What is Data Science?

Data science is a deep study of the massive amount of data, which involves extracting meaningful insights from raw, structured, and unstructured data that is processed using the scientific method, different technologies, and algorithms.

Data science uses the most powerful hardware, programming systems, and most efficient algorithms to solve the data related problems. It is the future of artificial intelligence.

In short, we can say that data science is all about:

* Asking the correct questions and analyzing the raw data.
* Modeling the data using various complex and efficient algorithms.
* Visualizing the data to get a better perspective.
* Understanding the data to make better decisions and finding the final result.

## Need for Data Science:

Some years ago, data was less and mostly available in a structured form, which could be easily stored in excel sheets, and processed using BI tools.

But in today's world, data is becoming so vast, i.e., approximately **2.5 quintals bytes** of data is generating on every day, which led to data explosion. It is estimated as per researches, that by 2020, 1.7 MB of data will be created at every single second, by a single person on earth. Every Company requires data to work, grow, and improve their businesses.

Now, handling of such huge amount of data is a challenging task for every organization. So to handle, process, and analysis of this, we required some complex, powerful, and efficient algorithms and technology, and that technology came into existence as data Science. Following are some main reasons for using data science technology:

* With the help of data science technology, we can convert the massive amount of raw and unstructured data into meaningful insights.
* Data science technology is opting by various companies, whether it is a big brand or a startup. Google, Amazon, Netflix, etc, which handle the huge amount of data, are using data science algorithms for better customer experience.
* Data science is working for automating transportation such as creating a self-driving car, which is the future of transportation.
* Data science can help in different predictions such as various survey, elections, flight ticket confirmation, etc.

## Data science Jobs:

1. Data Scientist
2. Data Analyst
3. Machine learning expert
4. Data engineer
5. Data Architect
6. Data Administrator
7. Business Analyst
8. Business Intelligence Manager

## Difference between BI (business intelligence) and Data Science

BI stands for business intelligence, which is also used for data analysis of business information: Below are some differences between BI and Data sciences:

**Business intelligence**

Business intelligence deals with structured data, e.g., data warehouse

Analytical (historical data)

Statistics and Visualization are the two skills required for business intelligence.

Business intelligence focuses on both Past and present data

**Data science**

Data science deals with structured and unstructured data, e.g., weblogs, feedback, etc.

Scientific (goes deeper to know the reason for the data report)

Statistics, Visualization, and Machine learning are the required skills for data science.

Data science focuses on past data, present data, and also future predictions.

What is statistics?

Statistics is the study of the collection, analysis, interpretation, presentation, and organization of data. In other words, it is a mathematical discipline to collect, summarize data. Also, we can say that statistics is a branch of applied mathematics.

**Statistics Definition:** Statistics is a branch that deals with every aspect of the data. Statistical knowledge helps to choose the proper method of collecting the data and employ those samples in the correct analysis process in order to effectively produce the results. In short, [statistics](https://byjus.com/maths/statistics/) is a crucial process which helps to make the decision based on the data.

**Statistics Example**

An example of statistical analysis is when we have to determine the number of people in a town who watch TV out of the total population in the town. The small group of people is called the sample here, which is taken from the population.

**Types of Statistics**

The two main branches of statistics are:

* Descriptive Statistics
* Inferential Statistics

**Descriptive Statistics** – Through graphs or tables, or numerical calculations, descriptive statistics uses the data to provide descriptions of the population.

**Inferential Statistics** – Based on the data sample taken from the population, inferential statistics makes the predictions and inferences.

**Importance of Statistics**

The important functions of statistics are:

* Statistics helps in gathering information about the appropriate quantitative data
* It depicts the complex data in graphical form, tabular form and in diagrammatic representation to understand it easily
* It provides the exact description and a better understanding
* It helps in designing the effective and proper planning of the statistical inquiry in any field
* It gives valid inferences with the reliability measures about the population parameters from the sample data
* It helps to understand the variability pattern through the quantitative observations

**Why statistics for data science**

Statistics serve as the backbone of data science, providing the necessary tools and techniques to extract meaningful insights from data.

Data science relies heavily on statistical methods to extract meaningful insights from data, while statistics find practical applications and gain new perspectives through data science. Both disciplines are interconnected, with statistics providing the foundational principles and techniques that enable data scientists to analyze and interpret data effectively.

Basic Statistical Concepts

Mean, Median, and Mode

What can we learn from looking at a group of numbers?

In Machine Learning (and in mathematics) there are often three values that interests us:

* **Mean** - The average value
* **Median** - The mid point value
* **Mode** - The most common value

Example: We have registered the speed of 13 cars:

speed = [99,86,87,88,111,86,103,87,94,78,77,85,86]

What is the average, the middle, or the most common speed value?

Mean

The mean value is the average value.

To calculate the mean, find the sum of all values, and divide the sum by the number of values:

(99+86+87+88+111+86+103+87+94+78+77+85+86) / 13 = 89.77

The NumPy module has a method for this. Learn about the NumPy module in our [NumPy Tutorial](https://www.w3schools.com/python/numpy/default.asp).

Use the NumPy mean() method to find the average speed:

import numpy  
  
speed = [99,86,87,88,111,86,103,87,94,78,77,85,86]  
  
x = numpy.mean(speed)  
  
print(x)

Median

The median value is the value in the middle, after you have sorted all the values:

77, 78, 85, 86, 86, 86, 87, 87, 88, 94, 99, 103, 111

It is important that the numbers are sorted before you can find the median.

The NumPy module has a method for this:

Example

Use the NumPy median() method to find the middle value:

import numpy  
  
speed = [99,86,87,88,111,86,103,87,94,78,77,85,86]  
  
x = numpy.median(speed)  
  
print(x)

If there are two numbers in the middle, divide the sum of those numbers by two.

77, 78, 85, 86, 86, 86, 87, 87, 94, 98, 99, 103  
  
(86 + 87) / 2 = 86.5

Example

Using the NumPy module:

import numpy  
  
speed = [99,86,87,88,86,103,87,94,78,77,85,86]  
  
x = numpy.median(speed)  
  
print(x)

Mode

The Mode value is the value that appears the most number of times:

99, 86, 87, 88, 111, 86, 103, 87, 94, 78, 77, 85, 86 = 86

The SciPy module has a method for this. Learn about the SciPy module in our [SciPy Tutorial](https://www.w3schools.com/python/scipy_intro.asp).

Example

Use the SciPy mode() method to find the number that appears the most:

from scipy import stats  
  
speed = [99,86,87,88,111,86,103,87,94,78,77,85,86]  
  
x = stats.mode(speed)  
  
print(x)

**--------------------------------------------------------**

## What are Percentiles?

Percentiles are used in statistics to give you a number that describes the value that a given percent of the values are lower than.

Example: Let's say we have an array of the ages of all the people that live in a street.

ages = [5,31,43,48,50,41,7,11,15,39,80,82,32,2,8,6,25,36,27,61,31]

What is the 75. percentile? The answer is 43, meaning that 75% of the people are 43 or younger.

The NumPy module has a method for finding the specified percentile:

### Example[Get your own Python Server](https://www.w3schools.com/python/python_server.asp)

Use the NumPy percentile() method to find the percentiles:

import numpy  
  
ages = [5,31,43,48,50,41,7,11,15,39,80,82,32,2,8,6,25,36,27,61,31]  
  
x = numpy.percentile(ages, 75)  
  
print(x)

### Example

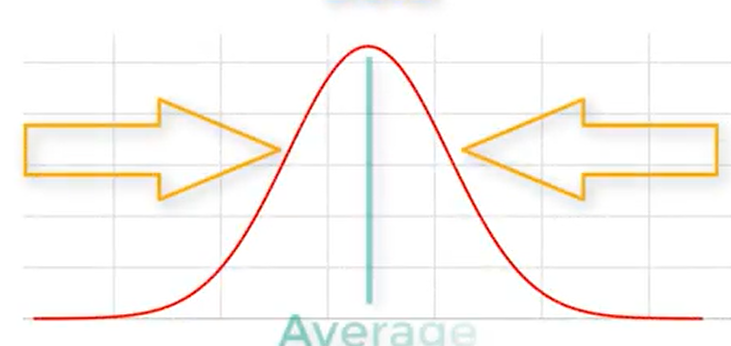
What is the age that 90% of the people are younger than?

import numpy  
  
ages = [5,31,43,48,50,41,7,11,15,39,80,82,32,2,8,6,25,36,27,61,31]  
  
x = numpy.percentile(ages, 90)  
  
print(x)

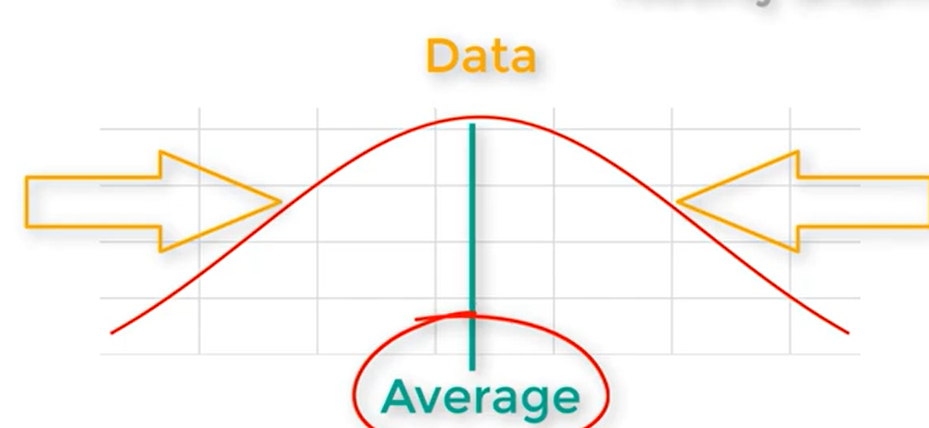
## What is Standard Deviation?

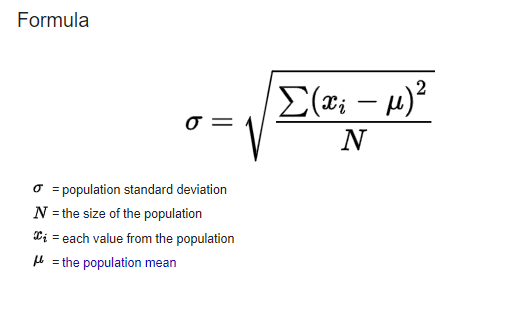
Standard deviation is a number that describes how spread out the values are.

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



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





**Why Use Standard deviation?**

Standard deviation is a useful statistic for a variety of reasons, including:

* Understanding data spread

Standard deviation is a numerical measure of how spread out data points are around the mean. It's useful when variability is as important as the average.

Example: This time we have registered the speed of 7 cars:

speed = [86,87,88,86,87,85,86]

The standard deviation is:

0.9

Meaning that most of the values are within the range of 0.9 from the mean value, which is 86.4.

Let us do the same with a selection of numbers with a wider range:

speed = [32,111,138,28,59,77,97]

The standard deviation is:

37.85

Meaning that most of the values are within the range of 37.85 from the mean value, which is 77.4.

As you can see, a higher standard deviation indicates that the values are spread out over a wider range.

The NumPy module has a method to calculate the standard deviation:

### Example[Get your own Python Server](https://www.w3schools.com/python/python_server.asp)

Use the NumPy std() method to find the standard deviation:

import numpy  
  
speed = [86,87,88,86,87,85,86]  
  
x = numpy.std(speed)  
  
print(x)

### Example

import numpy  
  
speed = [32,111,138,28,59,77,97]  
  
x = numpy.std(speed)  
  
print(x)

## Variance

Variance is another number that indicates how spread out the values are.

In fact, if you take the square root of the variance, you get the standard deviation!

Or the other way around, if you multiply the standard deviation by itself, you get the variance!

To calculate the variance you have to do as follows:

1. Find the mean:

(32+111+138+28+59+77+97) / 7 = 77.4

2. For each value: find the difference from the mean:

 32 - 77.4 = -45.4  
111 - 77.4 =  33.6  
138 - 77.4 =  60.6  
 28 - 77.4 = -49.4  
 59 - 77.4 = -18.4  
 77 - 77.4 = - 0.4  
 97 - 77.4 =  19.6

3. For each difference: find the square value:

(-45.4)2 = 2061.16  
 (33.6)2 = 1128.96  
 (60.6)2 = 3672.36  
(-49.4)2 = 2440.36  
(-18.4)2 =  338.56  
(- 0.4)2 =    0.16  
 (19.6)2 =  384.16

4. The variance is the average number of these squared differences:

(2061.16+1128.96+3672.36+2440.36+338.56+0.16+384.16) / 7 = 1432.2

Luckily, NumPy has a method to calculate the variance:

### Example

Use the NumPy var() method to find the variance:

import numpy  
  
speed = [32,111,138,28,59,77,97]  
  
x = numpy.var(speed)  
  
print(x)

**output - 1432.2448979591834**

## Standard Deviation

As we have learned, the formula to find the standard deviation is the square root of the variance:

√1432.25 = 37.85

Or, as in the example from before, use the NumPy to calculate the standard deviation:

### Example

Use the NumPy std() method to find the standard deviation:

import numpy  
  
speed = [32,111,138,28,59,77,97]  
  
x = numpy.std(speed)  
  
print(x)

## Symbols

Standard Deviation is often represented by the symbol Sigma: σ

Variance is often represented by the symbol Sigma Squared: σ2

**--------------------------------------------------------**

**Skewness Formula** एˈसिमट्रि( असममिति)

Skewness is a measure of the asymmetry of a distribution

The measure of skewness tells us the direction and the extent of skewness. In symmetrical distribution the mean, median, and mode are identical. the more the mean moves away from the mode, the larger the asymmetric or skewness.

Before learning let’s learn more about Mean, Median, and Mode first.

**Mean**

Mean is the average of the numbers in the data distribution, It is calculated by adding up all the values in the dataset and dividing the sum by the number of values in the dataset.

***Mean= Sum of all values in Dataset / Total number of values***

**Example: Find the mean of a dataset of exam scores: 70, 80, 85, 90, and 95.**

**Solution:**

*Mean = (70 + 80 + 85 + 90 + 95) / 5 = 84*

*So the mean of this dataset is 84.*

**Median**

When arranging all the data in order (ascending and descending) the comes in the middle of the data is called the median.

**Median is the middle value of a dataset when the values are arranged in order from smallest to largest.**

**Examples for Odd Numbers in the Dataset**

**Example 1: Find the median of a dataset of exam scores: 70, 85, 80, 95, 90**

**Solution:**

*Firstly arrange all no. in order from smallest to largest: 70, 80, 85, 90, 95.*

*The mid value is 85. so, the median is 85.*

**Example 2: Find the median of a dataset: 5, 10, 15, 20, 25.**

**Solution:**

*Firstly arrange all no. in order from smallest to largest: 5, 10, 15, 20, 25.*

*The mid value is 15. so, the median is 15.*

**If there are an even number of values in the dataset, the median is calculated by taking the average of the two middle values.**

**Examples for Even Numbers in the Dataset**

**Example 1: Find the median of a dataset of exam scores: 70, 80, 85, 90.**

**Solution:**

*The median is calculated as (80 + 85) / 2 = 82.5*

*So the median of this dataset is 82.5.*

**Example 2: Find the median of a dataset: 2, 4, 6, 8, 10, 12.**

**Solution:**

*Firstly, we need to find the middle two numbers. So, 6, and 8 are mid values of the dataset*

*Median = (6 + 8) / 2 = 7*

*So the median of this dataset is 7.*

**Mode**

most frequently used number in data is called the mode of the data.

**Example 1: We have a data set representing the number of pets owned by 10 people: 3, 1, 0, 2, 1, 1, 4, 2, 2, 1. Find the mode.**

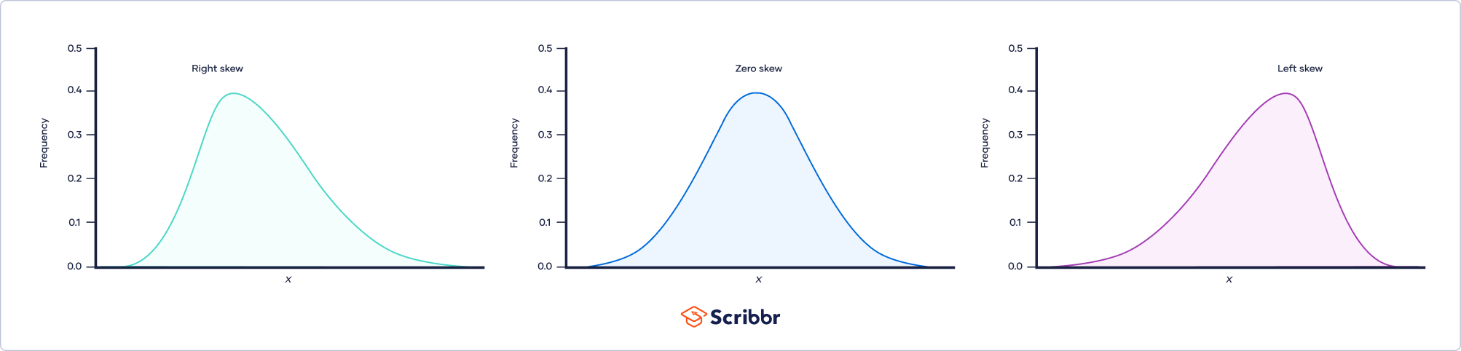
**Solution:**

So, the value that appears most frequently in the data set is 1. the value 1 appears four times. Therefore, the mode of this data set is 1.

**Skewness**

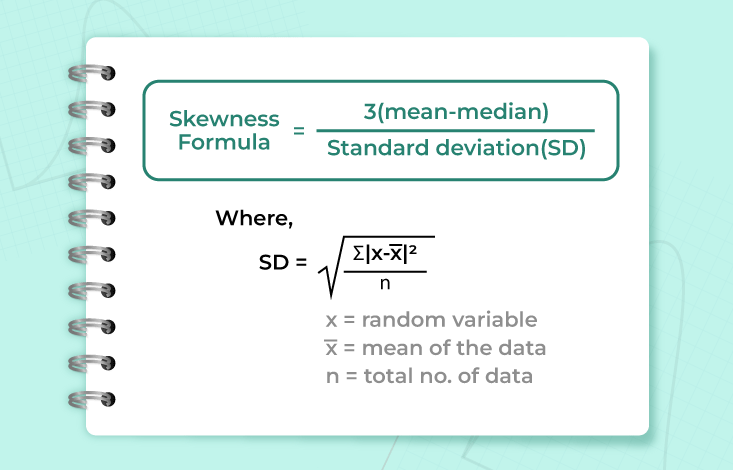
 is a measure of the asymmetry of a distribution. A distribution is asymmetrical when its left and right side are not mirror images.

A distribution can have right (or positive), left (or negative), or zero skewness. A right-skewed distribution is longer on the right side of its peak, and a left-skewed distribution is longer on the left side of its peak:



**Skewness Formula**

The skewness formula is discussed in the image below,



**Type of Skewness**

Various types of skewness used in mathematics are,

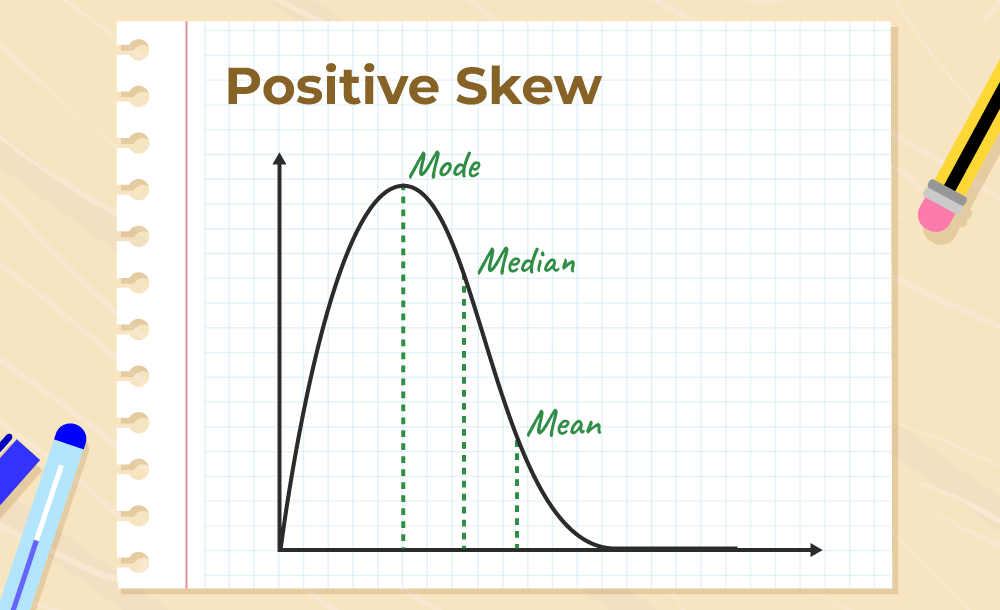
* Positive Skewness
* Negative Skewness
* Zero Skewness

**Positive Skewness**

Positive Skewness means the tail on the right side of the distribution is longer. The mean and median will be greater than the mode.

Condition for positive skewness = **Mean > Median >Mode**

The positive curve of skewness is shown in the image below,



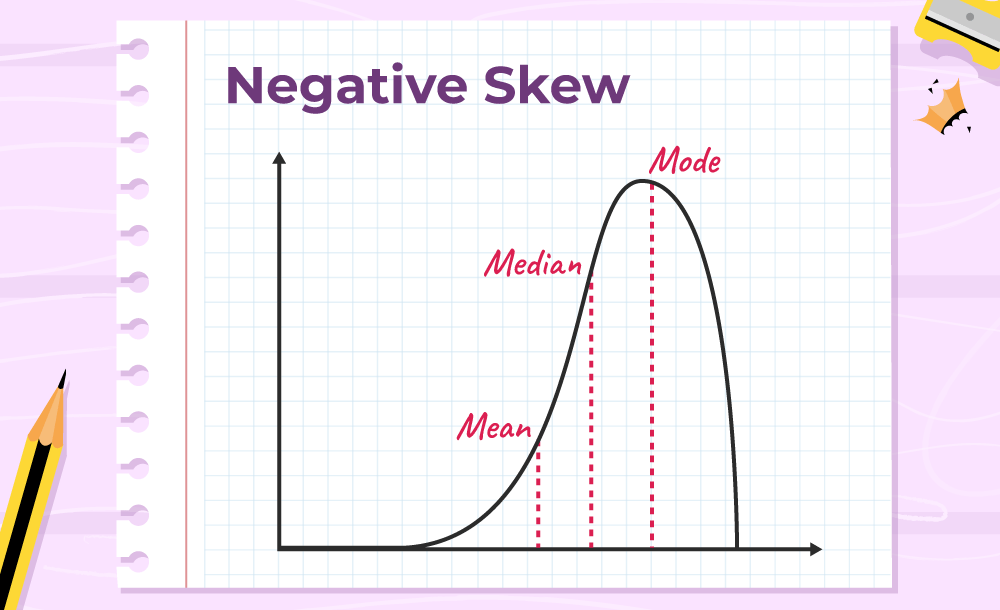
Let’s take an example of the income distribution where a few people earn very high incomes and the majority earn lower incomes. so, this is often positively skewed. Analyzing skewed data can provide valuable insights into the underlying causes and potential solutions or interventions.

**Negative Skewness**

Negative Skewness means when the tail of the left side of the distribution is longer than the tail on the right side. The mean and median will be less than the mode.

Condition for negative skewness is **Mode > Median > Mean**

The curve shows negative skewness in the image below,



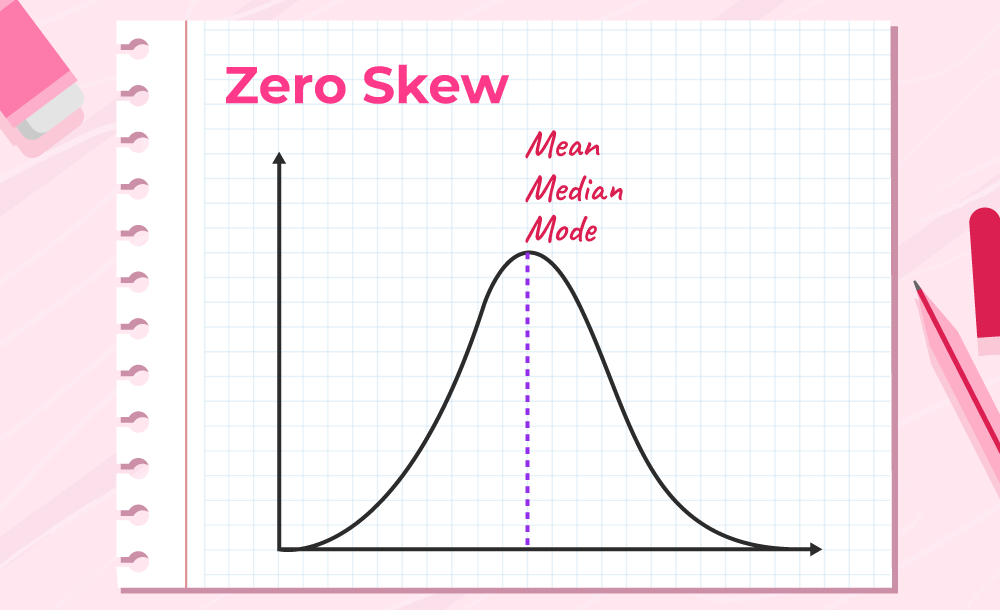
Let’s take an example of a match, during the match most of the players of a particular team scored runs above 50 and only a few of them scored below 10. In such a case, the data is generally represented with the help of a negatively skewed distribution. And this data is helpful to analyze the game’s performance.

**Zero Skewness**

It is also known as a “symmetric distribution”.It signifies that distribution of data is evenly distributed around the mean, with no long tails on either end of the distribution

Condition for zero skewness is **Mean = Mode = Median**

The curve for zero skews is shown in the image below,

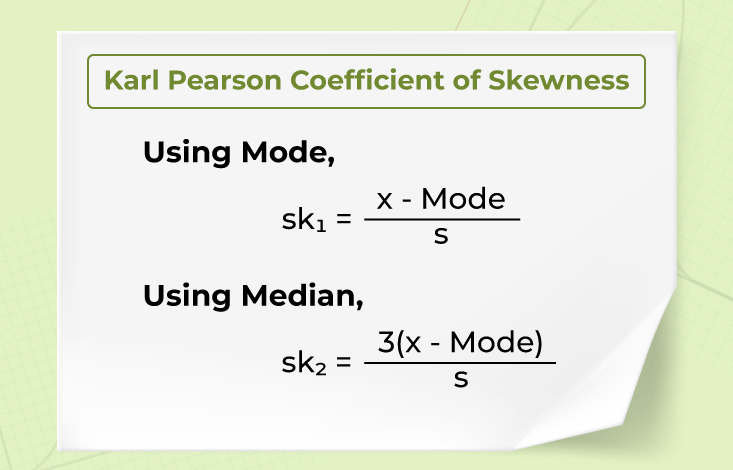


**Methods to Measure Skewness**

Skewness can be measured using Karl Pearson’s Coefficient of Skewness.

**Karl Pearson’s Co-efficient of Skewness**

The formula for measuring Skewness using Karl Pearson’s Co-efficient is discussed below in the image,



**Conditions**

* Mean = Mode = Median, then the coefficient of skewness is zero for symmetrical distribution.
* Mean > Mode, then the coefficient of skewness will be positive.
* Mean < Mode, then the coefficient of skewness will be negative.

Karl person`s coefficient of skewness has a positive sign for the positively skewed and a negative sign for the negatively skewed.

**Solved Examples on Skewness Formula**

**Example 1: Find the skewness for the given Data ( 2,4,6,6)**

**Solution:**

*Mean of Data = (2 + 4 + 6 + 6) / 4*

*= 18 / 4*

*= 4.5*

*Number of terms (n) = 4 (even)*

*Median of Data = {[n / 2]th + [n / 2 + 1]th}/2 term*

*= [(4 /2)th term + (4/2 +1)th term] / 2*

*= [2ndterm + 3rd term] / 2*

*= [4+6]/2*

*= 10/2*

*Median of Data  = 5*

*Mode of Data = Highest Frequency term = 6 (frequency 2)*

*S.D. = √[(2 – 5)2 + (4-5)2 + (6-5)2 + (6-5)2/4]*

*= √[(9 + 1 + 1 + 1)/4]*

*= √(3)*

*= 1.732*

***Skewness = 3(Mean – Median)/S.D.***

*By Applying Skewness Formula,*

*Skewness = 3(4.5 – 5)/1.732*

*= 3(-0.5)/ 1.732*

*Skewness = – 0.866*

*So, the skewness of these data is negative.*

**Example 2: A boy collects some rupees in a week as follows (25,28,26,30,40,50,40) and finds the skewness of the given Data in question with the help of the skewness formula.**

**Solution:**

*Mean of Data = (25+28+26+30+40+50+40) / 7*

*= 239 / 7*

*= 34.14*

*Number of terms (n) =7 (odd)*

*Arrange Data in ascending order = 25,26 ,28,30,40,40,50*

*The median of data is = 30*

*Mode of Data = Highest Frequency term = 40 (frequency 2)*

*S.D       = √(1/7 – 1) x ((25 – 34.1429)2 + (28 – 34.1429)2 + (26 – 34.1429)2 + (30 – 34.1429)2 + (40 – 34.1429)2 +(534.1429)2 + (40 – 34.1429)2)  
           = √(1/6) x ((-9.1429)2 + (-6.1429)2 + (-8.1429)2 + (-4.1429)2 + (5.8571)2 + (15.8571)2 + (5.8571)2)  
           = √(0.1667) x ((83.5926) + (37.7352) + (66.3068) + (17.1636) + (34.3056) + (251.4476) + (34.3056))  
           = √(0.1667) x 524.8571  
           = √87.4762  
         . = 9.3529*

***Skewness = 3(Mean – Median)/S.D.***

*By Applying Skewness Formula,*

*Skewness = 3(34.14 – 30)/9.3529*

*= 1.32*

*Skewness = 1.32*

*So skewness for these data is positive*

**Example 3: Attendance of all classes of a school are as follows find their skewness?**

**1st (35), 2nd(32), 3rd(38), 4th(39), 5th(43)**

| **Class Name** | **Number of students** |
| --- | --- |
| 1 st | 35 |
| 2 nd | 32 |
| 3 rd | 38 |
| 4 th | 39 |
| 5 th | 45 |

**Solution:**

*Mean of Data =  (35 + 32 + 38 + 39 + 42)/5*

*= 186/5*

*= 37.2*

*Number of terms (n) = 5 (odd)*

*Arrange Data in ascending order = 32,35,38,39,42*

*Median of Data  = 38*

*S.D. = √(1/5 – 1) x ((35 – 37.2)2 + (32 – 37.2)2 + (38 – 37.2)2 + (39 – 37.2)2 + (42 – 37.2)2)  
       = √(1/4) x ((-2.2)2 + (-5.2)2 + (0.8)2 + (1.8)2 + (4.8)2)  
       = √(0.25) x ((4.84) + (27.04) + (0.64) + (3.24) + (23.04))  
       = √(0.25) x 58.8  
       = √14.7  
       = 3.8341*

***Skewness = ∑(yi– ymean) / (n – 1) x (sd)³***

*Skewness =((35 – 37.2)³ + (32 – 37.2)³ + (38 – 37.2)³ + (39 – 37.2)³ + (42 – 37.2)³) / (5 – 1)³ x 3.8341*

*Skewness = ((-2.2)³ + (-5.2)³ + (0.8)³ + (1.8)³ + (4.8)³ )/ (4)³ x 3.8341*

*Skewness =((-10.648) + (-140.608) + (0.512) + (5.832) + (110.592)) / 64 x 3.8341*

*Skewness =-34.32 / 245.3824*

*Skewness = -0.1522*

*So, the skewness of these data is negative.*

*-----------------------------code-----------------------------------------*

*import pandas as pd*

*import matplotlib.pyplot as plt*

*# Create a DataFrame with positive skewness*

*data = {'Values': [1, 2, 2, 3, 3, 3, 4, 4, 4, 4, 5, 5, 5, 5, 6, 7, 8, 9, 10, 15]}*

*df = pd.DataFrame(data)*

*# Calculate skewness*

*skewness = df['Values'].skew()*

*print(f'Skewness: {skewness}')*

*# Plot histogram*

*df['Values'].plot(kind='hist', bins=15, color='orange', edgecolor='black')*

*# Show plot*

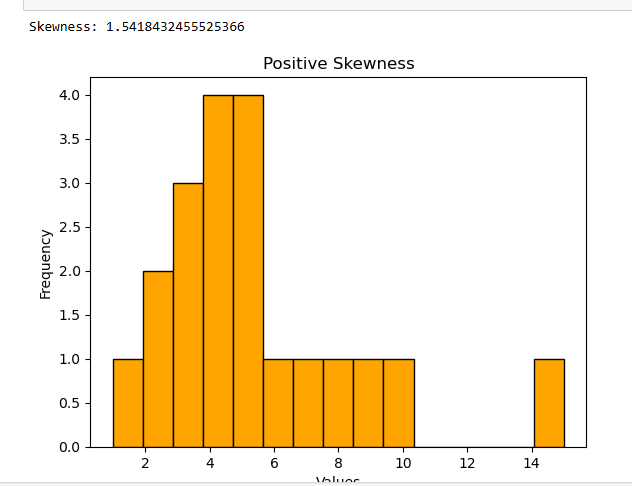
*plt.title('Positive Skewness')*

*plt.xlabel('Values')*

*plt.ylabel('Frequency')*

*plt.show()*

*Note->*

**

The plt.ylabel('Frequency') function in Python's Matplotlib library is used to label the y-axis of a plot. In this case, "Frequency" refers to the number of times each data value appears within the histogram bins.

Here's a quick breakdown:

* **plt.ylabel()**: Sets a label for the y-axis.
* **'Frequency'**: Indicates the count or frequency of the occurrences of the data values in each bin of the histogram.
* import numpy as np
* import numpy as np
* import pandas as pd
* import matplotlib.pyplot as plt
* import seaborn as sns

Incomes =  [2000, 2500, 3000, 3200, 3500, 3800, 4000, 4200, 4500, 10000]

df=np.array(Incomes)

df

np.mean(df)

np.median(df)

from scipy import stats

stats.mode(df)

np.std(df)

sd=pd.DataFrame(df)

sd

sd.columns=['Income']

sd

sd.skew()

plt.figure(figsize=(10, 6))

plt.hist(sd['Income'], bins=10, edgecolor='r', color='skyblue')

plt.title('Histogram of Monthly Income')

plt.xlabel('Income ($)')

plt.ylabel('Frequency')

plt.grid(axis='y', alpha=0.75)

plt.show()

plt.figure(figsize=(10, 6))

sns.kdeplot(sd['Income'], color='red')

plt.title('Histogram of Monthly Income')

plt.xlabel('Income ($)')

plt.ylabel('Frequency')

plt.grid(axis='y', alpha=0.75)

plt.show()

#positive skewness

#mean>median>mode

Scores = [90, 92, 93, 94, 95, 96, 97, 98, 98, 50]

cv=np.array(Scores)

cv

np.mean(cv)

np.median(cv)

stats.mode(cv)

np.std(cv)

vb=pd.DataFrame(cv)

vb.columns=['Scores']

vb

vb.skew()

plt.figure(figsize=(10, 6))

plt.hist(vb['Scores'], bins=10, edgecolor='r', color='skyblue')

plt.title('Histogram of Scores')

plt.xlabel('Scores')

plt.ylabel('Frequency')

plt.grid(axis='y', alpha=0.75)

plt.show()

plt.figure(figsize=(10, 6))

sns.kdeplot(vb['Scores'], color='red')

plt.title('Histogram of Scores')

plt.xlabel('Scores')

plt.ylabel('Frequency')

plt.grid(axis='y', alpha=0.75)

plt.show()

#negative skewness

#mean<median<mode

data=[2,3,3,4,4,4,5,5,5,5,6,6,6,6,6,7,7,7,7,7,7,8,8,8,8,8,9,9,9,9,10,10,10,11,11,12]

gf=np.array(data)

gf

np.mean(gf)

np.median(gf)

stats.mode(gf)

np.std(gf)

bf=pd.DataFrame(gf)

bf

bf.columns=['Data']

bf

bf.skew()

plt.figure(figsize=(10, 6))

plt.hist(bf['Data'], bins=36, edgecolor='r', color='skyblue')

plt.title('Histogram of Data')

plt.xlabel('Data')

plt.ylabel('Frequency')

plt.grid(axis='y', alpha=0.75)

plt.show()

plt.figure(figsize=(10, 6))

sns.kdeplot(bf['Data'], color='red')

plt.title('Histogram of Data')

plt.xlabel('Data')

plt.ylabel('Frequency')

plt.grid(axis='y', alpha=0.75)

plt.show()

# ---------------------------------------------------

# Sample and Population Deviation

The main difference between population and sample standard deviation is the amount of data used to calculate them:

* **Population standard deviation**

Uses data from the entire population. This is used when the research question requires it, or when it's possible to collect data from every member of the population.

* **Sample standard deviation**

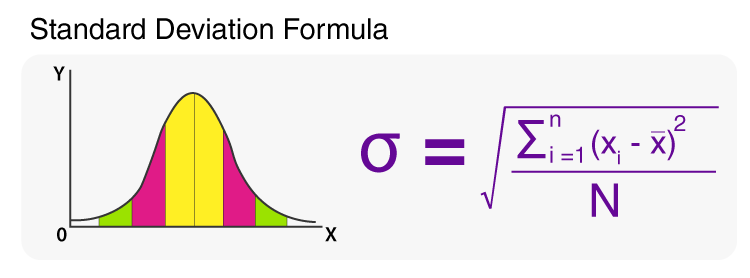
Uses data from a sample, which is a smaller group selected from the population. This is used when working with a sample

# Standard Deviation Formula

Standard deviation formula is used to find the values of a particular data that is dispersed. In simple words, the standard deviation is defined as the deviation of the values or data from an average mean. Lower standard deviation concludes that the values are very close to their average. Whereas higher values mean the values are far from the mean value. It should be noted that the [standard deviation](https://byjus.com/maths/standard-deviation/) value can never be negative.

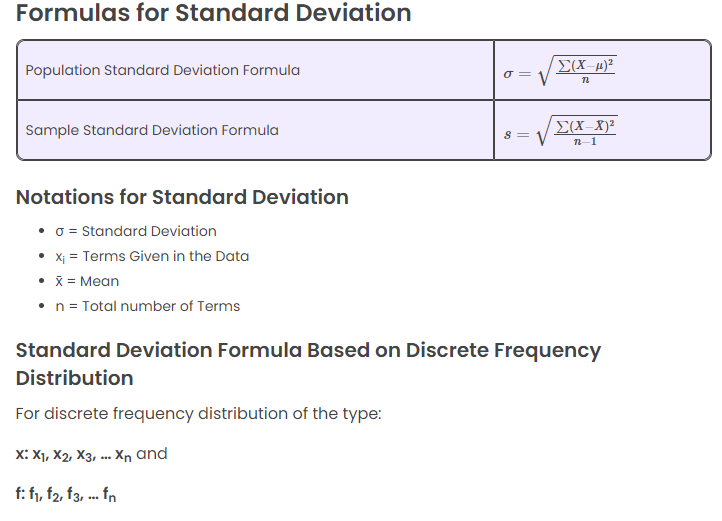
**Standard Deviation is of two types:**

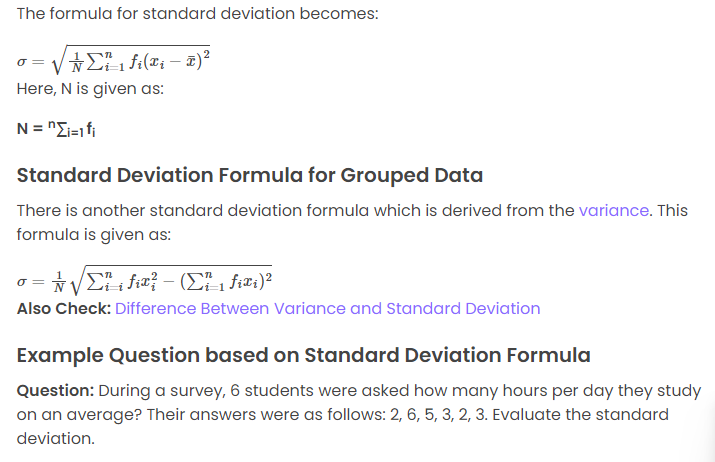
1. Population Standard Deviation
2. Sample Standard Deviation

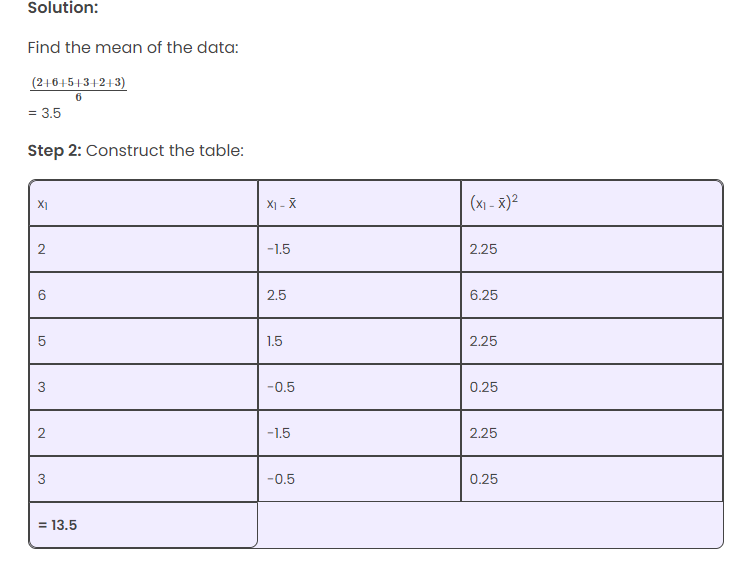


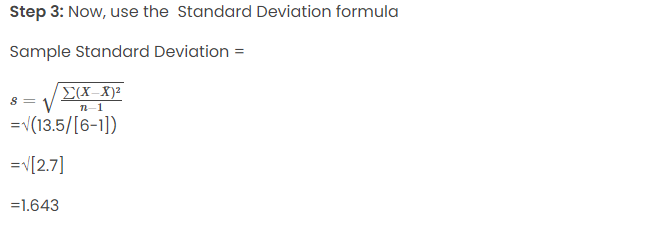
Formula to Calculate Standard Deviation

**Formulas for Standard Deviation**









### Standard Deviation Overview

**Standard Deviation (SD)** is a measure of how spread out the values in a dataset are. It quantifies the amount of variation or dispersion from the mean (average) of the dataset.

* **Low SD** means that most of the data points are close to the mean.
* **High SD** means that the data points are spread out over a wider range of values.

The formula for standard deviation is:

σ=1N∑i=1N(xi−μ)2\sigma = \sqrt{\frac{1}{N}\sum\_{i=1}^{N} (x\_i - \mu)^2}σ=N1​i=1∑N​(xi​−μ)2​

Where:

* σ\sigmaσ is the standard deviation.
* NNN is the number of data points.
* xix\_ixi​ represents each data point.
* μ\muμ is the mean of the data.

### First, Second, and Third Standard Deviations

The **first**, **second**, and **third standard deviations** refer to ranges around the mean in increments of the standard deviation, usually in the context of a normal distribution (bell curve). These ranges help quantify the probability of data points falling within a certain distance from the mean.

#### 1. **First Standard Deviation (±1σ)**

* The range from **mean - 1 standard deviation** to **mean + 1 standard deviation**.
* In a **normal distribution**, about **68%** of the data lies within this range.
* Example: If the mean is 50 and the standard deviation is 10, the first standard deviation range would be from 40 to 60 (mean ± 10).

#### 2. **Second Standard Deviation (±2σ)**

* The range from **mean - 2 standard deviations** to **mean + 2 standard deviations**.
* In a normal distribution, about **95%** of the data lies within this range.
* Example: If the mean is 50 and the standard deviation is 10, the second standard deviation range would be from 30 to 70 (mean ± 20).

#### 3. **Third Standard Deviation (±3σ)**

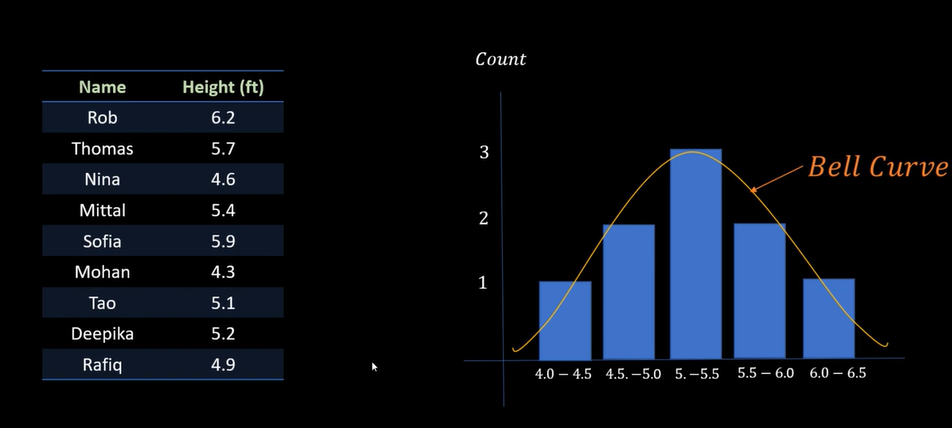
* The range from **mean - 3 standard deviations** to **mean + 3 standard deviations**.
* In a normal distribution, about **99.7%** of the data lies within this range.
* Example: If the mean is 50 and the standard deviation is 10, the third standard deviation range would be from 20 to 80 (mean ± 30).

### Usage of Standard Deviation

1. **Data Distribution**:
   * Standard deviation tells you how much variability exists in the data compared to the mean.
   * A large standard deviation indicates that the data points are far from the mean, while a small standard deviation shows that they are clustered around the mean.

**NORMAL DISTRIBUTION (HIST Graph)**

**What is Normal Distribution?**

****

**Above Image**

**If you data show in bell curve then is called normal distribution.**

**3 people are average height**

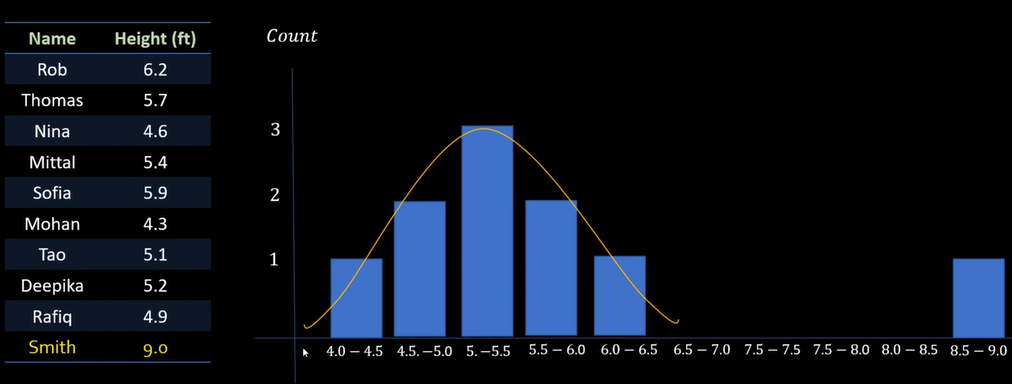
**2 people are other height**

**1 people are other height**

**Mostly majority is average height then you can call is normal distribution**

**How normal distribution is used on Data Analysis**

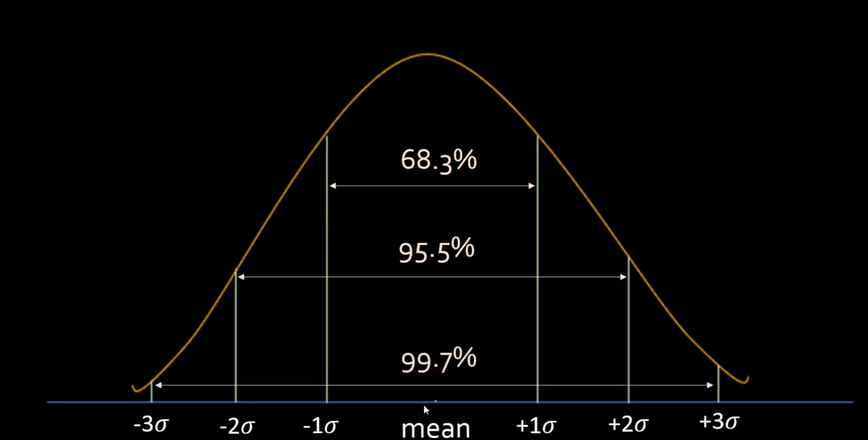
**Ans- Data Cleaning: Outlier Removal**

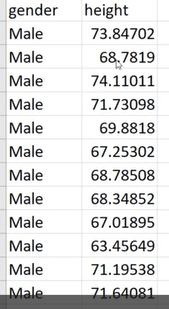
****

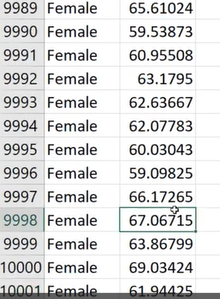
**Above Image show smith height is 9 which is not correct for dataset. Then you should be remove outlier.**

**If you have big dataset then first check outlier with help of histro graph plot. And clean the data**

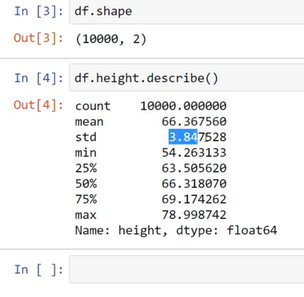
**Then prediction should be apply.**

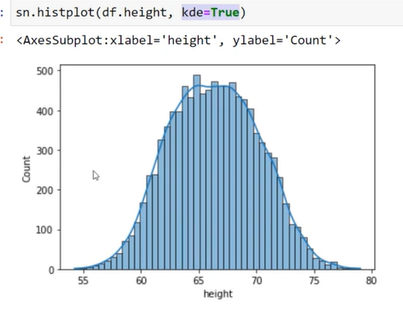
****

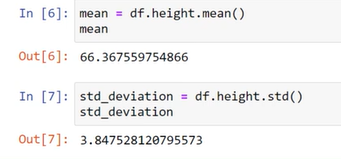
****

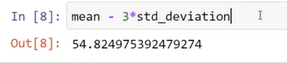
****

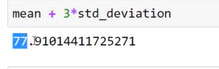
****

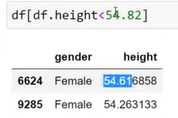
****

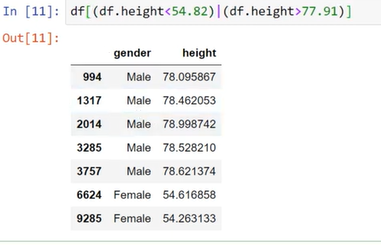
****

****

****

****

****

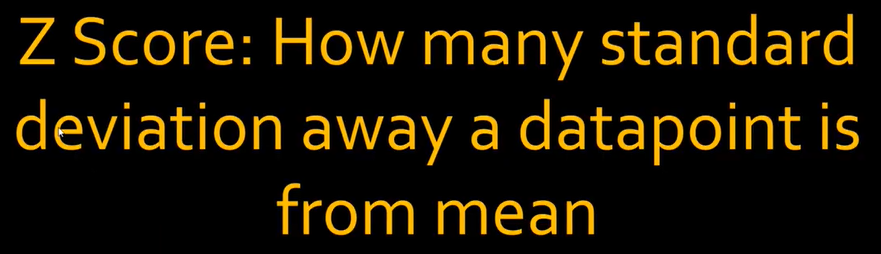
****

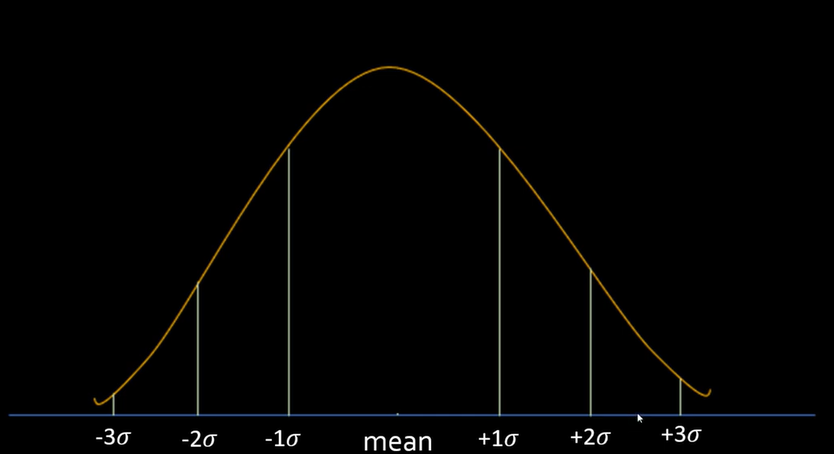
****

**Above show <77.91 and >54.82 between height me outerlier nahi hai.**

**Refer Video**

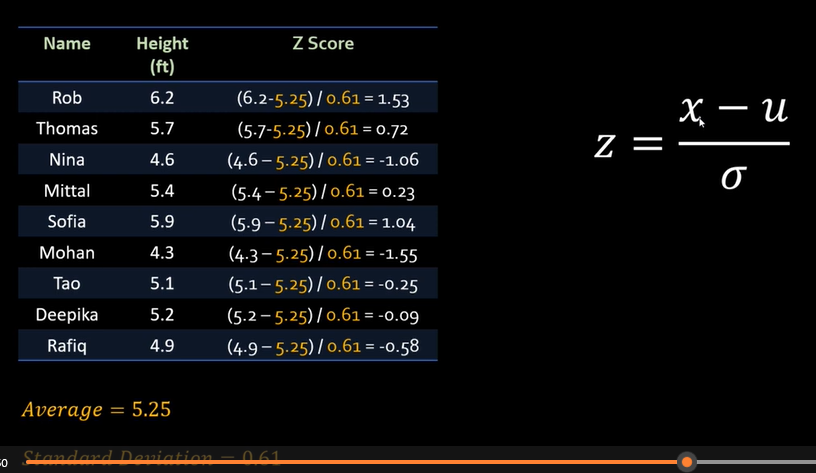
**2.What is Z-Score?**

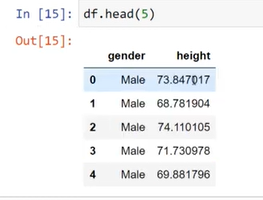
****

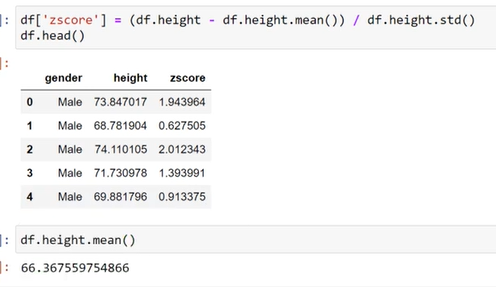
****

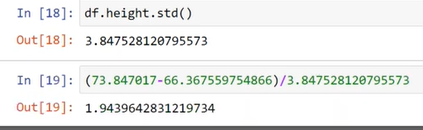
**-3 is kam and +3 is zyda sd is outlier hai**

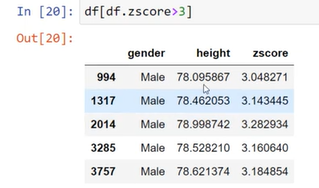
**Mean dataset se nikal dena hai.taki proper data analysis hoga.**

****

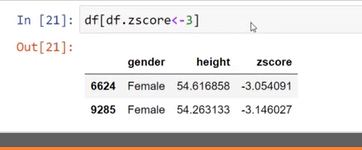
****

****

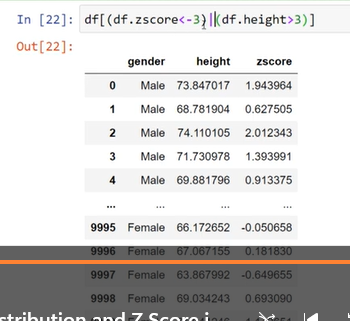
****

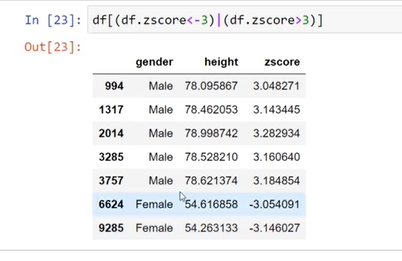
****

**Above is Outlier bez greater than 3**

****

**Above is outlier less than -3**

****

****

# How to Calculate Z-Scores in Python ( scipy.stats as stats)

In statistics, a z-score tells us how many standard deviations away a value is from the mean. We use the following formula to calculate a z-score:

z = (X – μ) / σ

where:

X is a single raw data value  
μ is the population mean  
σ is the population standard deviation

This tutorial explains how to calculate z-scores for raw data values in Python.

**How to Calculate Z-Scores in Python**

We can calculate z-scores in Python using scipy.stats.zscore, which uses the following syntax:

scipy.stats.zscore(a, axis=0, ddof=0, nan\_policy=’propagate’)

where:

a: an array like object containing data  
axis: the axis along which to calculate the z-scores. Default is 0.  
ddof: degrees of freedom correction in the calculation of the standard deviation. Default is 0.  
nan\_policy: how to handle when input contains nan. Default is propagate, which returns nan. ‘raise’ throws an error and ‘omit’ performs calculations ignoring nan values.  
The following examples illustrate how to use this function to calculate z-scores for one-dimensional numpy arrays, multi-dimensional numpy arrays, and Pandas DataFrames.

**Numpy One-Dimensional Arrays**

**Step 1: Import modules.**

import pandas as pd  
import numpy as np  
import scipy.stats as stats

**Step 2: Create an array of values.**

data = np.array([6, 7, 7, 12, 13, 13, 15, 16, 19, 22])

**Step 3: Calculate the z-scores for each value in the array.**

stats.zscore(data)

outpu: [-1.394, -1.195, -1.195, -0.199, 0, 0, 0.398, 0.598, 1.195, 1.793]

Each z-score tells us how many standard deviations away an individual value is from the mean. For example:

The first value of “6” in the array is 1.394 standard deviations below the mean.  
The fifth value of “13” in the array is 0 standard deviations away from the mean, i.e. it is equal to the mean.  
The last value of “22” in the array is 1.793 standard deviations above the mean.

Example

import pandas as pd

import numpy as np

import scipy.stats as stats

data = np.array([6, 7, 7, 12, 13, 13, 15, 16, 19, 22])

print(data)

print("\n")

print("mean:",np.mean(data))

print("std:",np.std(data))

# 6-mena(13)=-7

z=-7/np.std(data)

print("First Value of Z score :",z)

print("\n\n")

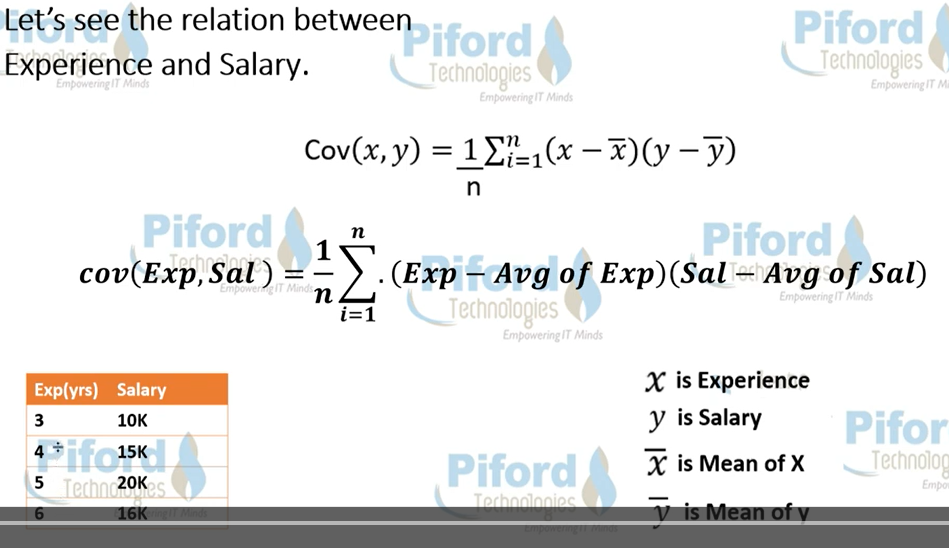
print(stats.zscore(data))

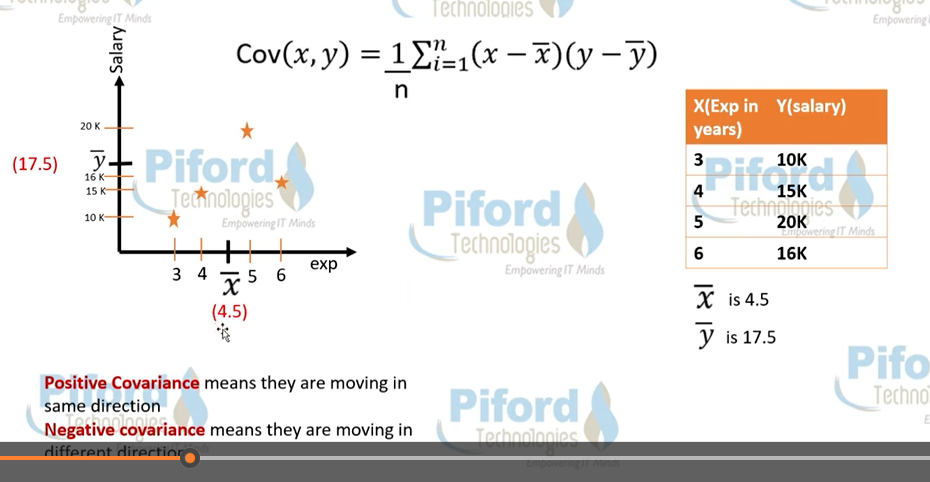
**What is Covariance?**

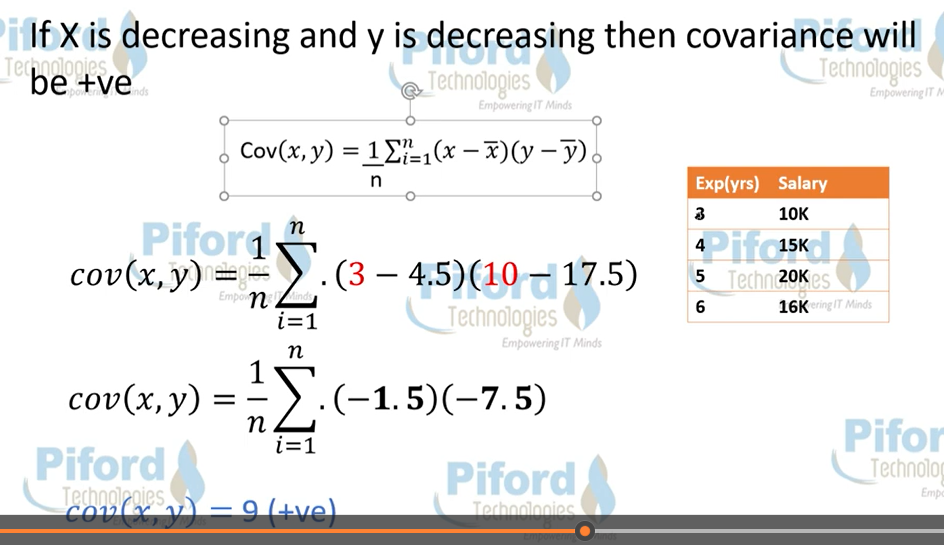
In statistics and probability theory, covariance deals with the joint variability of two random variables: x and y. Generally, it is treated as a statistical tool used to define the relationship between two variables. In this article, covariance meaning, formula, and its relation with correlation are given in detail.

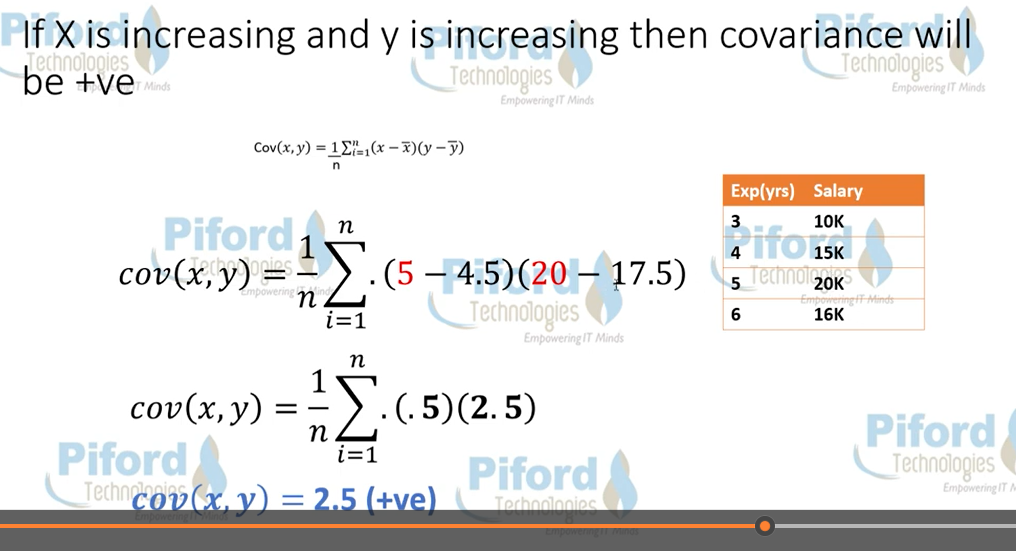
## Covariance Meaning

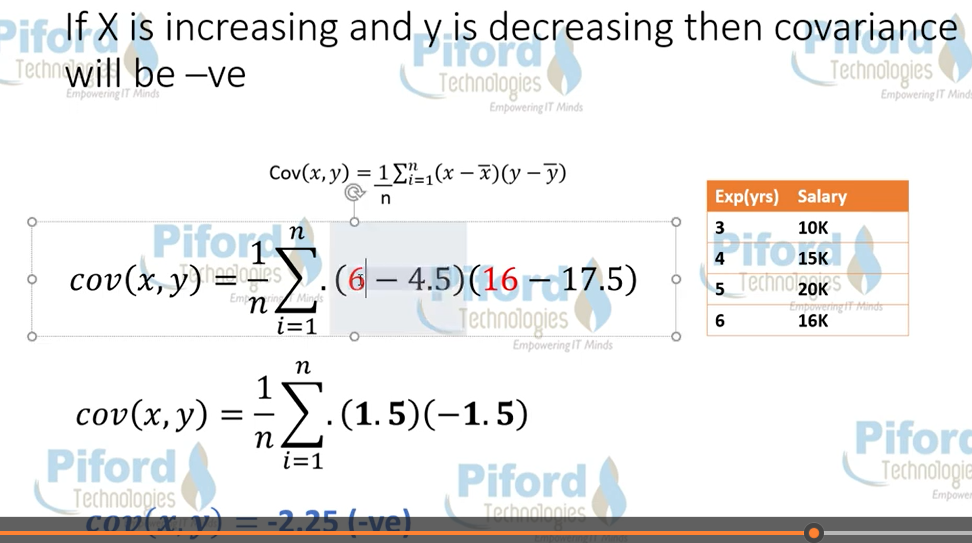
**Covariance** is a measure of the relationship between two random variables and to what extent, they change together. Or we can say, in other words, it defines the changes between the two variables, such that change in one variable is equal to change in another variable

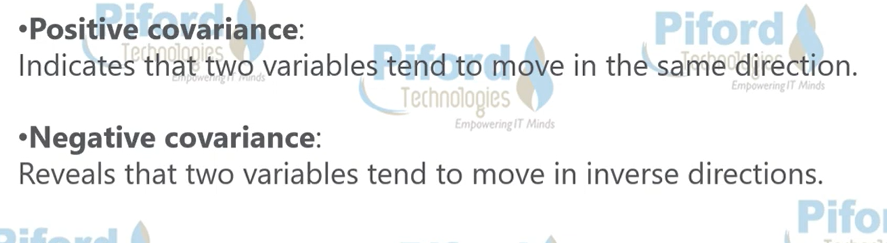












Sure, let's calculate the covariance step by step.

Given: Data set 𝑥*x*: [3, 4, 5, 6] Data set 𝑦*y*: [10, 15, 20, 16]

Step 1: Calculate the means of both datasets. Mean of 𝑥*x*:

𝑥ˉ=3+4+5+64=184=4.5*x*ˉ=43+4+5+6​=418​=4.5

Mean of 𝑦*y*:

𝑦ˉ=10+15+20+164=614=15.25*y*ˉ​=410+15+20+16​=461​=15.25

Step 2: Calculate the deviations of each data point from their respective means. For dataset 𝑥*x*:

𝑥𝑖−𝑥ˉ*xi*​−*x*ˉ

For dataset 𝑦*y*:

𝑦𝑖−𝑦ˉ*yi*​−*y*ˉ​

Step 3: Multiply the deviations for each pair of data points and sum them up.

Cov(𝑥,𝑦)=∑𝑖=1𝑛(𝑥𝑖−𝑥ˉ)(𝑦𝑖−𝑦ˉ)𝑛−1Cov(*x*,*y*)=*n*−1∑*i*=1*n*​(*xi*​−*x*ˉ)(*yi*​−*y*ˉ​)​

Let's calculate:

(𝑥1−𝑥ˉ)×(𝑦1−𝑦ˉ)=(3−4.5)×(10−15.25)=(−1.5)×(−5.25)=7.875(*x*1​−*x*ˉ)×(*y*1​−*y*ˉ​)=(3−4.5)×(10−15.25)=(−1.5)×(−5.25)=7.875

(𝑥2−𝑥ˉ)×(𝑦2−𝑦ˉ)=(4−4.5)×(15−15.25)=(−0.5)×(−0.25)=0.125(*x*2​−*x*ˉ)×(*y*2​−*y*ˉ​)=(4−4.5)×(15−15.25)=(−0.5)×(−0.25)=0.125

(𝑥3−𝑥ˉ)×(𝑦3−𝑦ˉ)=(5−4.5)×(20−15.25)=(0.5)×(4.75)=2.375(*x*3​−*x*ˉ)×(*y*3​−*y*ˉ​)=(5−4.5)×(20−15.25)=(0.5)×(4.75)=2.375

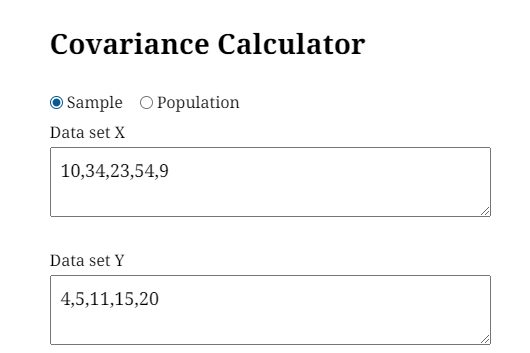
(𝑥4−𝑥ˉ)×(𝑦4−𝑦ˉ)=(6−4.5)×(16−15.25)=(1.5)×(0.75)=1.125(*x*4​−*x*ˉ)×(*y*4​−*y*ˉ​)=(6−4.5)×(16−15.25)=(1.5)×(0.75)=1.125

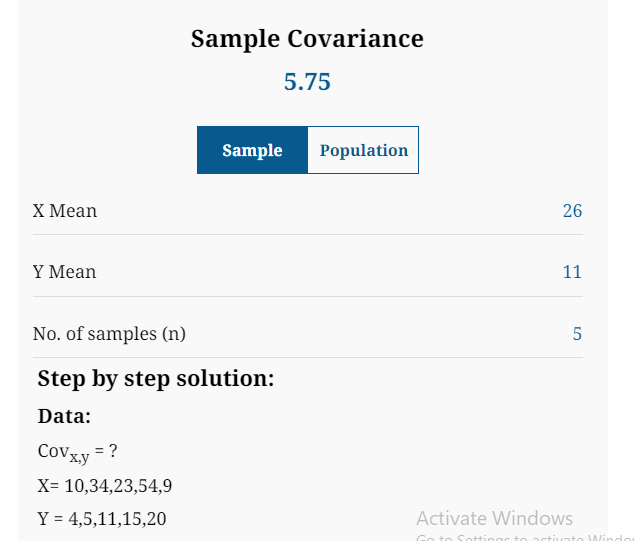
Step 4: Sum up these values and divide by 𝑛−1*n*−1 (where 𝑛*n* is the number of data points).

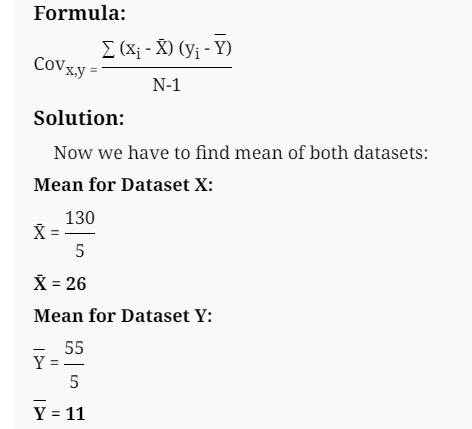
Cov(𝑥,𝑦)=7.875+0.125+2.375+1.1254−1=11.53≈3.8333Cov(*x*,*y*)=4−17.875+0.125+2.375+1.125​=311.5​≈3.8333

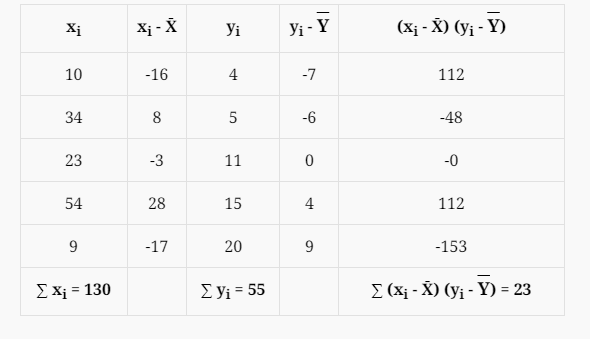
So, the covariance of datasets 𝑥*x* and 𝑦*y* is approximately 3.83333.8333.

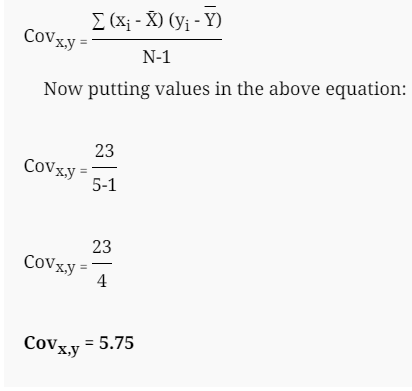
Top of Form











Use Below Link to calculate covariance.

[**https://www.standarddeviationcalculator.io/covariance-calculator**](https://www.standarddeviationcalculator.io/covariance-calculator)

**-------------------code---------------------------------**

**import numpy as np**

**import matplotlib.pyplot as plt**

**# Sample data**

**temperature = np.array([20, 25, 30, 35, 40, 45, 50])**

**ice\_cream\_sales = np.array([100, 120, 150, 180, 200, 220, 250])**

**covariance = np.cov(temperature, ice\_cream\_sales)[0, 1]**

**# Print the covariance**

**print("Covariance:", covariance) #Sample Covariance 583.333**

**# Create a scatter plot**

**plt.scatter(temperature, ice\_cream\_sales)**

**plt.xlabel("Temperature")**

**plt.ylabel("Ice Cream Sales")**

**plt.title("Temperature vs. Ice Cream Sales")**

**plt.show()**

The code covariance = np.cov(temperature, ice\_cream\_sales)[0, 1] calculates the covariance between the temperature and ice\_cream\_sales arrays. Here's a breakdown of what each part does:

1. **np.cov(temperature, ice\_cream\_sales):**
   * This part calculates the covariance matrix between the two arrays, temperature and ice\_cream\_sales. A covariance matrix is a square matrix that shows the covariance between each pair of variables.
   * In this case, since we have two variables, the covariance matrix will be a 2x2 matrix.
2. **[0, 1]:**
   * This part accesses a specific element from the covariance matrix. The indices 0 and 1 refer to the row and column, respectively.

**# Python code to demonstrate the**

**# use of numpy.cov**

**import numpy as np**

**x = [3,4,5,6]**

**mx=(3+4+5+6)/4**

**print("mean X:",mx)**

**y = [10,15,20,16]**

**my=(10+15+20+16)/4**

**print("mean y:",my)**

**print(x)**

**print(y)**

**s=0**

**for i in range(0,4):**

**s=s+((x[i]-4.5)\*(y[i]-15.25))**

**print("s ",s)**

**result=s/(len(x)-1)**

**print("Result:",result)**

**Covariance and variance are both statistical terms that are used to measure different things**:

* **Covariance**

Measures the relationship between two variables, or how much they change together. Covariance is used to determine if two variables are related, and whether they move in the same or opposite directions. Covariance is measured in two dimensions and has a unit of XY.

* **Variance**

Measures the spread of a set of numbers from their mean value. Variance is used to determine how far each number in a set is from the mean and from every other number in the set. Variance is measured in one dimension.

# How to Calculate Covariance in 6 Steps (With Examples)

## What is covariance?

Covariance is a measurement used in statistics to determine if two variables are changing in the same direction. It is a measurement of the difference between the two variables, and the two variables used to determine covariance are unrelated.

You can measure covariance in terms of units related to the two variables in the data sets. For example, in finance, two data sets could be the cost of one company's stock, while the other could be an unrelated company's stock. Since they represent both values in terms of dollars, the units for the measurement would be dollars.

Covariance compares the two variables in terms of positive and negative. If the value for covariance is negative, then, the two variables move in opposite directions. If the value for covariance is positive, then the two variables move in the same direction.

Notably, this means that two variables could decrease in the same direction, and the covariance would still return as a positive value. For example, if two companies both have stocks that are becoming cheaper over time, then their covariance would be positive.

**Related:** [**How To Calculate Sample Covariance**](https://www.indeed.com/career-advice/career-development/calculate-sample-covariance)

## Covariance versus variance

Variance is a measurement of the distance between a variable and the average value of a set of data. Unlike covariance, one data point or trend is the average, while the other is a point or trend of interest that you decide to measure.

Using the above example, if the first company has stock that is growing over time, but the general trend for all stock is falling, then, the variance between the average and the company's stock may increase. If the second company's stock is also increasing at a similar rate to the first, then the covariance would be positive.

**Related:** [**What Is Variance?**](https://www.indeed.com/career-advice/career-development/what-is-variance)

## How to calculate covariance

To calculate covariance, you can use the formula:

**Cov(X, Y) = Σ(Xi-µ)(Yj-v) / n**

Where the parts of the equation are:

* Cov(X, Y) represents the covariance of variables X and Y.
* Σ represents the sum of other parts of the formula.
* (Xi) represents all values of the X-variable.
* µ represents the average value of the X-variable.
* Yj represents all values of the Y-variable.
* v represents the average value of the Y-variable.
* Σ represents the sum of the values for both (Xi-µ) and (Yj-v).
* n represents the total number of data points across both variables.

You can use the following steps and the covariance formula to find the covariance of your data:

### 1. Get the data

The first step in finding the covariance of two variables is to gather the data for both sets. For example, the table below shows the values of two new company stocks between 2015 and 2020:

| **Year** | **Company X Stock Value ($)** | **Company Y Stock Value ($)** |
| --- | --- | --- |
| 2015 | 1,245 | 100 |
| 2016 | 1,415 | 123 |
| 2017 | 1,312 | 129 |
| 2018 | 1,427 | 143 |
| 2019 | 1,510 | 150 |
| 2020 | 1,590 | 197 |

### 2. Calculate the average value for each variable

To find the average value for each stock, add all the X-values together and divide by the total number of X-values. Then, do the same for the Y-values:

1. µ = 1,245 + 1,415 + 1,312 + 1,427 + 1,510 + 1,590 / 6
2. µ = 1,416.5
3. v = 100 + 123 + 129 + 143 + 150 + 197 / 6
4. v = 140.3

### 3. Find the difference between each value and the mean for both variables

Subtract the mean value for each set of variables from each variable within that set. For example:

| **Year** | **Company X (Xi-µ)** | **Company Y (Yj-v)** |
| --- | --- | --- |
| 2015 | 1,245 - 1,416.5 = -171.5 | 100 - 140.3 = -40.3 |
| 2016 | 1,415 - 1,416.5 = -1.5 | 123 - 140.3 = -17.3 |
| 2017 | 1,312 - 1,416.5 = -104.5 | 129 - 140.2 = -11.2 |
| 2018 | 1,427 - 1,416.5 = 10.5 | 143 - 140.3 = 2.7 |
| 2019 | 1,510 - 1,416.5 = 93.5 | 150 - 140.3 = 9.7 |
| 2020 | 1,590 - 1,416.5 = 173.5 | 197 - 140.3 = 56.7 |

### 4. Multiply the values for the two variables

Once you have found the values for both variables in the previous step, you can multiply them together. For example:

| **Year** | **Company X (Xi-µ)** | **Company Y (Yj-v)** | **(Xi-µ)(Yj-v)** |
| --- | --- | --- | --- |
| 2015 | 1,245 - 1,416.5 = -171.5 | 100 - 140.3 = -40.3 | (-171.5)(-40.3) = 6,911.45 |
| 2016 | 1,415 - 1,416.5 = -1.5 | 123 - 140.3 = -17.3 | (-1.5)(-17.3) = 25.95 |
| 2017 | 1,312 - 1,416.5 = -104.5 | 129 - 140.2 = -11.2 | (-104.5)(-11.3) = 1,180.85 |
| 2018 | 1,427 - 1,416.5 = 10.5 | 143 - 140.3 = 2.7 | (10.5)(2.7) = 28.35 |
| 2019 | 1,510 - 1,416.5 = 93.5 | 150 - 140.3 = 9.7 | (93.5)(9.7) = 906.95 |
| 2020 | 1,590 - 1,416.5 = 173.5 | 197 - 140.3 = 56.7 | (173.5)(56.7) = 9,837.45 |

### 5. Add the values together

After you have calculated the product of the two variables together, you can add the values to get the second to last part of the equation. For example, you can add the product values from the companies above to get the summation of all values:

*6,911.45 + 25.95 + 1,180.85 + 28.35 + 906.95 + 9,837.45 = 18,891*

### 6. Use the values from previous steps to find the covariance of the data

Once you have calculated the parts of the equation, you can put your values into it. For example, you can put the stocks of the company from above into the equation as shown below:

*Cov(X, Y) = 18,891 / 6*

Where the values are:

* *18,891 = Σ(Xi-µ)(Yj-v)*
* *6 = n*

As calculated above, the covariance of company X's stock and company Y's stock is 3,148.5. The positive nature of the covariance value shows that the two companies' stocks move in the same direction.

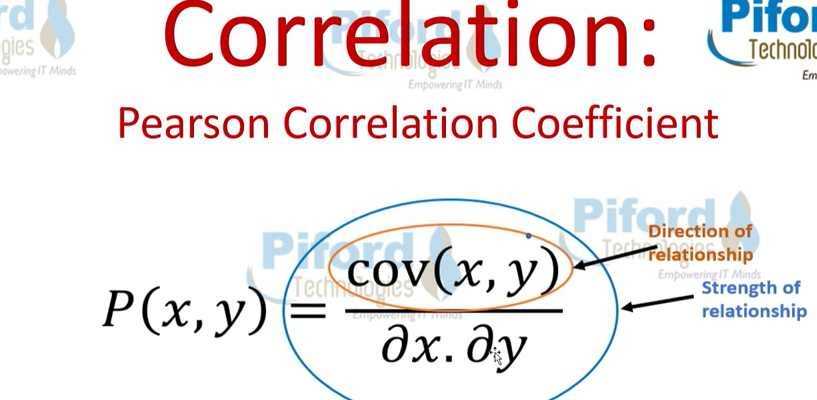
**------------------------------------------------------------------**

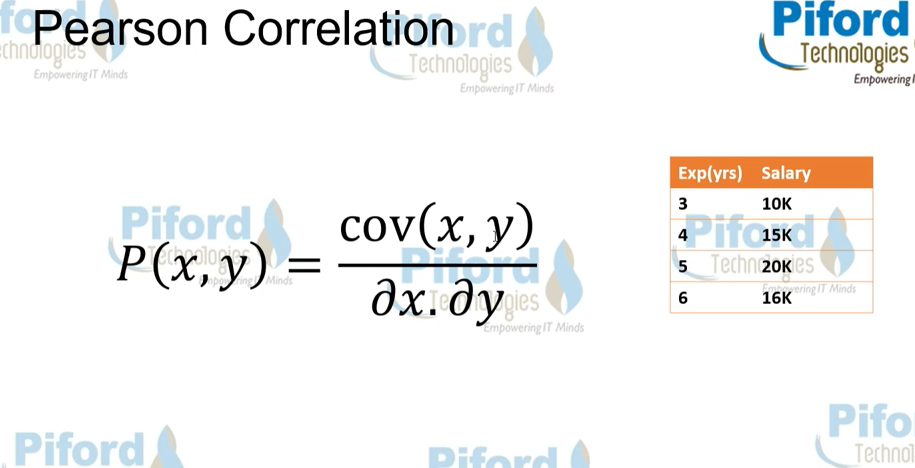
**Correlation**

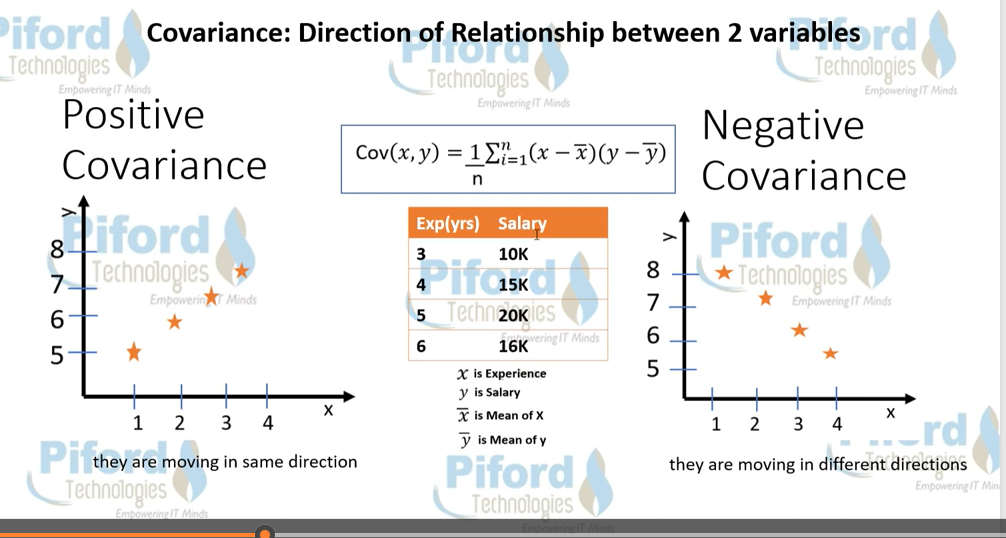
**Pearson Correlation Coefficient**

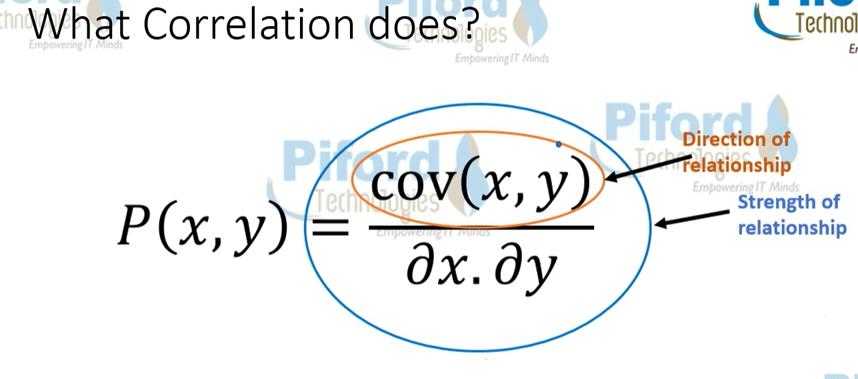
**Correlation** refers to a process for establishing the relationships between two variables.

 The correlation coefficient is usually represented using the symbol r, and it ranges from -1 to +1.

****

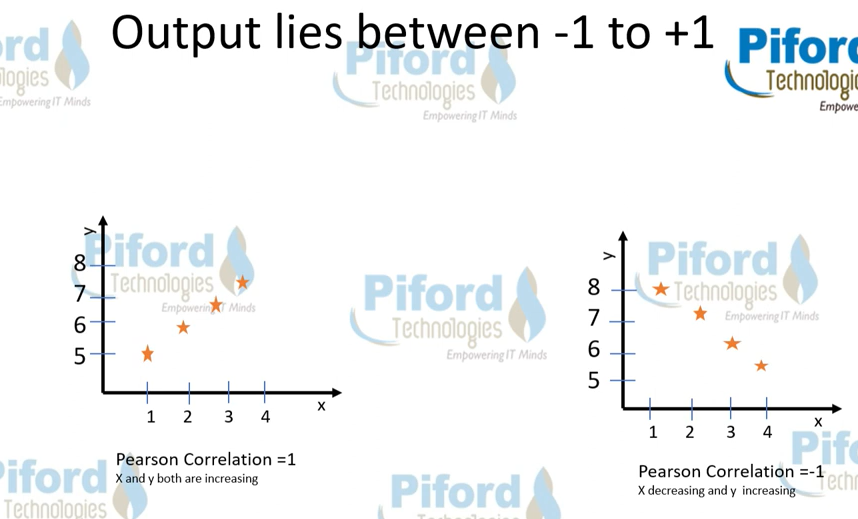
****

****

****

**Covariance only check direction of relationship between two variable.**

**While Correlation check the how much positive or negative direction move the data is called strength of relationship**

****

**1 means higher similarity or equal between two variable.**

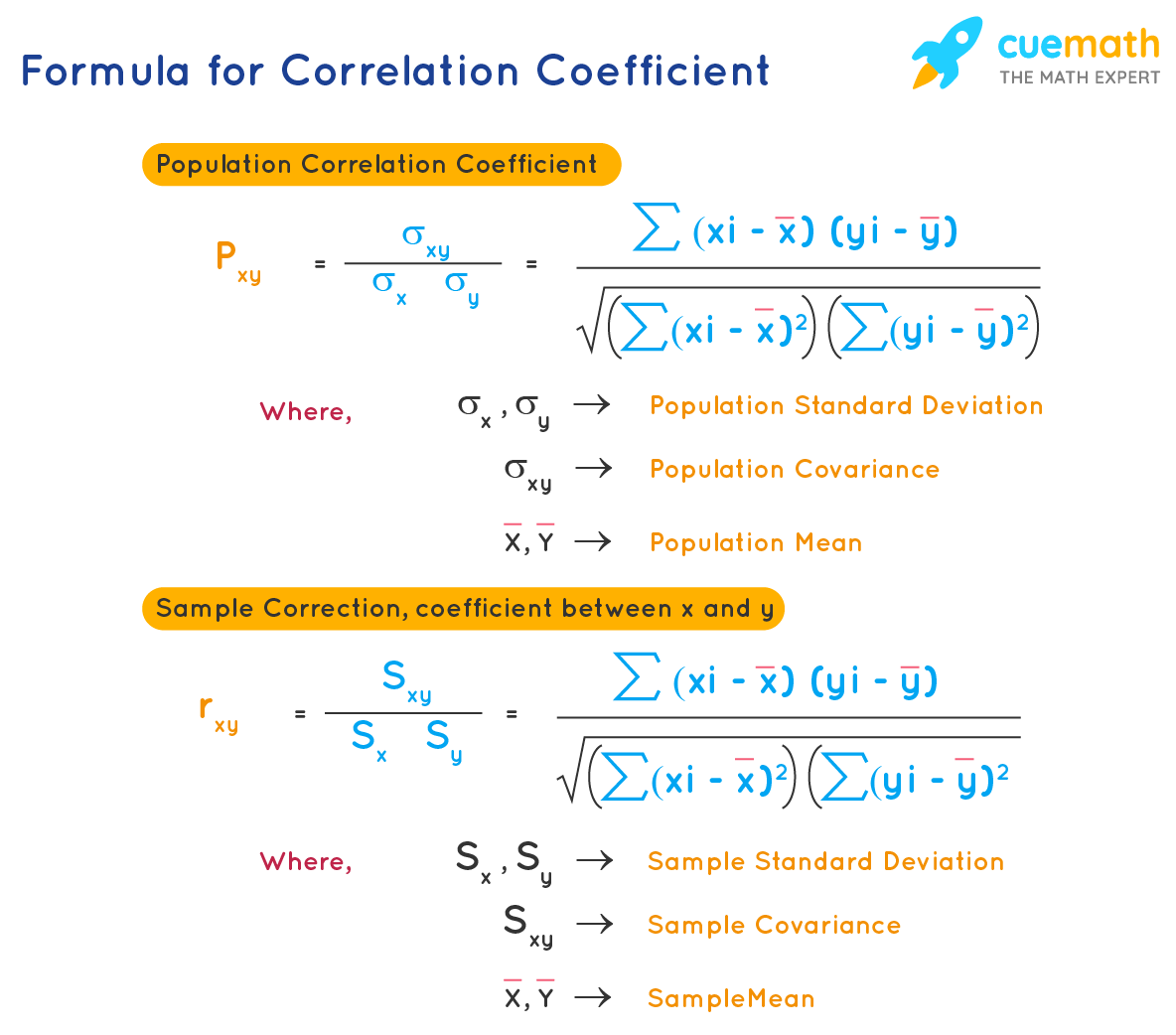
**-1 mean low similarity between two variable.**

**Why Use Correlation?**

**Use to feature selection**

**Suppose you have two or three column have same feature then you can drop two column in ML. You can use single column and save the time of model also or model work as fast.**

# Correlation Coefficient Formula



**Exercise and solution below Link**

**https://www.cuemath.com/correlation-coefficient-formula/**

**--------------------------code--------------------------**

**import numpy as np**

**import matplotlib.pyplot as plt**

**# Sample data**

**#x = np.array([1, 2, 3, 4, 5])**

**#y = np.array([2, 4, 5, 4, 5])**

**# Sample data**

**x = np.array([1, 2, 3, 4, 5])**

**y = np.array([2, 4, 6, 7, 8])**

**# Calculate correlation coefficient**

**correlation = np.corrcoef(x, y)[0, 1]**

**# Print the correlation coefficient**

**print("Correlation coefficient:", correlation)**

**# Create a scatter plot**

**plt.scatter(x, y)**

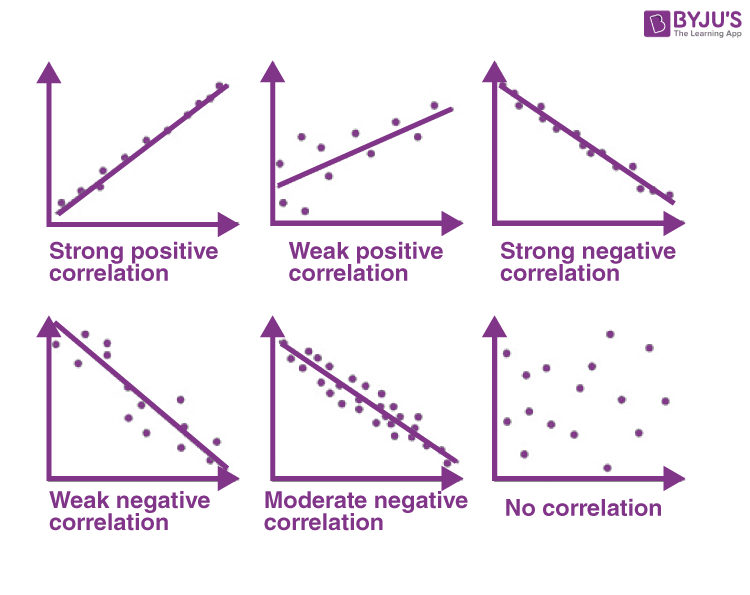
**plt.xlabel("Variable X")**

**plt.ylabel("Variable Y")**

**plt.title("Correlation Example")**

**plt.show()**

## Types of Correlation



**Regression** is a statistical method used to analyze the relationship between a dependent variable and one or more independent variables. It's a powerful tool for modeling, prediction, and understanding how changes in one variable affect another.

**How does regression work?**

* **Dependent Variable:** This is the variable you're trying to predict or understand.
* **Independent Variables:** These are the variables that are believed to influence the dependent variable.
* **Model:** Regression creates a mathematical equation that represents the relationship between these variables. This equation is used to make predictions or understand the impact of changes in the independent variables on the dependent variable.

**Common Types of Regression**

1. **Linear Regression:**
   * Assumes a linear relationship between the dependent and independent variables.
   * The model is represented by a straight line equation.
   * Used for simple relationships or when the data points roughly follow a straight line.
2. **Multiple Linear Regression:**
   * Extends linear regression to include multiple independent variables.
   * Used when the dependent variable is influenced by more than one factor.
3. **Polynomial Regression:**
   * Allows for more complex relationships by fitting a polynomial curve to the data.
   * Used when the relationship is not linear but can be represented by a polynomial function.
4. **Logistic Regression:**
   * Used when the dependent variable is categorical (e.g., yes/no, true/false).
   * Predicts the probability of belonging to a particular category based on the independent variables.

## What is Hypothesis Testing?

Hypothesis testing is a statistical method that is used to make a statistical decision using experimental data. Hypothesis testing is basically an assumption that we make about a population parameter. It evaluates two mutually exclusive statements about a population to determine which statement is best supported by the sample data.

***Example:****You say an average height in the class is 30 or a boy is taller than a girl. All of these is an assumption that we are assuming, and we need some statistical way to prove these. We need some mathematical conclusion whatever we are assuming is true.*

### ****Defining Hypotheses****

* **Null hypothesis (H0):**In statistics, the null hypothesis is a general statement or default position that there is no relationship between two measured cases or no relationship among groups. In other words, it is a basic assumption or made based on the problem knowledge.  
  **Example**: A company’s mean production is 50 units/per da H0:= 50.
* **Alternative hypothesis (H1):**The alternative hypothesis is the hypothesis used in hypothesis testing that is contrary to the null hypothesis.   
  Example: A company’s production is not equal to 50 units/per day i.e. H1:50.

### Key Terms of Hypothesis Testing

* **Level of significance**: It refers to the degree of significance in which we accept or reject the null hypothesis. 100% accuracy is not possible for accepting a hypothesis, so we, therefore, select a level of significance that is usually 5%. This is normally denoted with and generally, it is 0.05 or 5%, which means your output should be 95% confident to give a similar kind of result in each sample.
* **P-value:**The [P value](https://www.geeksforgeeks.org/p-value/), or calculated probability, is the probability of finding the observed/extreme results when the null hypothesis(H0) of a study-given problem is true. If your P-value is less than the chosen significance level then you reject the null hypothesis i.e. accept that your sample claims to support the alternative hypothesis.
* **Test Statistic:**The test statistic is a numerical value calculated from sample data during a hypothesis test, used to determine whether to reject the null hypothesis. It is compared to a critical value or p-value to make decisions about the statistical significance of the observed results.
* **Critical value**: The critical value in statistics is a threshold or cutoff point used to determine whether to reject the null hypothesis in a hypothesis test.
* **Degrees of freedom:** Degrees of freedom are associated with the variability or freedom one has in estimating a parameter. The degrees of freedom are related to the sample size and determine the shape.

## ****Why do we use Hypothesis Testing?****

Hypothesis testing is an important procedure in statistics. Hypothesis testing evaluates two mutually exclusive population statements to determine which statement is most supported by sample data. When we say that the findings are statistically significant, thanks to hypothesis testing.

## One-Tailed and Two-Tailed Test

One tailed test focuses on one direction, either greater than or less than a specified value. We use a one-tailed test when there is a clear directional expectation based on prior knowledge or theory. The critical region is located on only one side of the distribution curve. If the sample falls into this critical region, the null hypothesis is rejected in favor of the alternative hypothesis.

### One-Tailed Test

There are two types of one-tailed test:

* **Left-Tailed (Left-Sided) Test:** The alternative hypothesis asserts that the true parameter value is less than the null hypothesis. Example: H0​:and H1:
* **Right-Tailed (Right-Sided) Test**: The alternative hypothesis asserts that the true parameter value is greater than the null hypothesis. Example: H0 :and H1:

### Two-Tailed Test

A two-tailed test considers both directions, greater than and less than a specified value.We use a two-tailed test when there is no specific directional expectation, and want to detect any significant difference.

Example: H0:  50 and H1:

## What are Type 1 and Type 2 errors in Hypothesis Testing?

In hypothesis testing,[Type I and Type II errors](https://www.geeksforgeeks.org/alpha-and-beta-test/)are two possible errors that researchers can make when drawing conclusions about a population based on a sample of data. These errors are associated with the decisions made regarding the null hypothesis and the alternative hypothesis.

* **Type I error:** When we reject the null hypothesis, although that hypothesis was true. Type I error is denoted by alpha().
* **Type II errors:** When we accept the null hypothesis, but it is false. Type II errors are denoted by beta().

|  | **Null Hypothesis is True** | **Null Hypothesis is False** |
| --- | --- | --- |
| **Null Hypothesis is True (Accept)** | Correct Decision | Type II Error (False Negative) |
| **Alternative Hypothesis is True (Reject)** | Type I Error (False Positive) | Correct Decision |

## How does Hypothesis Testing work?

### ****Step 1: Define Null and Alternative Hypothesis****

State the null hypothesis (), representing no effect, and the alternative hypothesis (​), suggesting an effect or difference.

We first identify the problem about which we want to make an assumption keeping in mind that our assumption should be contradictory to one another, assuming [Normally distributed data.](https://www.geeksforgeeks.org/normal-distribution/)

### ****Step 2 – Choose significance level****

Select a significance level (), typically 0.05, to determine the threshold for rejecting the null hypothesis. It provides validity to our hypothesis test, ensuring that we have sufficient data to back up our claims. Usually, we determine our significance level beforehand of the test. The [p-value](https://www.geeksforgeeks.org/p-value/) is the criterion used to calculate our significance value.

### ****Step 3**** – ****Collect and Analyze data.****

Gather relevant data through observation or experimentation. Analyze the data using appropriate statistical methods to obtain a test statistic.

### ****Step 4-Calculate Test Statistic****

The data for the tests are evaluated in this step we look for various scores based on the characteristics of data. The choice of the test statistic depends on the type of hypothesis test being conducted.

There are various hypothesis tests, each appropriate for various goal to calculate our test. This could be a [Z-test](https://www.geeksforgeeks.org/z-test/), [Chi-square](https://www.geeksforgeeks.org/ml-chi-square-test-for-feature-selection/), [T-test](https://www.geeksforgeeks.org/t-test/), and so on.

1. **Z-test**: If population means and standard deviations are known. Z-statistic is commonly used.
2. **t-test**: If population standard deviations are unknown. and sample size is small than t-test statistic is more appropriate.
3. **Chi-square test**: Chi-square test is used for categorical data or for testing independence in contingency tables
4. **F-test**: F-test is often used in analysis of variance (ANOVA) to compare variances or test the equality of means across multiple groups.

We have a smaller dataset, So, T-test is more appropriate to test our hypothesis.

*T-statistic is a measure of the difference between the means of two groups relative to the variability within each group. It is calculated as the difference between the sample means divided by the standard error of the difference. It is also known as the t-value or t-score.*

### ****Step 5 – Comparing Test Statistic:****

In this stage, we decide where we should accept the null hypothesis or reject the null hypothesis. There are two ways to decide where we should accept or reject the null hypothesis.

#### **Method A: Using Crtical values**

Comparing the test statistic and tabulated critical value we have,

* If Test Statistic>Critical Value: Reject the null hypothesis.
* If Test Statistic≤Critical Value: Fail to reject the null hypothesis.

**Note:**Critical values are predetermined threshold values that are used to make a decision in hypothesis testing.To determine [critical values](https://www.geeksforgeeks.org/how-to-find-the-z-critical-value-in-python/) for hypothesis testing, we typically refer to a statistical distribution table , such as the normal distribution or t-distribution tables based on.

#### Method B: Using P-values

We can also come to an conclusion using the p-value,

* If the p-value is less than or equal to the significance level i.e. (), you reject the null hypothesis. This indicates that the observed results are unlikely to have occurred by chance alone, providing evidence in favor of the alternative hypothesis.
* If the p-value is greater than the significance level i.e. (), you fail to reject the null hypothesis. This suggests that the observed results are consistent with what would be expected under the null hypothesis.

**Note**: The p-value is the probability of obtaining a test statistic as extreme as, or more extreme than, the one observed in the sample, assuming the null hypothesis is true. To determine [p-value](https://www.geeksforgeeks.org/p-value/) for hypothesis testing, we typically refer to a statistical distribution table , such as the normal distribution or t-distribution tables based on.

### ****Step 7- Interpret the Results****

At last, we can conclude our experiment using method A or B.

## Calculating test statistic

To validate our hypothesis about a population parameter we use[statistical functions](https://www.geeksforgeeks.org/introduction-of-statistical-data-distributions/). We use the z-score, p-value, and level of significance(alpha) to make evidence for our hypothesis for [normally distributed data](https://www.geeksforgeeks.org/normal-distribution/).

### 1. Z-statistics:

When population means and standard deviations are known.

Here Formula but not show check from internet

where,

* is the sample mean,
* μ represents the population mean,
* σ is the standard deviation
* and n is the size of the sample.

### 2. T-Statistics

T test is used when n<30,

t-statistic calculation is given by:

Here Formula but not show check from internet

where,

* t = t-score,
* x̄ = sample mean
* μ = population mean,
* s = standard deviation of the sample,
* n = sample size

### 3. Chi-Square Test

Chi-Square Test for Independence categorical Data (Non-normally distributed) using:

Here Formula but not show check from internet

where,

* is the observed frequency in cell
* i,j are the rows and columns index respectively.
* is the expected frequency in cell, calculated as :

## Real life Hypothesis Testing example

Let’s examine hypothesis testing using two real life situations,

### ****Case A:**** D****oes a New Drug Affect Blood Pressure?****

Imagine a pharmaceutical company has developed a new drug that they believe can effectively lower blood pressure in patients with hypertension. Before bringing the drug to market, they need to conduct a study to assess its impact on blood pressure.

**Data:**

* Before Treatment: 120, 122, 118, 130, 125, 128, 115, 121, 123, 119
* After Treatment: 115, 120, 112, 128, 122, 125, 110, 117, 119, 114

#### **Step 1**: Define the Hypothesis

* **Null Hypothesis**: (H0)The new drug has no effect on blood pressure.
* **Alternate Hypothesis**: (H1)The new drug has an effect on blood pressure.

#### **Step 2:** Define the Significance level

Let’s consider the Significance level at 0.05, indicating rejection of the null hypothesis.

If the evidence suggests less than a 5% chance of observing the results due to random variation.

#### **Step 3**: Compute the test statistic

Using [paired T-test](https://www.geeksforgeeks.org/paired-t-test-a-detailed-overview/) analyze the data to obtain a test statistic and a p-value.

The test statistic (e.g., T-statistic) is calculated based on the differences between blood pressure measurements before and after treatment.

**t = m/(s/√n)**

Where:

* **m** = mean of the difference i.e Xafter, Xbefore
* **s** = standard deviation of the difference (d) i.e di​=Xafter,i​−Xbefore,
* **n** = sample size,

then, m= -3.9, s= 1.8 and n= 10

we, calculate the , T-statistic = -9 based on the formula for paired t test

#### Step 4: Find the p-value

The calculated t-statistic is -9 and degrees of freedom df = 9, you can find the p-value using statistical software or a t-distribution table.

thus, p-value = 8.538051223166285e-06

**Step 5: Result**

* If the p-value is less than or equal to 0.05, the researchers reject the null hypothesis.
* If the p-value is greater than 0.05, they fail to reject the null hypothesis.

**Conclusion:** Since the p-value (8.538051223166285e-06) is less than the significance level (0.05), the researchers reject the null hypothesis. There is statistically significant evidence that the average blood pressure before and after treatment with the new drug is different.

### Python Implementation of Hypothesis Testing

Let’s create hypothesis testing with python, where we are testing whether a new drug affects blood pressure. For this example, we will use a paired T-test. We’ll use the **scipy.stats** library for the T-test.

[Scipy](https://www.geeksforgeeks.org/data-analysis-with-scipy/) is a mathematical library in Python that is mostly used for mathematical equations and computations.

We will implement our first real life problem via python,

See Example and calculation.

<https://www.geeksforgeeks.org/z-test/>

<https://www.javatpoint.com/hypothesis-testing-python>