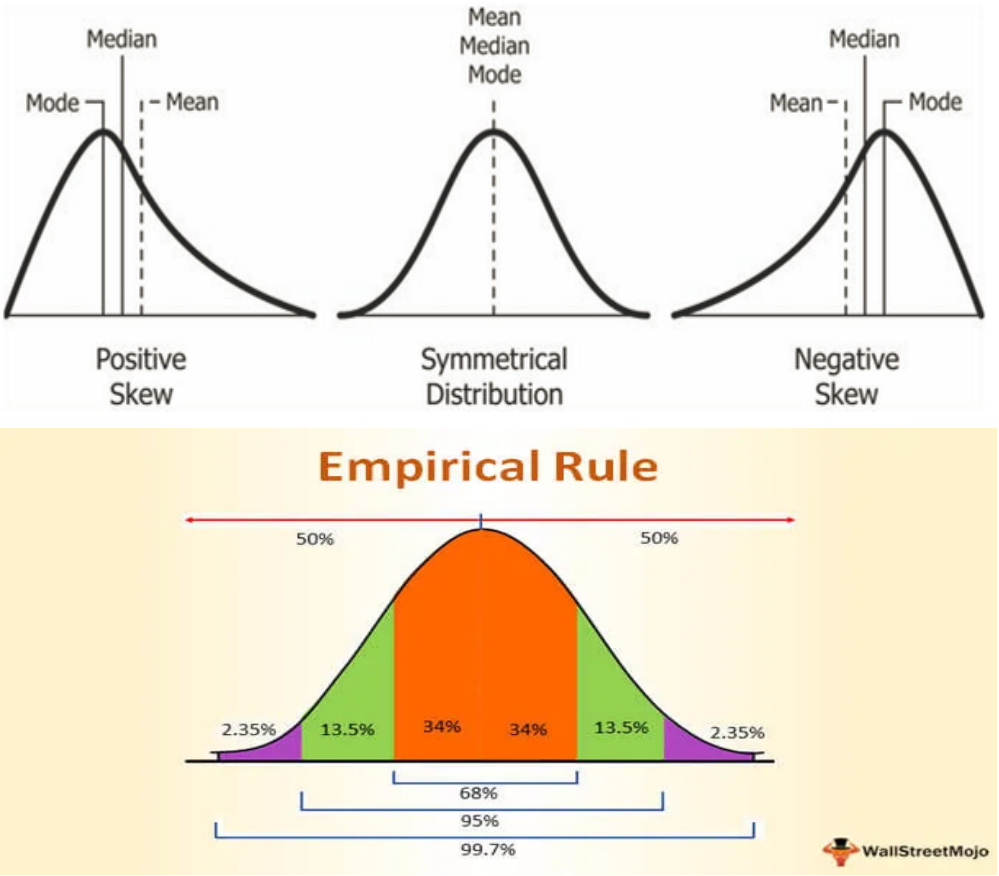
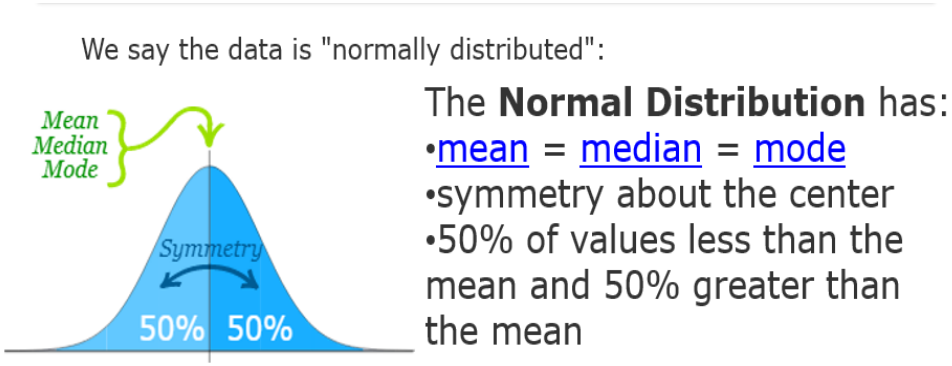



Figure 2.3: Empirical rule

**Many things closely follow a Normal Distribution**

Example:

- Heights of people
- Size of things produced by machines
- Errors in measurements
- Blood pressure
- Marks on a test




The normal distribution is a bell-shaped, symmetrical distribution in which the mean, median and mode are all equal. If the mean, median and mode are unequal, the distribution will be either positively or negatively skewed. Consider the illustration below:

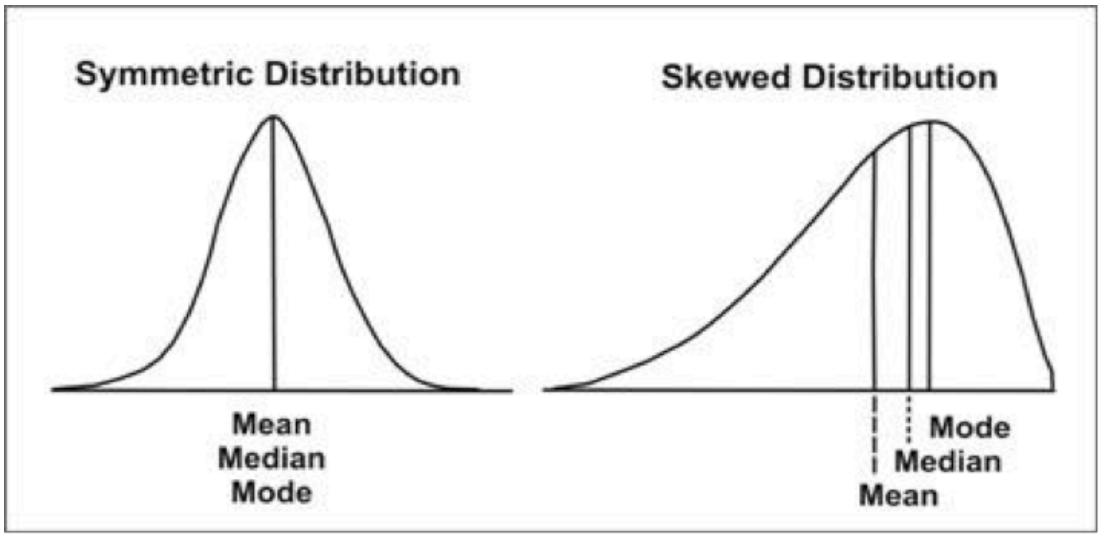



Figure 2.4: Symmetric and Skewed distribution

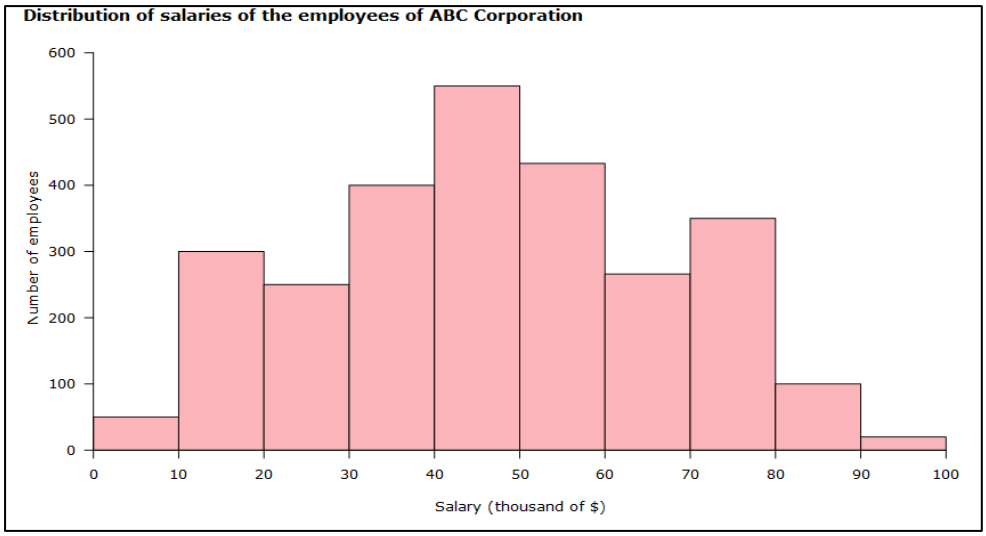
**1.5.Probability Density**

- Given a random variable, we are interested in the density of its probabilities.
- For example, given a random sample of a variable, we might want to know things like the shape of the probability distribution, the most likely value, the spread of values, and other properties.
- Knowing the probability distribution for a random variable can help to **calculate** moments of the distribution, like the **mean and variance**, but can also be useful for other more general

considerations, like determining whether an observation is unlikely or very unlikely and might be an outlier or anomaly.

- The problem is, we may not know the probability distribution for a random variable.
- We rarely do know the distribution because we don't have access to all possible outcomes for a random variable. In fact, all we have access to is a sample of observations. As such, we must select a probability distribution.
- This problem is referred to as probability density estimation, or **simply "*density estimation*,"** as we are using the observations in a random sample to estimate the general density of probabilities beyond just the sample of data, we have available.
- A random variable *x* has a probability distribution *p(x)*.
- The relationship between the outcomes of a random variable and its probability is referred to as the probability density, or simply the "*density*."
- If a random variable is continuous, then the probability can be calculated via probability density function, or PDF for short.
- The shape of the probability density function across the domain for a random variable is referred to as the probability distribution and common probability distributions have names, such as uniform, normal, exponential, and so on.
- There are a few **steps in the process of density estimation** for a random variable.
- The first step is to review the density of observations in the random sample with a simple histogram.
- From the histogram, we might be able to identify a common and well-understood probability distribution that can be used, such as a normal distribution. If not, we may have to fit a model to estimate the distribution.

**Histogram**



**Density With a Histogram**

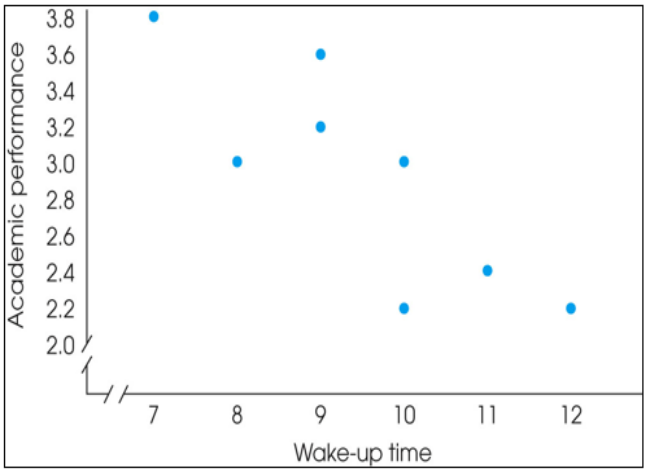
- The first step in density estimation is to create a histogram of the observations in the random sample.
- A histogram is a plot that involves first grouping the observations into bins and counting the number of events that fall into each bin.
- The counts, or frequencies of observations, in each bin are then plotted as a bar graph with the bins on the x-axis and the frequency on the y-axis.
- The choice of the number of bins is important as it controls the coarseness of the distribution (number of bars) and, in turn, how well the density of the observations is plotted.
- It is a good idea to experiment with different bin sizes for a given data sample to get multiple perspectives or views on the same data.

**Correlational Studies**

- The goal of a **correlational** study is to determine whether there is a relationship between two variables and to describe the relationship.
- A **correlational** study simply observes the two variables as they exist naturally.

**Correlational Studies**

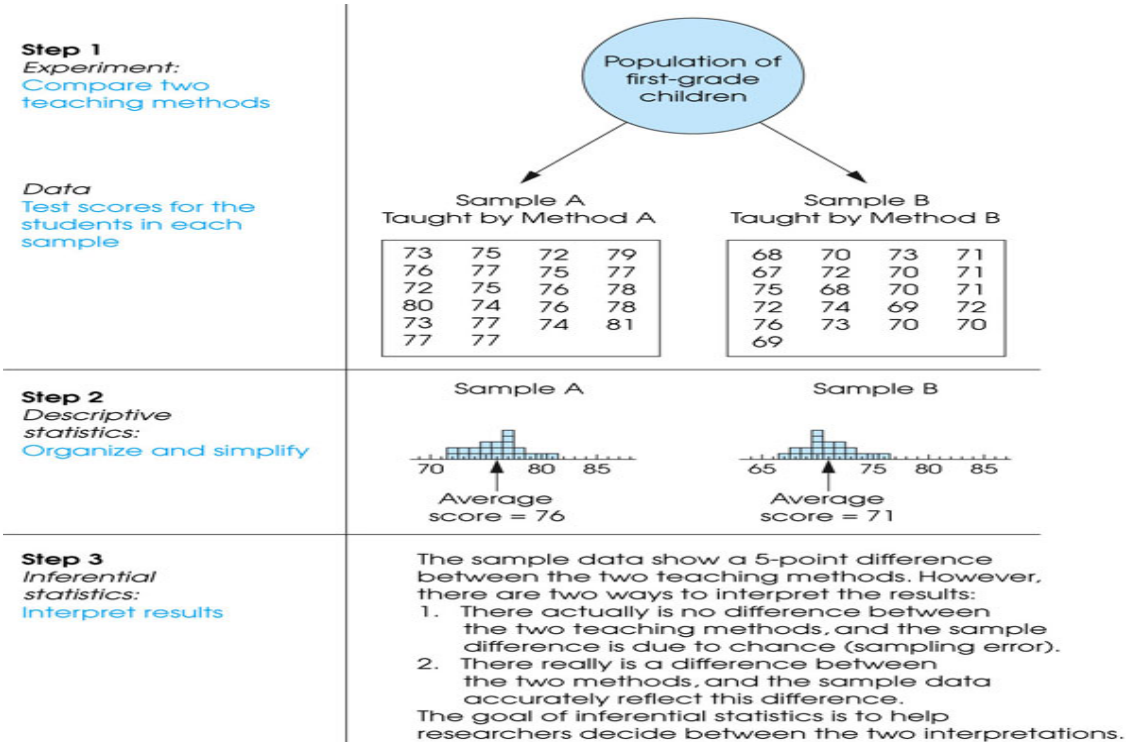
| Child | Wake-up Time | Academic Performance |
|---|---|---|
| A | 11 | 2.4 |
| B | 9 | 3.6 |
| C | 9 | 3.2 |
| D | 12 | 2.2 |
| E | 7 | 3.8 |
| F | 10 | 2.2 |
| G | 10 | 3.0 |
| H | 8 | 3.0 |




**Experiment**

- The goal of an **experiment** is to demonstrate a cause-and-effect relationship between two variables; that is, to show that changing the value of one variable causes change to occur in a second variable.
- In an **experiment**, one variable is manipulated to create treatment conditions.
- A second variable is observed and measured to obtain scores for a group of individuals in each of the treatment conditions.
- The measurements are then compared to see if there are differences between treatment conditions.
- All other variables are controlled to prevent them from influencing the results.

- In an experiment, the manipulated variable is called the **independent variable** and the observed variable is the **dependent variable**.




## 1.6. What Is a Probability Density Function (PDF)?

A probability distribution can be described in various forms, such as by a probability density function or a cumulative distribution function. Probability density functions, or PDFs, are mathematical functions that usually apply to continuous and discrete values. PDFs are very commonly used in statistical analysis, and thus are quite commonly used for Data Science. Generally, PDFs are a necessary tool when studying data with applied science using statistics. However, there are some PDFs that extend beyond this basic usage and have slightly different uses than one might be assume on first glance. For example, the PDF of the T distribution is often used to calculate a T-statistic. This T statistic, along with the degrees of freedom (n minus one) (v,) are then usually put into the regularized lower incomplete beta function, which happens to be the cumulative distribution function for the T distribution. While the absolute likelihood for a continuous random variable to take on any particular value is 0, the value of the PDF can be used to infer, in any particular sample of random variables, how much more likely it is statistically that the random variable would equal one sample compared to the other sample.

A function that defines the relationship between a random variable and its probability, such that you can find the probability of the variable using the function, is called a Probability Density Function (PDF) in statistics.

The different types of variables. They are mainly of two types:

1. Discrete Variable: A variable that can only take on a certain finite value within a specific range is called a discrete variable. It usually separates the values by a finite interval, e.g.,

a sum of two dice. On rolling two dice and adding up the resulting outcome, the result can only belong to a set of numbers not exceeding 12 (as the maximum result of a dice throw is 6). The values are also definite.

2. Continuous Variable: A continuous random variable can take on infinite different values within a range of values, e.g., amount of rainfall occurring in a month. The rain observed can be 1.7cm, but the exact value is not known. It can, in actuality, be 1.701, 1.7687, etc. As such, you can only define the range of values it falls into. Within this value, it can take on infinite different values.

Now, consider a continuous random variable x, which has a probability density function, that defines the range of probabilities taken by this function as f(x). After plotting the pdf, you get a graph as shown below:

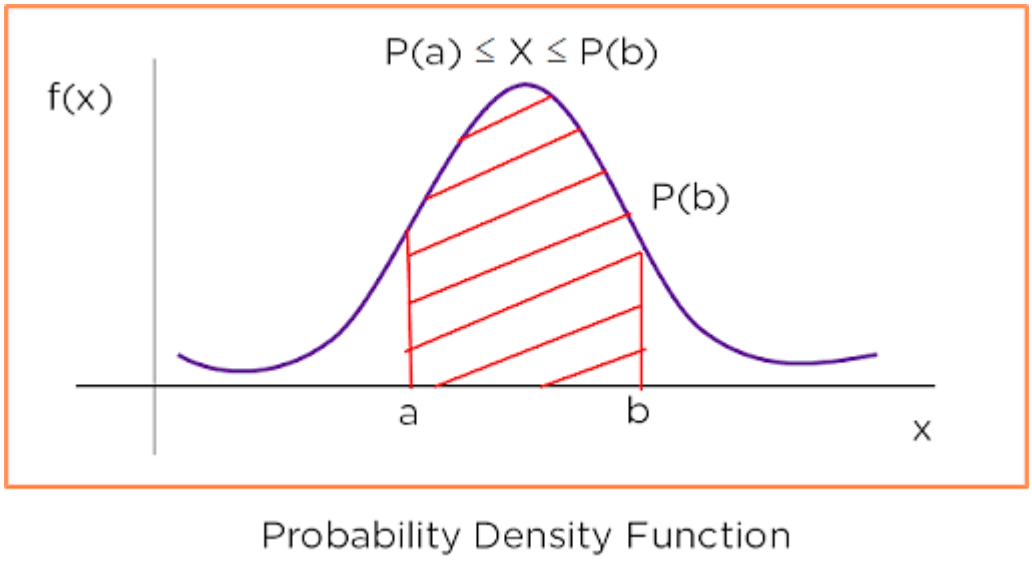



Figure 2.5: Probability Density Function

In the above graph, you get a bell-shaped curve after plotting the function against the variable. The blue curve shows this. Now consider the probability of a point b. To find it, you need to find the area under the curve to the left of b. This is represented by P(b). To find the probability of a variable falling between points a and b, you need to find the area of the curve between a and b. As the probability cannot be more than P(b) and less than P(a), you can represent it as:

P(a) <= X <= P(b).

Consider the graph below, which shows the rainfall distribution in a year in a city. The x-axis has the rainfall in inches, and the y-axis has the probability density function. The probability of some amount of rainfall is obtained by finding the area of the curve on the left of it.

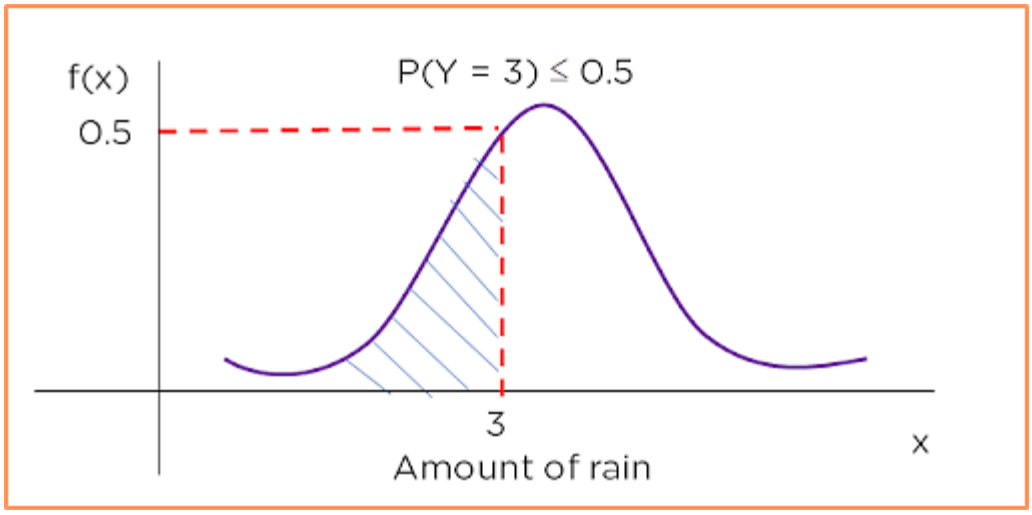



Figure 2.6: Probability Density Function of the amount of rainfall

For the probability of 3 inches of rainfall, you plot a line that intersects the y-axis at the same point on the graph as a line extending from 3 on the x-axis does. This tells you that the probability of 3 inches of rainfall is less than or equal to 0.5.

## 1.7. Descriptive Statistics

### What is Statistics?

Statistics is the science of collecting data and analyzing them to infer proportions (sample) that are representative of the population. In other words, statistics is interpreting data in order to make predictions for the population.

### Descriptive Statistics

Descriptive Statistics is summarizing the data at hand through certain numbers like mean, median etc. so as to make the understanding of the data easier. It does not involve any generalization or inference beyond what is available. This means that the descriptive statistics are just the representation of the data (sample) available and not based on any theory of probability.

### Commonly Used Measures

1. Measures of Central Tendency
2. Measures of Dispersion (or Variability)

### Measures of Central Tendency

A Measure of Central Tendency is a one number summary of the data that typically describes the center of the data. These one number summary is of three types.

1. **Mean :** Mean is defined as the ratio of the sum of all the observations in the data to the total number of observations. This is also known as Average. Thus mean is a number around which the entire data set is spread.
2. **Median :** Median is the point which divides the entire data into two equal halves. One-half of the data is less than the median, and the other half is greater than the same. Median is calculated by first arranging the data in either ascending or descending order.

- If the number of observations are odd, median is given by the middle observation in the sorted form.

- If the number of observations are even, median is given by the mean of the two middle observation in the sorted form.

An important point to note that the order of the data (ascending or descending) does not effect the median.

**3. Mode :** Mode is the number which has the maximum frequency in the entire data set, or in other words,mode is the number that appears the maximum number of times. A data can have one or more than one mode.

- If there is only one number that appears maximum number of times, the data has one mode, and is called **Uni-modal**.
- If there are two numbers that appear maximum number of times, the data has two modes, and is called **Bi-modal**.
- If there are more than two numbers that appear maximum number of times, the data has more than two modes, and is called **Multi-modal**.

*Example to compute the Measures of Central Tendency*

Consider the following data points.

**17, 16, 21, 18, 15, 17, 21, 19, 11, 23**

- Mean — Mean is calculated as

$$Mean = \frac{17+16+21+18+15+17+21+19+11+23}{10} = \frac{178}{10} = 17.8$$

- Median — To calculate Median, lets arrange the data in ascending order.

11, 15, 16, 17, 17, 18, 19, 21, 21, 23

Since the number of observations is even (10), median is given by the average of the two middle observations (5th and 6th here).

$$Median = \frac{5^{th}\ Obs + 6^{th}\ Obs}{2} = \frac{17+18}{2} = 17.5$$

- Mode — Mode is given by the number that occurs maximum number of times. Here, 17 and 21 both occur twice. Hence, this is a Bimodal data and the modes are 17 and 21.
- Since Median and Mode does not take all the data points for calculations, these are robust to outliers, i.e. these are not effected by outliers.
- At the same time, Mean shifts towards the outlier as it considers all the data points. This means if the outlier is big, mean overestimates the data and if it is small, the data is underestimated.
- If the distribution is symmetrical, Mean = Median = Mode. Normal distribution is an example.

### 1.8.Notion of probability, distributions, mean, variance, covariance, covariance matrix,

Probability and Statistics form the basis of Data Science. The probability theory is very much helpful for making the prediction. Estimates and predictions form an important part of Data science. With the help of statistical methods, we make estimates for the further analysis. Thus, statistical methods are largely dependent on the theory of probability. And all of probability and statistics is dependent on Data.

### 1.8.1. Data

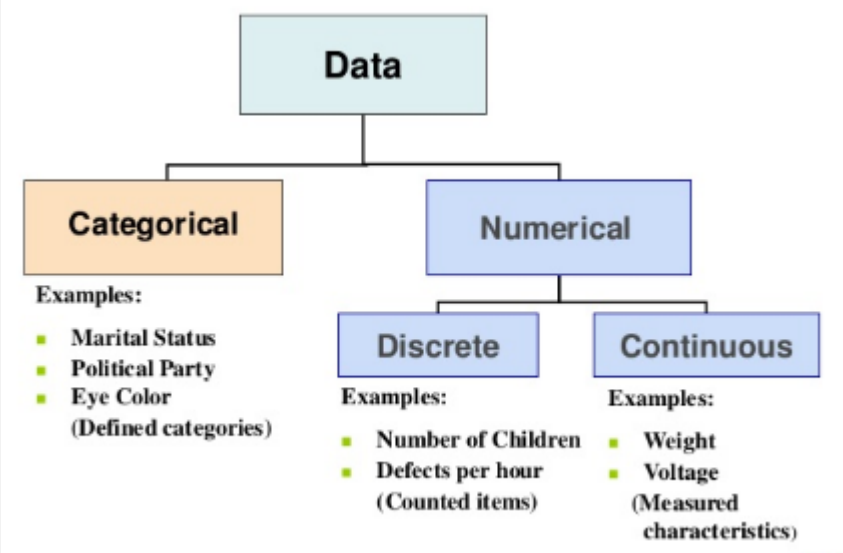
Data is the collected information(observations) we have about something or facts and statistics collected together for reference or analysis.

*Data — a collection of facts (numbers, words, measurements, observations, etc) that has been translated into a form that computers can process*

**Why does Data Matter?**

- Helps in understanding more about the data by identifying relationships that may exist between 2 variables.
- Helps in predicting the future or forecast based on the previous trend of data.
- Helps in determining patterns that may exist between data.
- Helps in detecting fraud by uncovering anomalies in the data.

Data matters a lot nowadays as we can infer important information from it. Now let's delve into how data is categorized. Data can be of 2 types categorical and numerical data. For Example in a bank, we have regions, occupation class, gender which follow categorical data as the data is within a fixed certain value and balance, credit score, age, tenure months follow numerical continuous distribution as data can follow an unlimited range of values.

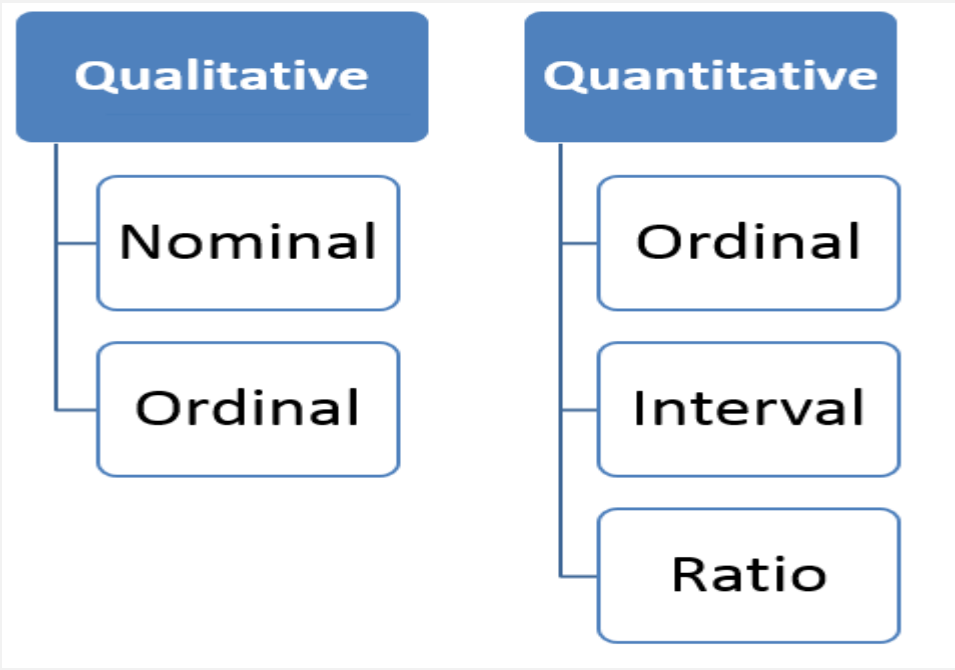



Note: Categorical Data can be visualized by Bar Plot, Pie Chart, Pareto Chart. Numerical Data can be visualized by Histogram, Line Plot, Scatter Plot

### 1.8.2. Descriptive Statistics

A descriptive statistic is a summary statistic that quantitatively describes or summarizes features of a collection of information. It helps us in knowing our data better. It is used to describe the characteristics of data.

**Measurement level of Data**




The qualitative and quantitative data is very much similar to the above categorical and numerical data.

**Nominal**: Data at this level is categorized using names, labels or qualities. eg: Brand Name, ZipCode, Gender.

**Ordinal**: Data at this level can be arranged in order or ranked and can be compared. eg: Grades, Star Reviews, Position in Race, Date

**Interval**: Data at this level can be ordered as it is in a range of values and meaningful differences between the data points can be calculated. eg: Temperature in Celsius, Year of Birth

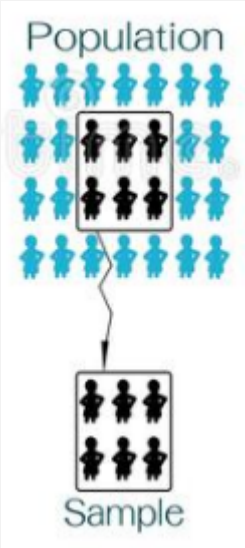
**Ratio**: Data at this level is similar to interval level with added property of an inherent zero. Mathematical calculations can be performed on these data points. eg: Height, Age, Weight

### 1.8.3. Population or Sample Data

Before performing any analysis of data, we should determine if the data we're dealing with is population or sample.

**Population:** Collection of all items (N) and it includes each and every unit of our study. It is hard to define and the measure of characteristic such as mean, mode is called parameter.

**Sample:** Subset of the population (n) and it includes only a handful units of the population. It is selected at random and the measure of the characteristic is called as statistics.




For Example, say you want to know the mean income of the subscribers to a movie subscription service(parameter). We draw a random sample of 1000 subscribers and determine that their mean income($\bar{x}$) is $34,500 (statistic). We conclude that the population mean income ($\mu$) is likely to be close to $34,500 as well.

### 1.8.4. Measures of Central Tendency

The 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. They are also classed as summary statistics.

**1.8.5. Mean**: The mean is equal to the sum of all the values in the data set divided by the number of values in the data set i.e the calculated average. **It susceptible to outliers** when unusual values are added it gets skewed i.e deviates from the typical central value.

$$\bar{x} = \frac{(x_1 + x_2 + \cdots + x_n)}{n}$$

**1.8.6. Median**: The median is the middle value for a dataset that has been arranged in order of magnitude. Median is a better alternative to mean as it is less affected by outliers and skewness of the data. The median value is much closer than the typical central value.

If the total number of values is odd then

$$\text{Median} = \left(\frac{n+1}{2}\right)^{th} \text{term}$$

If the total number of values is even then

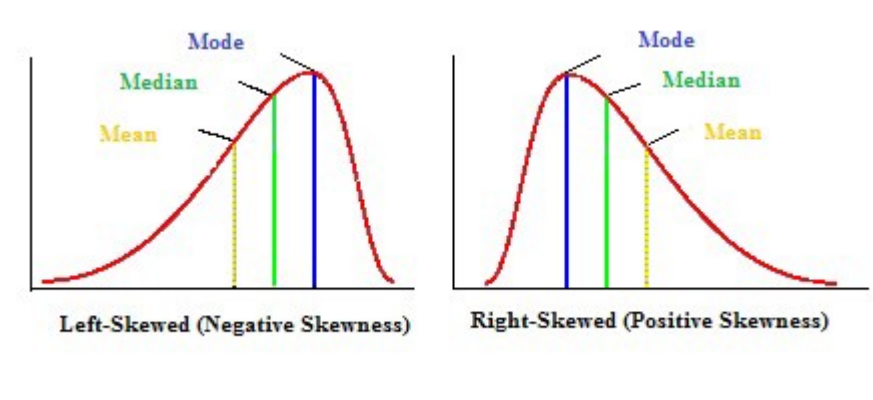
$$\text{Median} = \left(\frac{\left(\frac{n}{2}\right)^{th} term + \left(\frac{n}{2}+1\right)^{th} term}{2}\right)^{\text{th}} term$$

**1.8.7. Mode:** The mode is the most commonly occurring value in the dataset. The mode can, therefore sometimes consider the mode as being the most popular option.

For Example, In a dataset containing {13,35,54,54,55,56,57,67,85,89,96} values. Mean is 60.09. Median is 56. Mode is 54.

### 1.8.8. Measures of Asymmetry

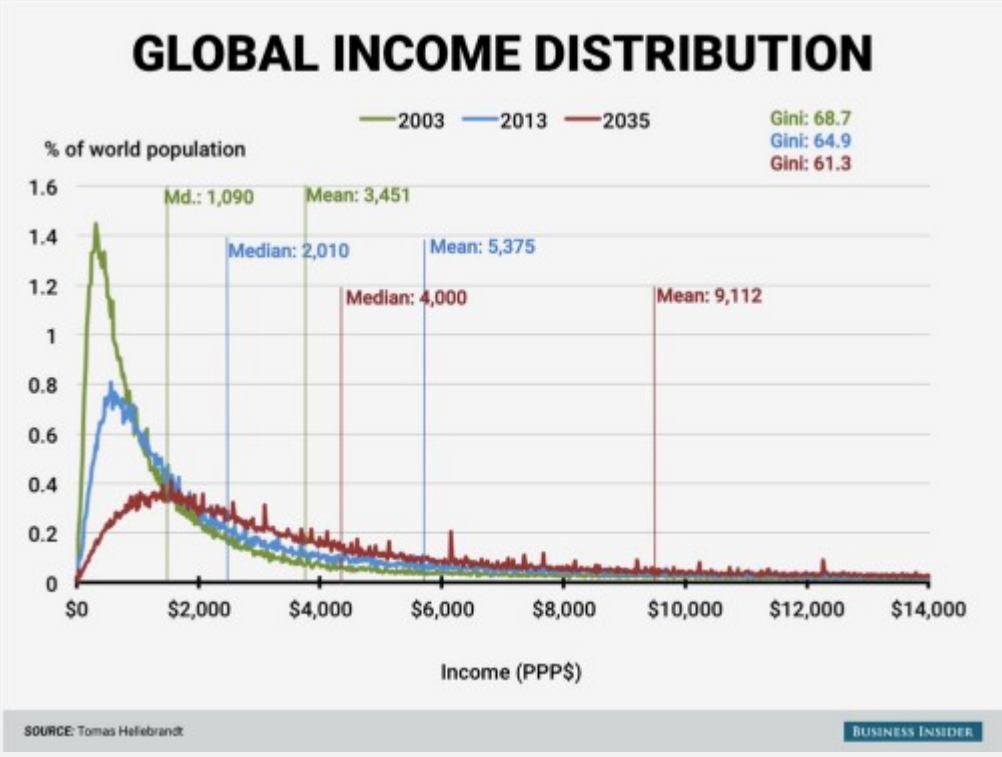
**Skewness:** Skewness is the asymmetry in a statistical distribution, in which the curve appears distorted or skewed towards to the left or to the right. Skewness indicates whether the data is concentrated on one side.




**Positive Skewness:** Positive Skewness is when the mean>median>mode. The outliers are skewed to the right i.e the tail is skewed to the right.

**Negative Skewness:** Negative Skewness is when the mean<median<mode. The outliers are skewed to the left i.e the tail is skewed to the left.

Skewness is important as it tells us about where the data is distributed.




For eg: Global Income Distribution in 2003 is highly right-skewed.We can see the mean $3,451 in 2003(green) is greater than the median $1,090. It suggests that the global income is not evenly distributed. Most individuals incomes are less than $2,000 and less number of people with income

above $14,000, so the skewness. But it seems in 2035 according to the forecast income inequality will decrease over time.

### 1.8.9. Measures of Variability(Dispersion)

The measure of central tendency gives a single value that represents the whole value; however, the central tendency cannot describe the observation fully. The measure of dispersion helps us to study the variability of the items i.e the spread of data.

*Remember: Population Data has N data points and Sample Data has (n-1) data points. (n-1) is called Bessel's Correction and it is used to reduce bias.*

**1.8.10. Range**: The difference between the largest and the smallest value of a data, is termed as the range of the distribution. Range does not consider all the values of a series, i.e. it takes only the extreme items and middle items are not considered significant. eg: For {13,33,45,67,70} the range is 57 i.e(70–13).

**1.8.11. Variance:** Variance measures how far is the sum of squared distances from each point to the mean i.e the dispersion around the mean.

*Variance is the average of all squared deviations.*

$$\sigma^2 = \frac{\sum (x_i - \mu)^2}{n} \text{ for populations}$$

$$s^2 = \frac{\sum (x_i - \bar{x})^2}{n-1} \text{ for samples}$$

Note: *The units of values and variance is not equal so we use another variability measure.*

**1.8.12. Standard Deviation:** As Variance suffers from unit difference so standard deviation is used. The square root of the variance is the standard deviation. It tells about the concentration of the data around the mean of the data set.

**Population standard deviation:** $\sigma$

= square root of the population variance

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

**Sample standard deviation:** $s$

= square root of the sample variance, so that

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

For eg: {3,5,6,9,10} are the values in a dataset.

$$\text{Mean} = \frac{3+5+6+9+10}{5} = 6.6$$

$$\text{Variance} = \frac{(3-6.6)^2+(5-6.6)^2+(6-6.6)^2+(9-6.6)^2+(10-6.6)^2}{5}$$

$$= \frac{12.96+2.56+0.36+5.76+11.56}{5} = \frac{33.2}{5} = 6.64$$

$$\text{Standard Deviation} = \sqrt{\text{Variance}} = \sqrt{6.64} = 2.576$$

**Given measurements on a sample, what is the difference between a standard deviation and a standard error?**

A **standard deviation is** a sample estimate of the population parameter; that is, it is an estimate of the variability of the observations. Since the population is unique, it has a unique standard deviation, which may be large or small depending on how variable the observations are. We would not expect the sample standard deviation to get smaller because the sample gets larger. However, a large sample would provide a more precise estimate of the population standard deviation than a small sample.
A **standard error**, on the other hand, is a measure of precision of an estimate of a population parameter. A standard error is always attached to a parameter, and one can have standard errors of any estimate, such as mean, median, fifth centile, even the standard error of the standard deviation. Since one would expect the precision of the estimate to increase with the sample size, the standard error of an estimate will decrease as the sample size increases.

**1.8.13. Coefficient of Variation(CV):** It is also called as the relative standard deviation. It is the ratio of standard deviation to the mean of the dataset.

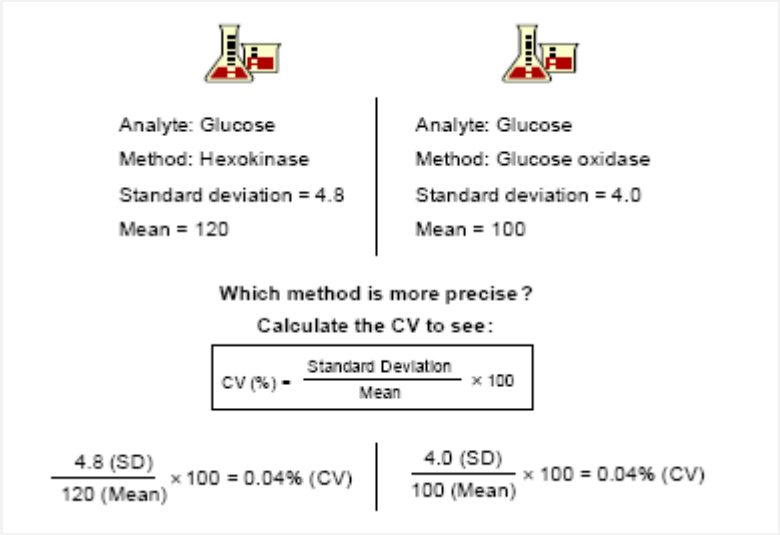
CV for a population:

$$CV = \frac{\sigma}{\mu} * 100\%$$

CV for a sample:

$$CV = \frac{s}{\bar{x}} * 100\%$$

Standard deviation is the variability of a single dataset. Whereas the coefficient of variance can be used for comparing 2 datasets.



From the above example, we can see that the CV is the same. Both methods are precise. So it is perfect for comparisons.

### 1.8.14. Measures of Quartiles

Quartiles are better at understanding as every data point considered.

### Measures of Relationship

Measures of relationship are used to find the comparison between 2 variables.

**1.8.15. Covariance:** Covariance is a measure of the relationship between the variability of 2 variables i.e It measures the degree of change in the variables, when one variable changes, will there be the same/a similar change in the other variable.

A population covariance is

$$Cov(x,y) = \sigma_{xy} = \frac{\sum_{i=1}^{n}(x_i - \mu_x)(y_i - \mu_y)}{N}$$

where $x_i$ and $y_i$ are the observed values, $\mu_x$ and $\mu_y$ are the population means, and N is the population size.

A sample covariance is

$$Cov(x,y) = s_{xy} = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{n-1}$$

where $x_i$ and $y_i$ are the observed values, $\bar{x}$ and $\bar{y}$ are the sample means, and n is the sample size.

Covariance does not give effective information about the relation between 2 variables as it is not normalized.

**1.8.16. Correlation:** Correlation gives a better understanding of covariance. It is normalized covariance. Correlation tells us how correlated the variables are to each other. It is also called as Pearson Correlation Coefficient.

$$Correlation = \rho = \frac{\text{cov}(X,Y)}{\sigma_X \sigma_Y}$$

The value of correlation ranges from -1 to 1. -1 indicates negative correlation i.e with an increase in 1 variable independent there is a decrease in the other dependent variable.1 indicates positive correlation i.e with an increase in 1 variable independent there is an increase in the other dependent variable.0 indicates that the variables are independent of each other.

For Example,

| Height | Weight | $x-\bar{x}$ | $y-\bar{y}$ | $(x-\bar{x})(y-\bar{y})$ | $(x-\bar{x})^2$ | $(y-\bar{y})^2$ |
|---|---|---|---|---|---|---|
| 5 | 45 | −0.14 | −5 | 0.7 | 0.019 | 25 |
| 5.5 | 53 | −0.36 | 3 | −1.08 | 0.129 | 9 |
| 6 | 70 | 0.86 | 20 | 17.2 | 0.739 | 400 |
| 4.7 | 42 | −0.44 | −8 | 3.52 | 0.193 | 64 |
| 4.5 | 40 | −0.64 | −10 | 6.4 | 0.409 | 100 |

Sum(Height) = 25.7 Mean(Height) = 5.14

Sum(Weight) = 250 Mean(Weight) = 50

$\sum(x-\bar{x})(y-\bar{y}) = 26.74$

$\sum(x-\bar{x})^2 = 1.489$

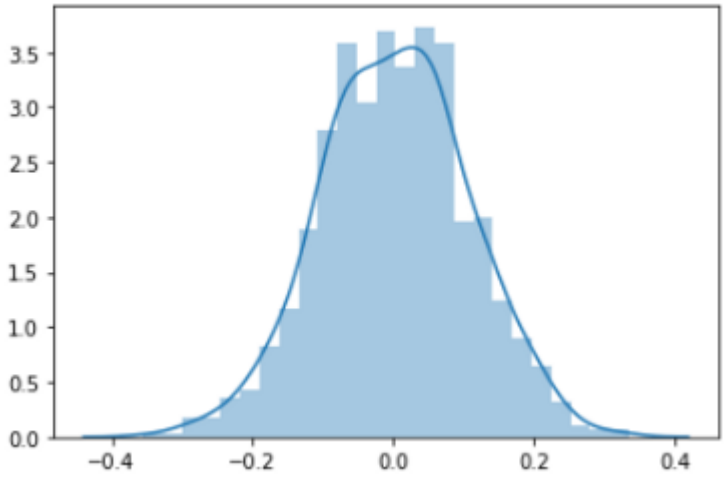
$\sum(y-\bar{y})^2 = 598$

$$\text{Correlation} = \frac{\sum(x-\bar{x})(y-\bar{y})}{\sqrt{\sum(x-\bar{x})^2}\sqrt{\sum(y-\bar{y})^2}} = \frac{26.74}{\sqrt{1.489}\sqrt{598}} = \frac{26.54}{1.220 * 24.454} = 0.889$$

Correlation 0.889 tells us Height and Weight has a positive correlation. It is obvious that as the height of a person increases weight too increases.

### 1.9.Understanding Univariate and Multivariate Normal Distribution

**Gaussian distribution** is a synonym for normal distribution. S is a set of random values whose probability distribution looks like the picture below.




This is a bell-shaped curve. If a probability distribution plot forms a bell-shaped curve like above and the mean, median, and mode of the sample are the same that distribution is called normal distribution or Gaussian distribution.

The Gaussian distribution is parameterized by two parameters:

- The mean and The variance

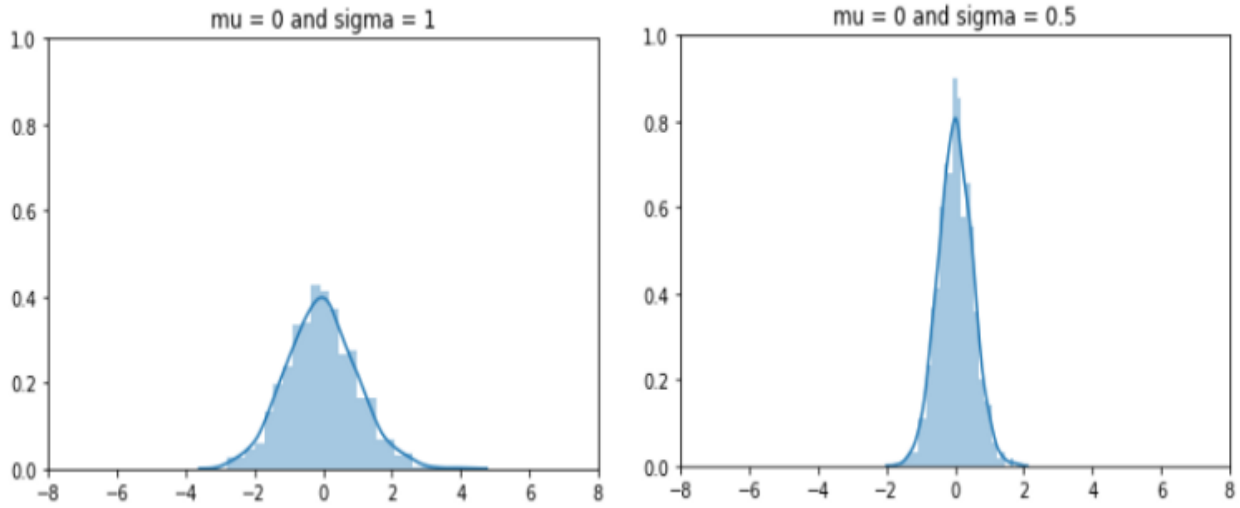
So, the Gaussian density is the highest at the point of μ or mean, and further, it goes from the mean, the Gaussian density keeps going lower.

Here is the formula for the Gaussian distribution:

$$p(x;\ \mu, \sigma^2) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(x-\ \mu)^2}{2\sigma^2}\right)$$

This is the formula for the bell-shaped curve where sigma square is called the variance.

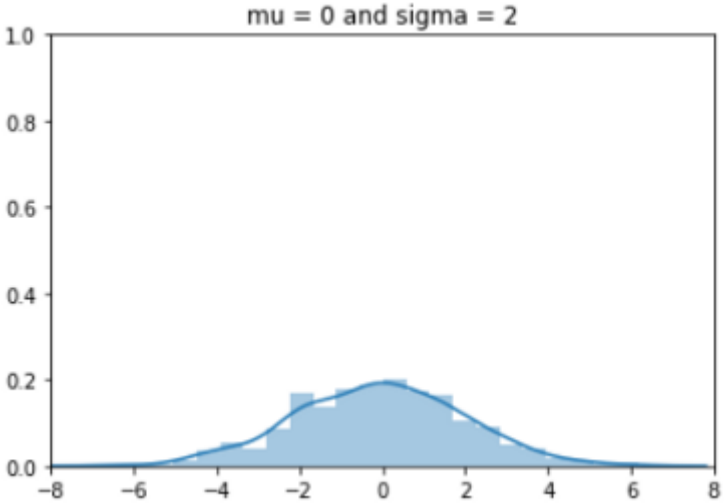
Mean =0, and different sigmas



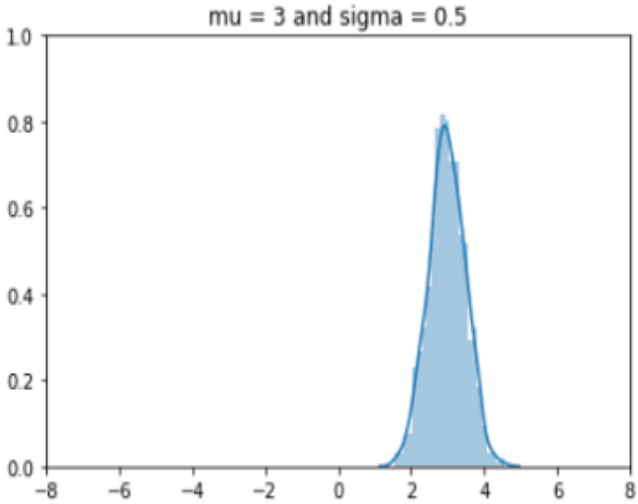
This is the probability distribution of a set of random numbers with μ is equal to 0 and sigma is 1. In the first picture, μ is 0 which means the highest probability density is around 0 and the sigma is one. means the width of the curve is 1.  the height of the curve is about 0.5 and the range is -4 to 4 (look at x-axis). The variance sigma square is 1.

Here is another set of random numbers that has a μ of 0 and sigma 0.5 in the second figure. Because the  μ is 0, like the previous picture the highest probability density is at around 0 and the sigma is 0.5. So, the width of the curve is 0.5. The variance sigma square becomes 0.25.

As the width of the curve is half the previous curve, the height became double. The range changed to -2 to 2 (x-axis) which is the half of the previous picture.




In this picture, sigma is 2 and μ is 0 as the previous two pictures. Compare it to figure 1 where sigma was 1. This time height became half of figure 1. Because the width became double as the sigma became double. The variance sigma square is 4, four times bigger than figure 1. Look at the range in the x-axis, it's -8 to 8.




Here, we changed μ to 3 and sigma is 0.5 as figure 2. So, the shape of the curve is exactly the same as figure 2 but the center shifted to 3. Now the highest density is at around 3.

It changes shapes with the different values of sigma but the area of the curve stays the same.

One important property of probability distribution is, the area under the curve is integrated to one.

**Parameter Estimation**

Calculating μ is straight forward. it's simply the average. Take the summation of all the data and divide it by the total number of data.

$$\mu = \frac{1}{m}\sum_{i=1}^{m} x^i$$

The formula for the variance (sigma square) is:

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

### 1.9.1. Univariate Normal Distributions

- Before defining the multivariate normal distribution, we will visit the univariate normal distribution. A random variable *X* is normally distributed with mean μ and variance σ2 if it has the probability density function of *X* as:
- ϕ(x)=12πσ2exp⁡{−12σ2(x−μ)2}
- This result is the usual bell-shaped curve that you see throughout statistics.
- In this expression, you see the squared difference between the variable *x* and its mean, μ. This value will be minimized when x is equal to μ. The quantity −σ−2(x−μ)2 will take its largest value when *x* is equal to μ or likewise, since the exponential function is a monotone function, the normal density takes a maximum value when x is equal to μ.
- The variance σ2 defines the spread of the distribution about that maximum. If σ2 is large, then the spread is going to be large, otherwise, if the σ2 value is small, then the spread will be small.

### 1.9.2. Multivariate Gaussian Distribution

- **Multivariate analysis** is a branch of statistics concerned with the analysis of multiple measurements, made on one or several samples of individuals. For example, we may wish to measure length, width, and weight of a product.
- **Multivariate statistical analysis** is concerned with data that consist of sets of measurements on a number of individuals or objects.

- The sample data may be heights and weights of some individuals drawn randomly from a population of school children in a given city, or the statistical treatment may be made on a collection of measurements

**"Why is the multivariate normal distribution so important? "**

- There are three reasons why this might be so:
- *Mathematical Simplicity*. It turns out that this distribution is relatively easy to work with, so it is easy to obtain multivariate methods based on this particular distribution.
- *Multivariate version of the Central Limit Theorem*. You might recall in the univariate course that we had a central limit theorem for the sample mean for large samples of random variables. A similar result is available in multivariate statistics that says if we have a collection of random vectors X1, X2, ⋯Xn that are independent and identically distributed, then the sample mean vector, $\bar{x}$, is going to be approximately multivariate normally distributed for large samples.
- Many natural phenomena may also be modeled using this distribution, just as in the univariate case.

*Multivariate normal model*

When multivariate data are analyzed, the multivariate normal model is the most commonly used model.

The multivariate normal distribution model extends the univariate normal distribution model to fit vector observations.

*Definition of multivariate normal distribution*

A $p$-dimensional vector of random variables,

$$\mathbf{X} = X_1, X_2, \ldots, X_p \qquad -\infty < X_i < \infty, \; i = 1, \ldots, p,$$

is said to have a multivariate normal distribution if its density function $f(\mathbf{X})$ is of the form

$$\begin{aligned} f(\mathbf{X}) &= f(X_1, X_2, \ldots, X_p) \\ &= \left(\frac{1}{2\pi}\right)^{p/2} |\mathbf{\Sigma}|^{-1/2} \exp\left[-\frac{1}{2}(\mathbf{X}-\mathbf{m})'\mathbf{\Sigma}^{-1}(\mathbf{X}-\mathbf{m})\right], \end{aligned}$$

where $\mathbf{m} = (m_1, \ldots, m_p)$ is the vector of means and $\mathbf{\Sigma}$ is the variance-covariance matrix of the multivariate normal distribution. The shortcut notation for this density is

$$\mathbf{X} = \mathrm{N}_p(\mathbf{m}, \mathbf{\Sigma}).$$

Instead of having one set of data, what if we have two sets of data and we need a multivariate Gaussian distribution. Suppose we have two sets of data; x1 and x2.

Separately modeling p(x1) and p(x2) is probably not a good idea to understand the combined effect of both the dataset. In that case, you would want to combine both the dataset and model only p(x).

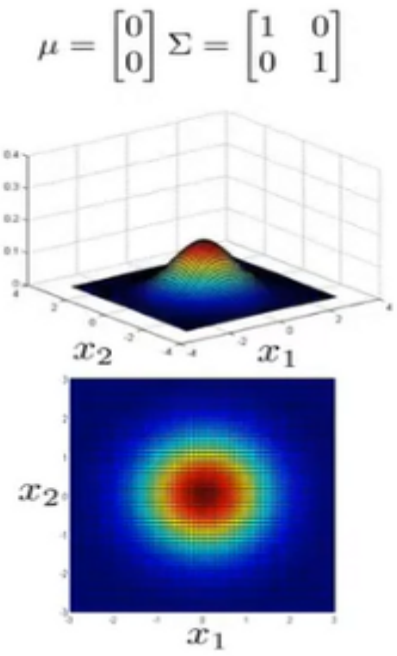
Here is the formula to calculate the probability for multivariate Gaussian distribution,

$$p(x;\ \mu, \Sigma) = \frac{1}{(2\pi)^{n/2}|\Sigma|^{\frac{1}{2}}} \exp\left(-\frac{1}{2}(x - \mu)^T \Sigma^{-1}(x - \mu)\right)$$

The summation symbol in this equation is the determinant of sigma which is actually an n x n matrix of sigma.

**Visual Representation of Multivariate Gaussian Distribution**

**Standard Normal Distribution**



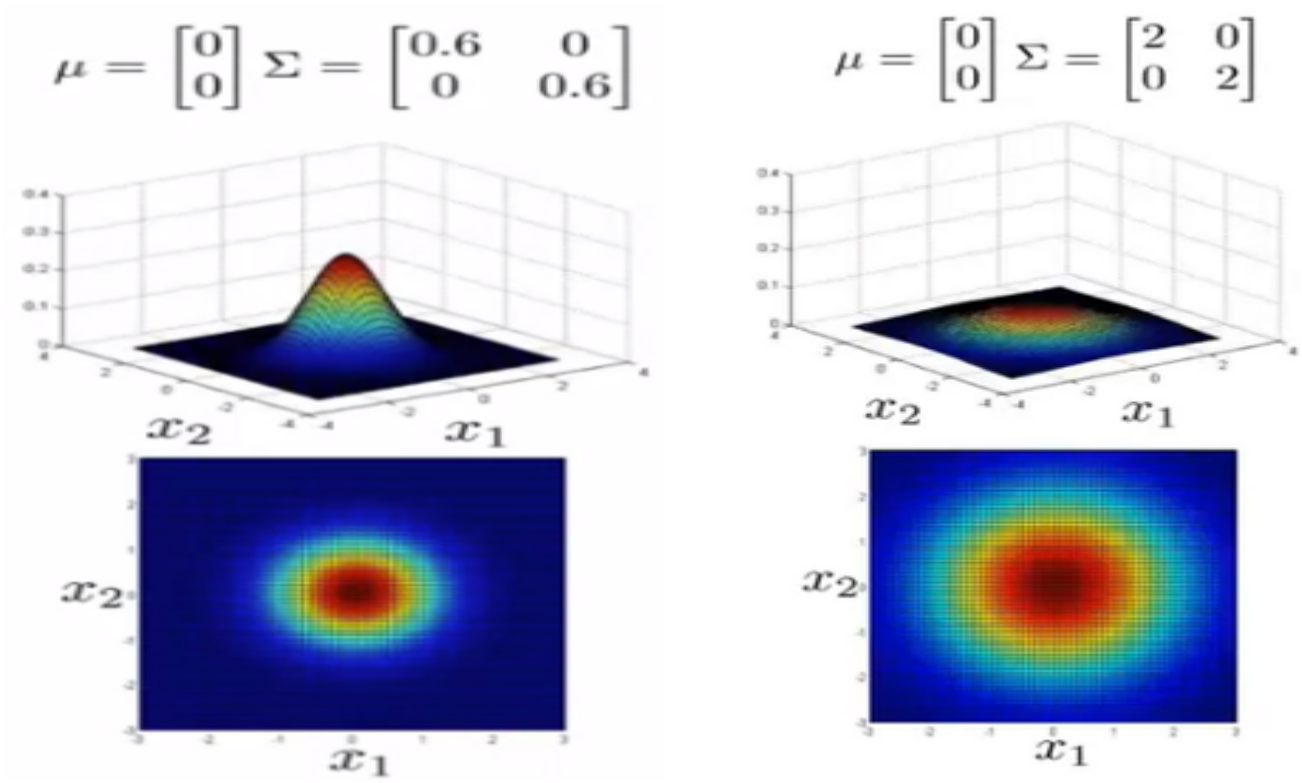

The picture represents a probability distribution of a multivariate Gaussian distribution where μ of both x1 and x2 are zeros.

Summation symbol is an identity matrix that contains sigma values as diagonals. The 1s in the diagonals are the sigma for both x1 and x2. And the zeros in the off diagonals show the correlation between x1 and x2. So, x1 and x2 are not correlated in this case.

In both x1 and x2 direction, the highest probability density is at 0 as the μ is zero.

The dark red color area in the center shows the highest probability density area. The probability density keeps going lower in the lighter red, yellow, green, and cyan areas. It's the lowest in the dark blue color zone.

**Changing the Standard Deviation - Sigma**




when the standard deviation sigma shrinks, the range also shrinks. At the same time, the height of the curve becomes higher to adjust the area.

In the contrast, when sigma is larger, the variability becomes wider. So, the height of the curve gets lower.

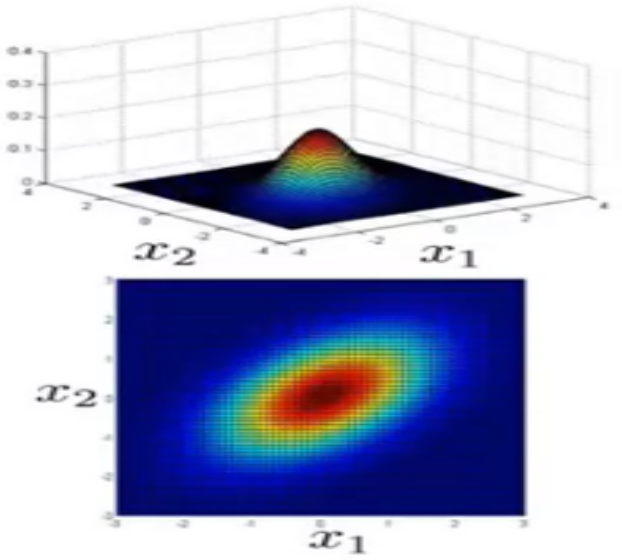
The sigma values for both x1 and x2 will not be the same always.

$$\mu = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \Sigma = \begin{bmatrix} 0.6 & 0 \\ 0 & 1 \end{bmatrix}$$

the range looks like an eclipse. It shrunk for the x1 as the standard deviation sigma is smaller for sigma.

**Change the Correlation Factor Between the Variables**

$$\mu = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \Sigma = \begin{bmatrix} 1 & 0.5 \\ 0.5 & 1 \end{bmatrix}$$




This is a completely different scenario. The off-diagonal values are not zeros anymore. It's 0.5. It shows that x1 and x2 are correlated by a factor of 0.5.

The eclipse has a diagonal direction now. x1 and x2 are growing together as they are positively correlated.

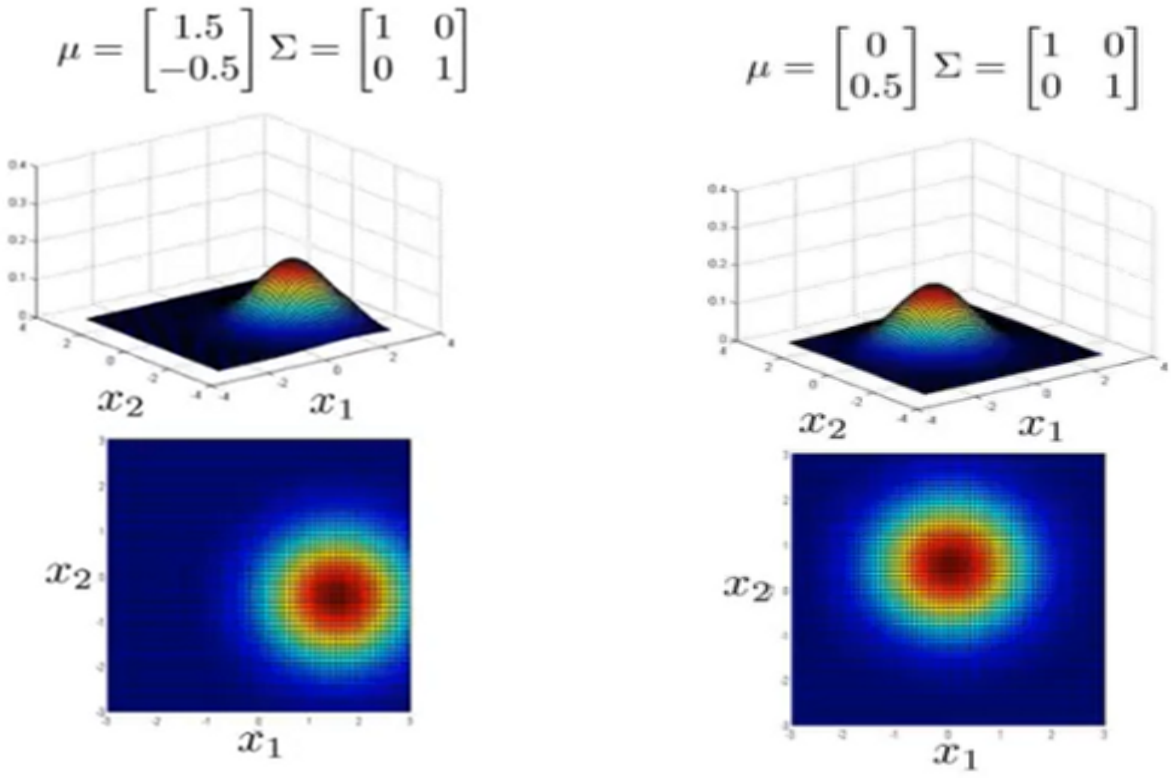
When x1 is large x2 also large and when x1 is small, x2 is also small.

**Different Means**

The center of the curve shifts from zero for x2 now.

The center position or the highest probability distribution area should be at 0.5 now.

The center of the highest probability in the x1 direction is 1.5. At the same time, the center of the highest probability is -0.5 for x2 direction.




### 1.10. Hypothesis Testing

Hypothesis testing is a part of statistical analysis, where we test the assumptions made regarding a population parameter.

It is generally used when we were to compare:

- a single group with an external standard
- two or more groups with each other

A Parameter is a number that describes the data from the population whereas, a Statistic is a number that describes the data from a sample.

**Terminologies**

**1.10.1. Null Hypothesis**: Null hypothesis is a statistical theory that suggests there is no statistical significance exists between the populations.

It is denoted by H0 and read as H-naught.

**1.10.2. Alternative Hypothesis**: An Alternative hypothesis suggests there is a significant difference between the population parameters. It could be greater or smaller. Basically, it is the contrast of the Null Hypothesis.

It is denoted by Ha or H1.

H0 must always contain equality(=). Ha always contains difference(≠, >, <).

For example, if we were to test the equality of average means (μ) of two groups:

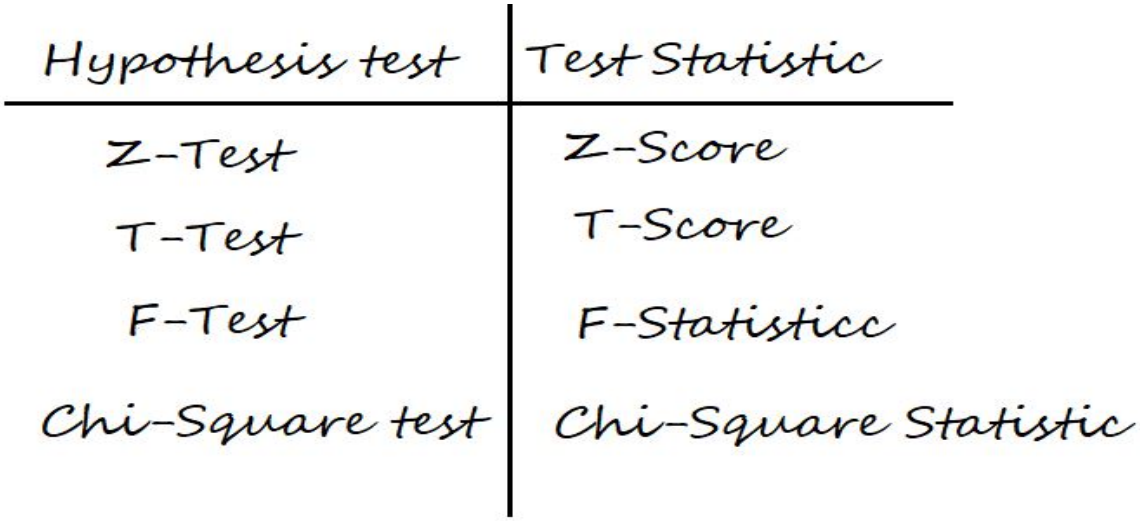
for a two-tailed test, we define H0: μ1 = μ2 and Ha: μ1≠μ2

for a one-tailed test, we define H0: μ1 = μ2 and Ha: μ1 > μ2 or Ha: μ1 < μ2

**1.10.3. Level of significance**: Denoted by alpha or α. It is a fixed probability of wrongly rejecting a True Null Hypothesis. For example, if α=5%, that means we are okay to take a 5% risk and conclude there exists a difference when there is no actual difference.

**1.10.4. Test Statistic**: It is denoted by t and is dependent on the test that we run. It is deciding factor to reject or accept Null Hypothesis.

The four main test statistics are given in the below table:




| Test Type | Distribution | Test Parameters |
|---|---|---|
| **Z-test** | Normal | Mean |
| **T-test** | Student-t | Mean |
| **ANOVA** | F distribution | Means |
| **Chi-Square** | Chi-squared distribution | Association between two categorical variables |

Each hypothesis test uses these basic principles.

| Element | Example | Description |
|---|---|---|
| **Hypothesis with hypothesized value** | $\mu > 45$ | The $H_o$ here is that the population mean is greater than 45 |
| **Test value** | 45 | This is used as a benchmark to test how likely a mean of 45 is given the population mean and SD |
| **Confidence interval** | $\alpha = 0.05$ | At the 95% confidence level (1-0.95 = 0.05), we can be certain that our test gets the true answer 95% of the time |
| **Test statistic** | $z\ score = 1.5$ | The test statistic gives you the standardized value of your test value on your test distribution |

| | | |
|---|---|---|
| **P-value** | $p - value = 0.06$ | The p-value is the calculated probability of your value occuring |

In hypothesis testing, the following rules are used to either reject or accept the hypothesis given a $\alpha$ of 0.05. Keep in mind that if you were to have an $\alpha$ of 0.1, you're results would be given with **90% confidence** and the example above, with a p-value of 0.06, would reject $H_o$.

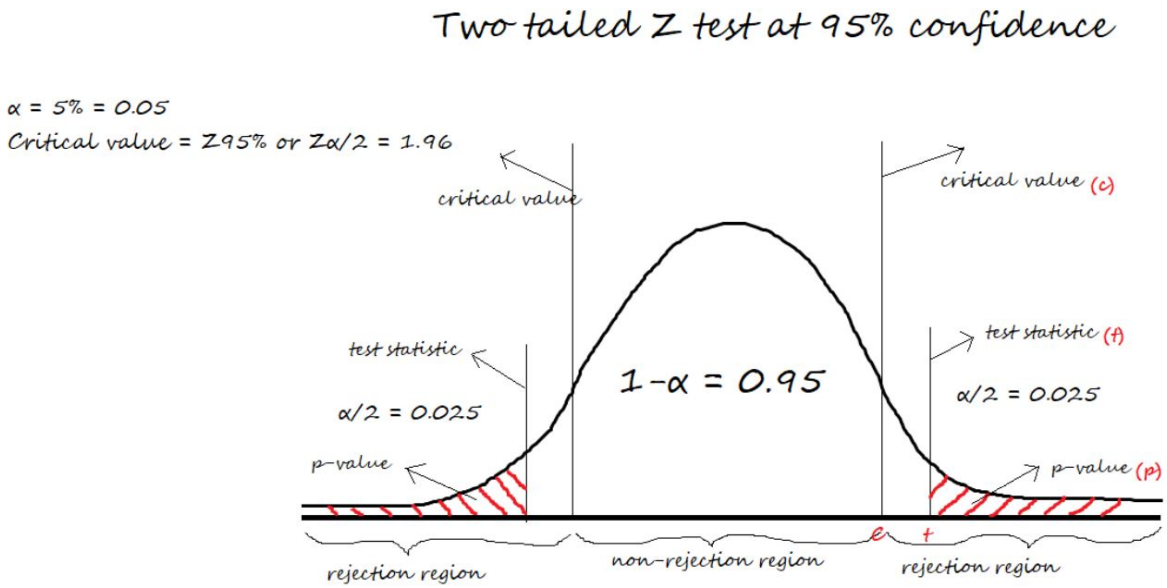
| | | |
|---|---|---|
| **P-value < 0.05** | Region of rejection | Reject $H_o$ |
| **P-value > 0.05** | Region of acceptance | Fail to reject $H_o$ |

**p-value**: It is the proportion of samples (assuming the Null Hypothesis is true) that would be as extreme as the test statistic. It is denoted by the letter p.

Now, assume we are running a two-tailed Z-Test at 95% confidence. Then, the level of significance (α) = 5% = 0.05. Thus, we will have (1-α) = 0.95 proportion of data at the center, and α = 0.05 proportion will be equally shared to the two tails. Each tail will have (α/2) = 0.025 proportion of data.

The critical value i.e., Z95% or Zα/2 = 1.96 is calculated from the Z-scores table.

Now, take a look at the below figure for a better understanding of critical value, test-statistic, and p-value.

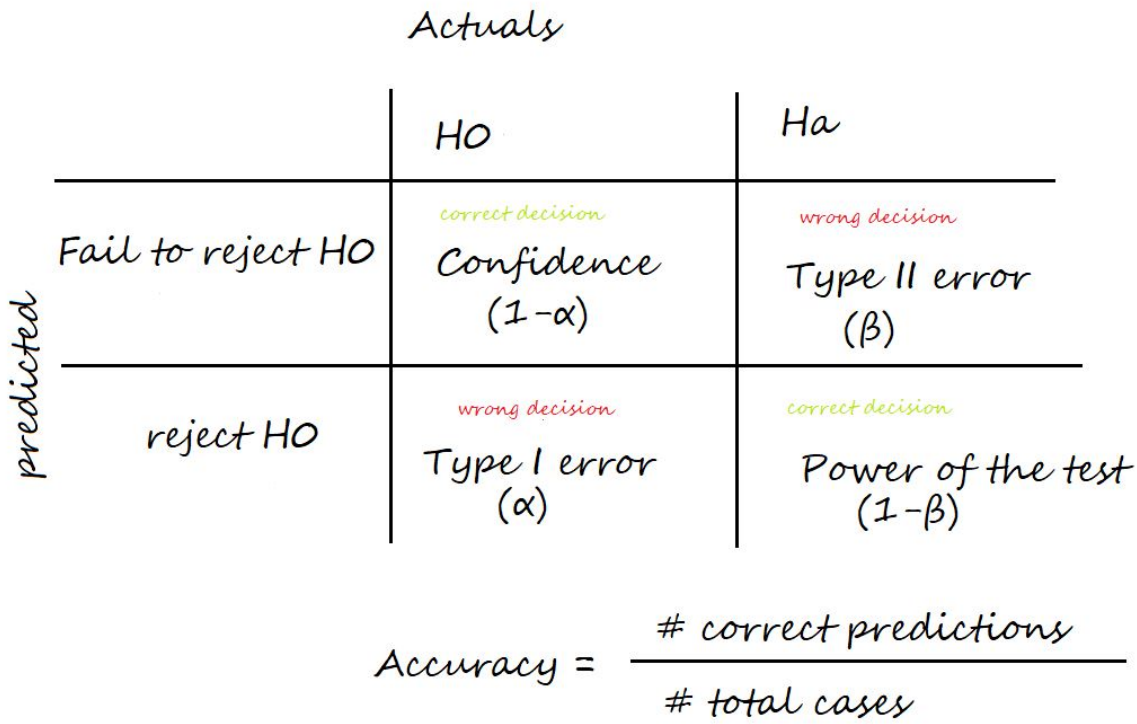



**Steps of Hypothesis testing**

For a given business problem,

1. Start with specifying Null and Alternative Hypotheses about a population parameter
2. Set the level of significance (α)
3. Collect Sample data and calculate the Test Statistic and P-value by running a Hypothesis test that well suits our data
4. Make Conclusion: Reject or Fail to Reject Null Hypothesis
5. Confusion Matrix in Hypothesis testing

To plot a confusion matrix, we can take actual values in columns and predicted values in rows or vice versa.




$$\text{Accuracy} = \frac{\#\ \text{correct predictions}}{\#\ \text{total cases}}$$

**Confidence**: The probability of accepting a True Null Hypothesis. It is denoted as (1-α)

**Power of test**: The probability of rejecting a False Null Hypothesis i.e., the ability of the test to detect a difference. It is denoted as (1-β) and its value lies between 0 and 1.

**Type I error**: Occurs when we reject a True Null Hypothesis and is denoted as α.

**Type II error:** Occurs when we accept a False Null Hypothesis and is denoted as β.

**Accuracy**: Number of correct predictions / Total number of cases

The factors that affect the power of the test are sample size, population variability, and the confidence (α).

Confidence and power of test are directly proportional. Increasing the confidence increases the power of the test.

Type 1 and 2 errors occur when we **reject** or **accept** our null hypothesis when, in reality, we shouldn't have. This happens because, while statistics is powerful, there is a certain chance that you may be wrong. The table below summarizes these types of errors.

| | **Accept** $H_o$ | **Reject** $H_o$ |
|---|---|---|
| **In reality,** $H_o$ **is actually true** | Correct: $H_o$ is true and statistical test accepts $H_o$ | Incorrect: Type 1 error - $H_o$ is true and statistical test rejects $H_o$ |
| **In reality,** $H_o$ **is actually false** | Incorrect: Type 2 error - $H_o$ is false and statistical test accepts $H_o$ | Correct: $H_o$ is false and statistical test rejects $H_o$ |

**Confidence Interval**

A confidence interval, in statistics, refers to the probability that a population parameter will fall between a set of values for a certain proportion of times.

A confidence interval is the mean of your estimate plus and minus the variation in that estimate. This is the range of values you expect your estimate to fall between if you redo your test, within a certain level of confidence.

Confidence, in statistics, is another way to describe probability. For example, if you construct a confidence interval with a 95% confidence level, you are confident that 95 out of 100 times the estimate will fall between the upper and lower values specified by the confidence interval.

The desired confidence level is usually one minus the alpha ( a ) value you used in the statistical test:

**Confidence level = 1 − a**

So if you use an alpha value of $p < 0.05$ for statistical significance, then your confidence level would be $1 − 0.05 = 0.95$, or 95%.

When to use confidence intervals?

Confidence intervals can be calculated for many kinds of statistical estimates, including:

- Proportions
- Population means
- Differences between population means or proportions
- Estimates of variation among groups

These are all point estimates, and don't give any information about the variation around the number. Confidence intervals are useful for communicating the variation around a point estimate.

**Example: Variation around an estimate**

You survey 100 Brits and 100 Americans about their television-watching habits, and find that both groups watch an average of 35 hours of television per week.

However, the British people surveyed had a wide variation in the number of hours watched, while the Americans all watched similar amounts.

Even though both groups have the same point estimate (average number of hours watched), the British estimate will have a wider confidence interval than the American estimate because there is more variation in the data.

**Calculating a confidence interval**

Most statistical programs will include the confidence interval of the estimate when you run a statistical test.

If you want to calculate a confidence interval on your own, you need to know:

- The point estimate you are constructing the confidence interval for
- The critical values for the test statistic
- The standard deviation of the sample
- The sample size

Once you know each of these components, you can calculate the confidence interval for your estimate by plugging them into the confidence interval formula that corresponds to your data.

**Point estimate**

The point estimate of your confidence interval will be whatever statistical estimate you are making (e.g. population mean, the difference between population means, proportions, variation among groups).

**Example**: Point estimate - In the TV-watching example, the point estimate is the mean number of hours watched: 35.

**Finding the critical value**

Critical values tell you how many standard deviations away from the mean you need to go in order to reach the desired confidence level for your confidence interval.

There are three steps to find the critical value.

1. **Choose your alpha ( a ) value.**

The alpha value is the probability threshold for statistical significance. The most common alpha value is p = 0.05, but 0.1, 0.01, and even 0.001 are sometimes used. It's best to look at the papers published in your field to decide which alpha value to use.

2. **Decide if you need a one-tailed interval or a two-tailed interval.**

You will most likely use a two-tailed interval unless you are doing a one-tailed t-test.

For a two-tailed interval, divide your alpha by two to get the alpha value for the upper and lower tails.

3. **Look up the critical value that corresponds with the alpha value.**

If your data follows a normal distribution, or if you have a large sample size ($n > 30$) that is approximately normally distributed, you can use the z-distribution to find your critical values.

For a z-statistic, some of the most common values are shown in this table:

| Confidence level | 90% | 95% | 99% |
|---|---|---|---|
| alpha for one-tailed CI | 0.1 | 0.05 | 0.01 |
| alpha for two-tailed CI | 0.05 | 0.025 | 0.005 |
| z-statistic | 1.64 | 1.96 | 2.57 |

If you are using a small dataset ($n \leq 30$) that is approximately normally distributed, use the t-distribution instead.

The t-distribution follows the same shape as the z-distribution, but corrects for small sample sizes. For the t-distribution, you need to know your degrees of freedom (sample size minus 1).

Check out this set of t tables to find your t-statistic. The author has included the confidence level and p-values for both one-tailed and two-tailed tests to help you find the t-value you need.

For normal distributions, like the t-distribution and z-distribution, the critical value is the same on either side of the mean.

**Example: Critical value** In the TV-watching survey, there are more than 30 observations and the data follow an approximately normal distribution (bell curve), so we can use the z-distribution for our test statistics.

For a two-tailed 95% confidence interval, the alpha value is 0.025, and the corresponding critical value is 1.96.

This means that to calculate the upper and lower bounds of the confidence interval, we can take the mean ±1.96 standard deviations from the mean.

**Finding the standard deviation**

Most statistical software will have a built-in function to calculate your standard deviation, but to find it by hand you can first find your sample variance, then take the square root to get the standard deviation.

**1.Find the sample variance**

Sample variance is defined as the sum of squared differences from the mean, also known as the mean-squared-error (MSE):

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

To find the MSE, subtract your sample mean from each value in the dataset, square the resulting number, and divide that number by $n - 1$ (sample size minus 1).

Then add up all of these numbers to get your total sample variance (s2). For larger sample sets, it's easiest to do this in Excel.

**2.Find the standard deviation.**

The standard deviation of your estimate (s) is equal to the square root of the sample variance/sample error (s2):

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

**Example: Standard deviation** In the television-watching survey, the variance in the GB estimate is 100, while the variance in the USA estimate is 25. Taking the square root of the variance gives us a sample standard deviation (s) of:

10 for the GB estimate.

5 for the USA estimate.

**Sample size**

The sample size is the number of observations in your data set.

**Example: Sample size** In our survey of Americans and Brits, the sample size is 100 for each group.

**Confidence interval for the mean of normally-distributed data**

Normally-distributed data forms a bell shape when plotted on a graph, with the sample mean in the middle and the rest of the data distributed fairly evenly on either side of the mean.

The confidence interval for data which follows a standard normal distribution is:

$$CI = \overline{X} \pm Z^* \frac{\sigma}{\sqrt{n}}$$

Where:

CI = the confidence interval

$\bar{X}$ = the population mean

Z* = the critical value of the z-distribution

σ = the population standard deviation

√n = the square root of the population size

The confidence interval for the t-distribution follows the same formula, but replaces the Z* with the t*.

In real life, you never know the true values for the population (unless you can do a complete census). Instead, we replace the population values with the values from our sample data, so the formula becomes:

$$CI = \hat{x} \pm Z^* \frac{s}{\sqrt{n}}$$

Where:

ˆx = the sample mean

s = the sample standard deviation

**Example: Calculating the confidence interval-** In the survey of Americans' and Brits' television watching habits, we can use the sample mean, sample standard deviation, and sample size in place of the population mean, population standard deviation, and population size.

To calculate the 95% confidence interval, we can simply plug the values into the formula.

For the USA:

$$\begin{aligned} CI &= 35 \pm 1.96 \frac{5}{\sqrt{100}} \\ &= 35 \pm 1.96(0.5) \\ &= 35 \pm 0.98 \end{aligned}$$

So for the USA, the lower and upper bounds of the 95% confidence interval are 34.02 and 35.98.

For GB:

$$\begin{aligned} CI &= 35 \pm 1.96 \frac{10}{\sqrt{100}} \\ &= 35 \pm 1.96(1) \\ &= 35 \pm 1.96 \end{aligned}$$

So for the GB, the lower and upper bounds of the 95% confidence interval are 33.04 and 36.96.

1. The confidence 'level' refers to the long term success rate of the method i.e. how often this type of interval will capture the parameter of interest.

2. A specific confidence interval gives a range of plausible values for the parameter of interest.

3. A larger margin of error produces a wider confidence interval that is more likely to contain the parameter of interest(increased confidence)

4. Increasing the confidence will increase the margin of error resulting in a wider interval.