DATA FOLKZ

CATAPULT DATA LEADERS

UNDERSTANDING

Statistics

TABLE OF CONTENT

Basics of Statistics **2**

What is statistics ? 2

How is statistics used in Data Science ? 2

Population and Sample 2

Data Types 3

Categories In Statistics 5

Sampling **6**

What is Sampling? 6

Sampling techniques 6

Understanding Descriptive Analysis **10**

Univariate and Bivariate Analysis 10

Measures of Descriptive Statistics 10

Measures of central tendency 11

Measures of dispersion 12

What are absolute measures and Relative measures? 13

Correlation and Causation 16

Skewness 19

Kurtosis 20

Box plots 21

Basics of Statistics

What is statistics ?

Statistics is the science of learning from data that is concerned with collection , presentation , description and analysis of data which is measurable in numerical terms. Statistics is used to process complex problems in the real world so that we can look for meaningful trends and changes in data.

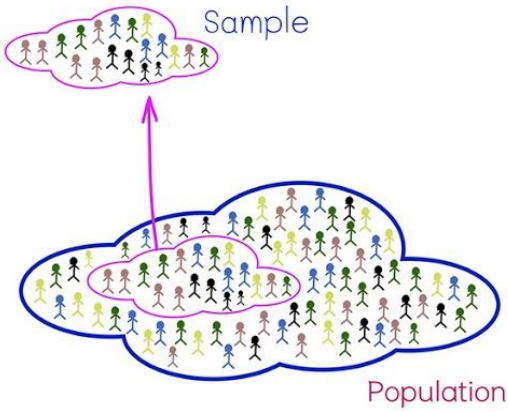
How is statistics used in Data Science ?

Statistical analysis is the science of collecting data and uncovering patterns and trends. It is needed for following tasks:

- Simplifying mass of data : Statistics helps to convert mass of data into significant figures that makes analysis easier.
- For Making future predictions based on past behavior.
- Presenting facts in definite format : Stats enables us to present general statements in a precise and definite format like numbers . Numerical values are more informative than statements.
- Facilitating comparisons of data : Statistics enables comparison between two similar entities by using their individual data and figures.
- Statistical methods help to formulate and test hypothesis to develop new theories

Population and Sample

- A population is the collection of all items of interest to our study and is usually denoted with an uppercase N. The numbers we obtain when using a population are called parameters.
- A Sample is a subset of the Population and is usually denoted with lowercase n, and the numbers we obtain when working with a sample are called statistics.



Data Types

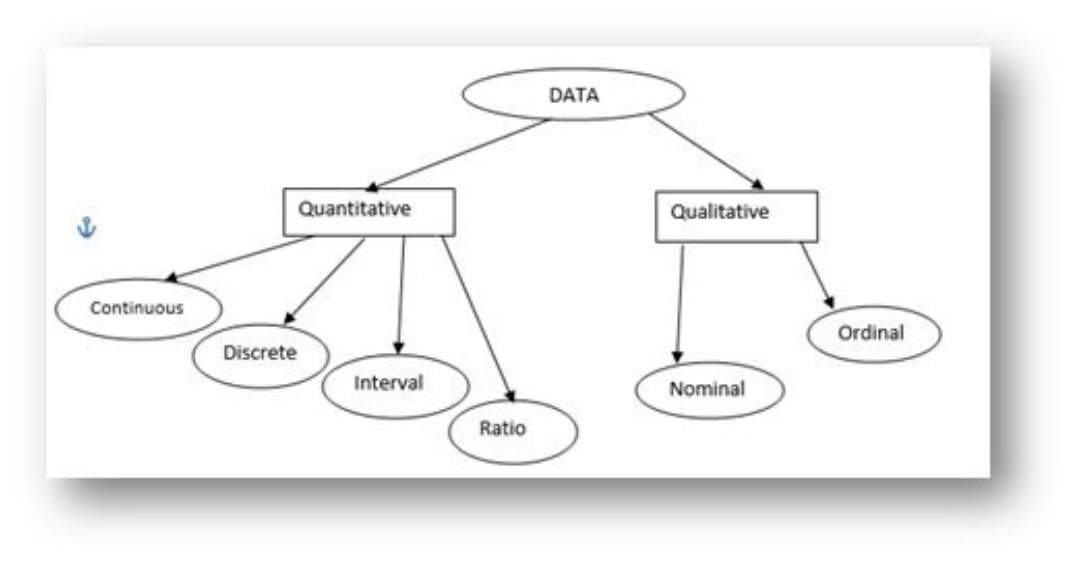
Data may be qualitative or quantitative

Quantitative : It is used to represent data that is associated with a numerical value. For example , the height of a student could be 1.80 m or weight could be 60 kg or the marks of a student could take any value till 100 .

Qualitative : They represent some characteristics or attributes. They depict descriptions that may be observed but cannot be computed or calculated. For example, data on attributes such as intelligence, honesty, wisdom or gender of a person or blood group .

Qualitative Data	Quantitative Data
Overview: • Deals with descriptions. • Data can be observed but not measured. • Colors, textures, smells, tastes, appearance, beauty, etc. • **Qualit**ative → **Qualit**y	**Overview:** • Deals with numbers. • Data which can be measured. • Length, height, area, volume, weight, speed, time, temperature, humidity, sound levels, cost, members, ages, etc. • **Quantit**ative → **Quantit**y

Quantitative and qualitative data can be broken into further sub-categories.



Quantitative

- Continuous
- Discrete
- Interval
- Ratio

Continuous Data: These are data that can take values between a certain range. For example, temperature can take values even in decimals and so is the case of the height and weights of the students.

Discrete Data: These are data that can take only certain specific values rather than a range of values. Their values can only be a whole number .
For example , number of students in class , number of mobiles sold in a month.

Interval Data: It is the type of data that can be measured or placed on the ordered and constant scale of measurement. The classic example of an interval scale is Celsius temperature.
For example, the difference between 60 and 50 degrees is a measurable 10 degrees, as is the difference between 80 and 70 degrees.

Ratio Data : It is the type of data that has natural zero as a starting point .For ratio data, it is not possible to have negative values.
For example , age, length , marks and income . Zero length , zero marks , zero income .

Qualitative

- Nominal
- Ordinal

Nominal data : Nominal data is related to identification of categories. Common examples include male/female, hair color, nationalities, and names of people.For example, race is a nominal variable having a number of categories, but there is no specific way to order from highest to lowest and vice versa.

Ordinal Data: Ordinal data is a type of categorical data with an order. The variables in ordinal data are listed in an orderly manner. For example, ranks (1,2,3)

Categories In Statistics

There are two main categories in Statistics, namely:

1. Descriptive Statistics
2. Inferential Statistics

Descriptive Statistics uses the data to provide descriptions of the population, either through numerical calculations or graphs or tables.

For example , We want to know the average marks of students in a classroom, In descriptive statistics we will record the marks of each and every student in the class and then find out the maximum, minimum and average marks of the class.

Inferential Statistics makes inferences and predictions about a population based on a sample of data taken from the population.

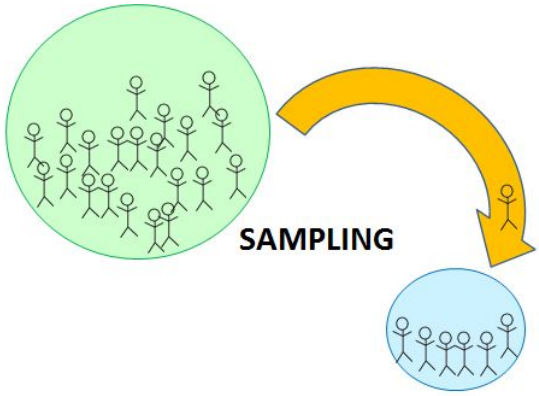
Example, Now we want to know the average marks of all the students studying in the same grade like 11th standard . Now it will be difficult to record marks of each student separately . So ,in Inferential Statistics, we will take a sample from the whole population and consider this sample for our statistical study (calculating average marks) for studying the population .

Sampling

What is Sampling?

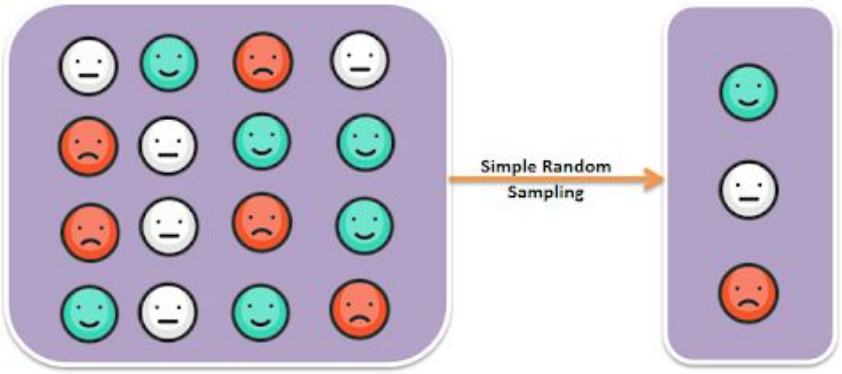
Sampling is a process used in statistical analysis in which a predetermined number of observations are taken from a larger population. The methodology used to sample from a larger population depends on the type of analysis being performed .

The chosen sample should be a fair representation of the entire population. When taking a sample from a larger population, it is important to consider how the sample is chosen.



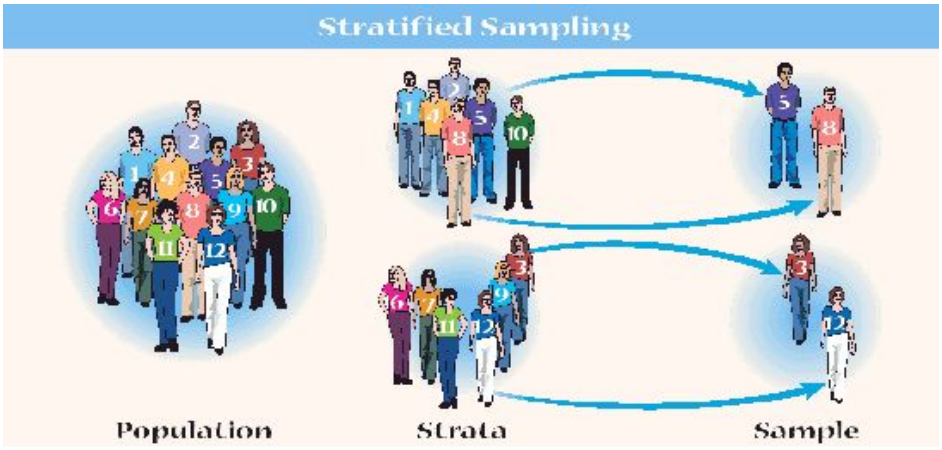
Sampling techniques

Simple Random Sampling : Every element has an equal chance of getting selected to be the part sample. It is used when we don't have any kind of prior information about the target population. It is removed from any potential bias because there is no human judgement involved in selecting the sample.
For example: Random selection of 20 students from a class of 50 students. Each student has an equal chance of getting selected. Here probability of selection is 1/50



Stratified Sampling : In this method, the population is first divided into subgroups (or strata) who all share a similar characteristic and then the elements are randomly selected from each of these strata. We need to have prior information about the population to create subgroups.

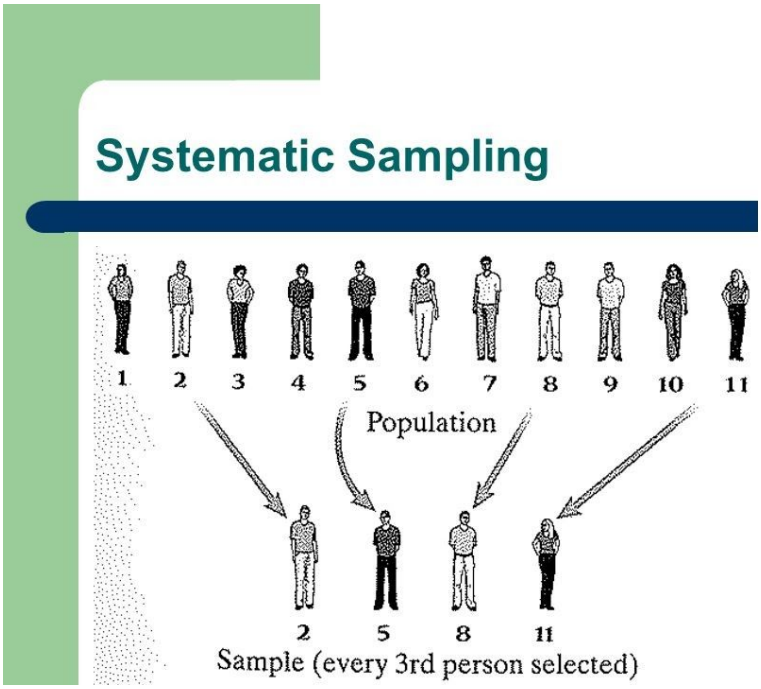
For example, in a study of stroke outcomes, we may stratify the population by sex, to ensure equal representation of men and women.



Systematic Clustering : Individuals are selected at regular intervals from the sampling frame. The intervals are chosen to ensure an adequate sample size. If you need a sample size *n* from a population of size *x*, you should select every *x/n*th individual for the sample.

For example, if you wanted a sample size of 100 from a population of 1000, select every 1000/100 = 10th member of the sampling frame.

Systematic sampling is often more convenient than simple random sampling. However, it may also lead to bias.

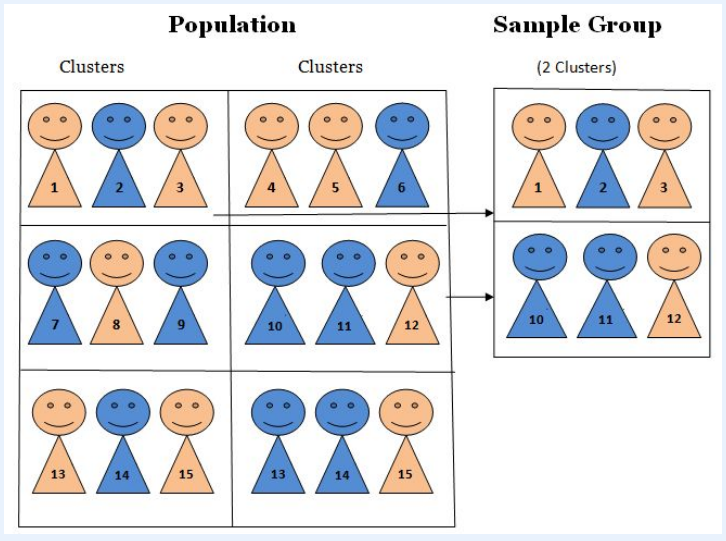


Cluster Sampling : Whole population is divided into subgroups, known as clusters, which are randomly selected to be included in the study. All the elements of the cluster are used for sampling.

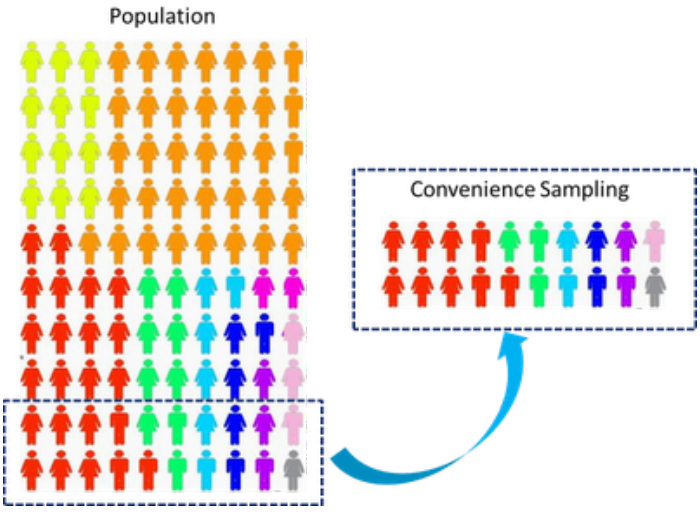
Cluster sampling can be more efficient than simple random sampling, especially where a study takes place over a wide geographical region.

In single-stage cluster sampling, Entire cluster is selected randomly for sampling.

In two-stage cluster sampling, first we randomly select clusters and then from those selected clusters we randomly select elements for sampling.



Convenience Sampling : It is one of the Non-Probability Sampling Methods . Convenience sampling is perhaps the easiest method of sampling, because participants are selected based on availability and willingness to take part.This method is used when the availability of samples is rare and also costly. So based on the convenience samples are selected.Useful results can be obtained, but the results are prone to significant bias, because those who volunteer to take part may be different from those who choose not to (volunteer bias), and the sample may not be representative of other characteristics, such as age or sex.



Understanding Descriptive Analysis

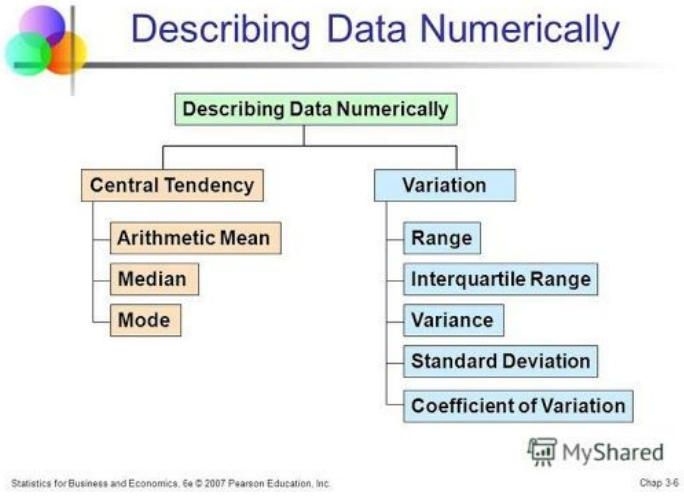
Descriptive statistics helps to describe and understand the features of a specific dataset . With descriptive statistics you are simply describing what is or what the data shows.

Univariate and Bivariate Analysis

1. Univariate data – This type of data consists of **only one variable**. The analysis of univariate data is thus the simplest form of analysis since the information deals with only one quantity that changes. It does not deal with causes or relationships and the main purpose of the analysis is to describe the data and find patterns that exist within it. For example, height .There is only one variable that is height and it is not dealing with any cause or relationship.

2. Bivariate data – This type of data involves **two different variables**. The analysis of this type of data deals with causes and relationships and the analysis is done to find out the relationship among the two variables.
Example of bivariate data can be temperature and ice cream sales in summer season.

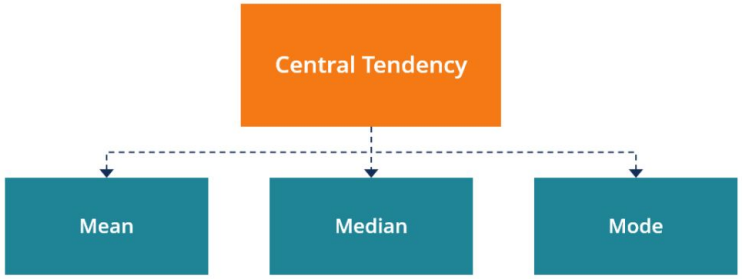
Measures of Descriptive Statistics



Descriptive statistics consists of two basic categories of measures: measures of central tendency and measures of variability or spread.

Measures of central tendency

They describe the center of a data set. Although it does not provide information regarding the individual values in the dataset, it delivers a comprehensive summary of the whole dataset.



Central Tendency can be measured in three different ways; namely mean, median and mode.

Mean: It is the average which is simply defined as the ratio of the summation of all values to the number of items. Let's look at an example of simple set of data representing the weight of 10 males,
55, 56, 56, 58, 60, 61, 63, 64, 70, 78.
The mean weight is calculated as,
Mean = (55 + 56 + 56 + 58 + 60 + 61 + 63 + 64 + 70 + 78) / 10 = 62.1

Mean Formula = $\sum X \div N$
$\sum X$= Sum of all the individual values,
N= Total number of items

Median: It is essentially known as the central value of a series. Median of a set of values can be arrived only after sorting the data in either ascending or descending order

X = 3, 6, 8, 3, 9, 1, 7

Sorted X = 1, 3, 3, 6, 7, 8, 9

Median of X = 1, 3, 3, **6**, 7, 8, 9

Median = 6

When the count of numbers is even: Median= **(n/2) +1**

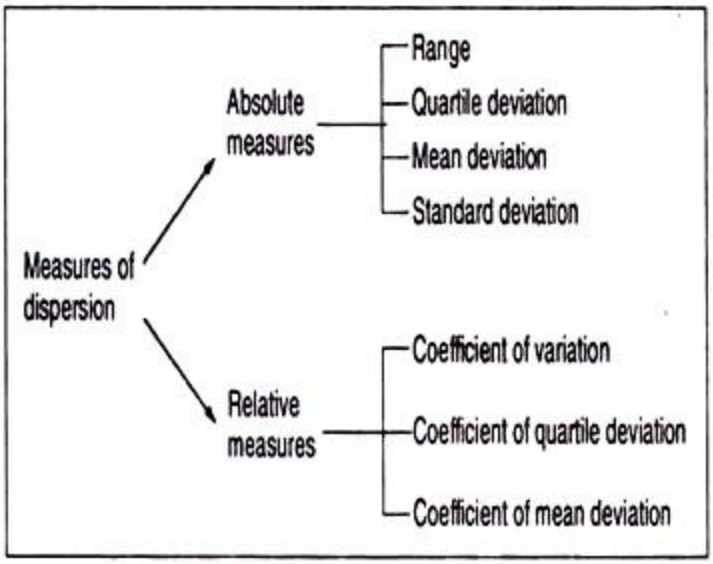
When the count of numbers is odd: Median= **(n+1)/2**
(n is the count of numbers in the given data)

Mode: Mode is the most frequently occurring number in the dataset.
Let us take an example of mode 89, 65, 11, 54, 11, 90, 56.Here in these varied observations the most occurring number is 11, hence Mode=11

Measures of dispersion

Dispersion is used to measure the variability in the data or to see how spread out the data is.In simple words dispersion in statistics is a way of describing how spread out a set of data is.

The spread of a data set can be understood by a range of descriptive statistics including variance, standard deviation, and interquartile range. Spread can also be shown in graphs: dot plots, box plots, and stem and leaf plots have a greater distance with samples that have a larger dispersion and vice versa.



Two types of method of dispersion :

- Absolute Measures
- Relative Measures

What are absolute measures and Relative measures?

An **absolute measure** is a term that defines the uses of numerical variations to determine the degree of error Absolute measures take the form of positive numbers, regardless of whether they represent high or low estimations. For example, they are used like cm, kg, Rs, etc. Most commonly used are standard deviation, mean deviation, range.

Relative measures are just an alternative to Absolute measures. They use statistical variations based on percentages to determine how far from reality a figure is within context .They are free from measuring units label's. Relative measures are coefficient of range, coefficient of standard deviation, coefficient of mean deviation etc

Application of dispersion : *For example*, pretend that you want to sell your house. You narrow your search to two companies: Magicbricks.com and Makaan.com. Both companies advertise that sellers receive, on average, 90% of their asking price. Does it matter which company you choose?

The real question is, does the **mean** (average), describe the data accurately enough to make an informed decision? No, it doesn't. The mean is not a reliable predictor; it only describes the data set as a whole and doesn't tell what's happening within the set.

Let's add some data to the example to illustrate the point. Assume the following shows the percent of asking price received on the previous nine sales for each company:

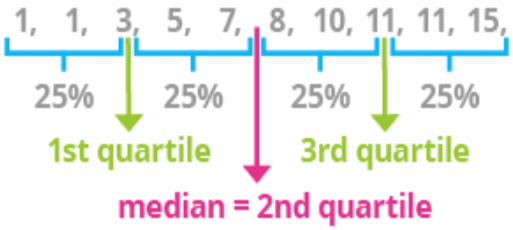
- Magic Bricks: 88, 92, 91, 89, 89, 91, 91, 89, and 90
- Makaan, 100, 100, 83, 100, 95, 86, and 90

How can you make an informed decision about which company will offer you the greatest benefit for the least risk?

You must analyze the **dispersion**, the amount of variation within a data set, of each set. Only then will you be able to truly compare these two companies

Range: It is the given measure of how spread apart the values in a data set are. It is measured as= (highest value – lowest value) of the variable.

Quartile deviation : A median divides a given dataset (which is already sorted) into two equal halves similarly, the quartiles are used to divide a given dataset into four equal halves**.**



Q1=the lowest 25% of numbers

Q2=the next lowest 25% of numbers (up to the median).

Q3=the second highest 25% of numbers (above the median).

Q4=the highest 25% of numbers.

$$Q_d = \frac{Q_3 - Q_1}{2}$$

$$\text{Coefficient of Quartile Deviation} = \frac{Q_3 - Q_1}{Q_3 + Q_1} \times 100$$

Standard deviation: Standard deviation is the square root of the mean of squared deviations from the arithmetic mean.

$$\sigma = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2}$$

Variance : Variance is the square of standard deviation.

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

Mean Deviation: It is same as standard deviation just mean of all deviation

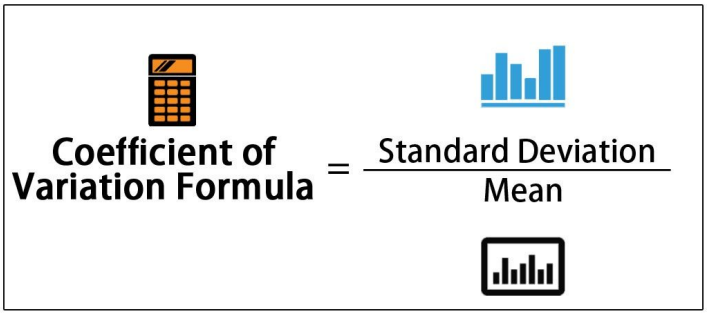
Mean Absolute Deviation Formula

$$\frac{\Sigma\ |x - \bar{x}|}{n}$$

Formula:

$$\text{Coefficient of MD} = \frac{Mean\ Deviation}{Median\ or\ Mean} \times 100$$

Coefficient of Variation : SD is the absolute measure of dispersion. The relative measure of dispersion based on standard deviation is known as coefficient of standard deviation.

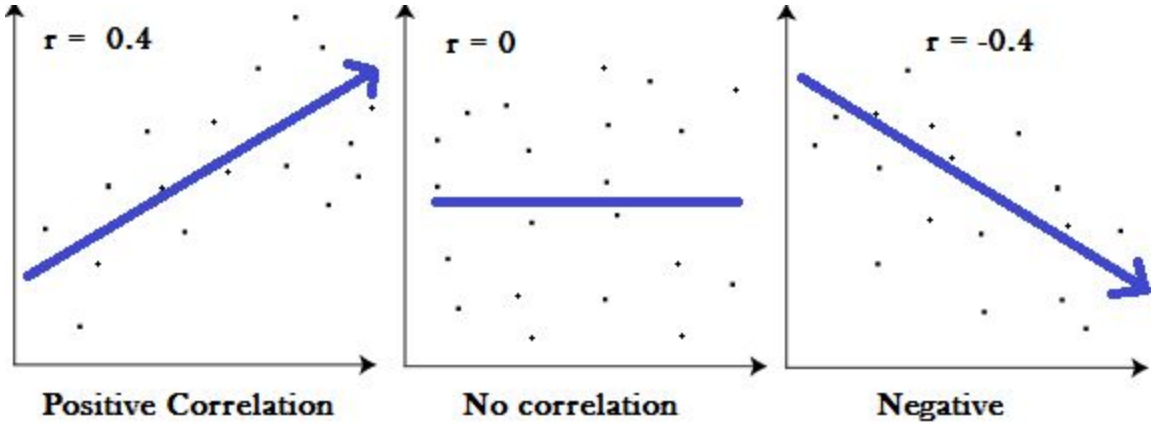


Correlation and Causation

Correlation is a statistical measure (expressed as a number) that describes the size and direction of a relationship between two or more variables. For example income and expenditure are correlated to each other.When income increases expenditure also increases. However expenditure does not increase in the proportion of income.Such relationships amongst the variables are determined by correlation.

Causation indicates that one event is the result of the occurrence of the other event; i.e. there is a causal relationship between the two events. This is also referred to as cause and effect.
(e.g. smoking causes an increase in the risk of developing lung cancer)

Correlation can be classified into following categories:
Positive correlation
Negative correlation
No correlation



The most common measure of correlation is Pearson's product-moment correlation, which is commonly referred to simply as the correlation, the correlation coefficient, or just the letter r
The value of the correlation coefficient ranges from [-1 – 1]. -1 stand for the negative relationship. 1 means a positive relationship. 0 means no relationship.

If the correlation coefficient has a negative value (below 0) it indicates a negative relationship between the variables. This means that the variables move in opposite directions (i.e. when one increases the other decreases, or when one decreases the other increases).
If the correlation coefficient has a positive value (above 0) it indicates a positive relationship between the variables meaning that both variables move in tandem, i.e. as one variable decreases the other also decreases, or when one variable increases the other also increases.
Where the correlation coefficient is 0 this indicates there is no relationship between the variables (one variable can remain constant while the other increases or decreases).

Formula of Correlation:

Pearson Correlation Coefficient

$$r = \frac{n(\Sigma xy) - (\Sigma x)(\Sigma y)}{\sqrt{[n\Sigma x^2 - (\Sigma x)^2][n\Sigma y^2 - (\Sigma y)^2]}}$$

Covariance is used to determine how much two random variables vary together. A positive covariance means that the two random variables move together while a negative covariance means they move inversely.
Covariance is not standardized, unlike the correlation coefficient. Therefore, covariance values can range from negative infinity to positive infinity.

$$COV(X,Y) = \frac{\sum_{i=1}^{n}\left(X_i - \bar{X}\right)\left(Y_i - \bar{Y}\right)}{n-1}$$

Example: Ice Cream Sales

The local ice cream shop keeps track of how much ice cream they sell versus the temperature on that day, here are their figures for the last 12 days:

Ice Cream Sales vs Temperature	
Temperature °C	**Ice Cream Sales**
14.2°	$215
16.4°	$325
11.9°	$185
15.2°	$332
18.5°	$406
22.1°	$522
19.4°	$412
25.1°	$614
23.4°	$544
18.1°	$421
22.6°	$445
17.2°	$408

Calculation of correlation:

Step 1: Find the mean of x, and the mean of y

Step 2: Subtract the mean of x from every x value (call them "a"), do the same for y (call them "b"

Step 3: Calculate: ab, a^2 and b^2 for every value

Step 4: Sum up ab, sum up a^2 and sum up b^2

Step 5: Divide the sum of ab by the square root of [(sum of a^2) × (sum of b^2)]

2 *Subtract Mean*

3 *Calculate ab, a^2 and b^2*

Temp °C	Sales	"a"	"b"	a×b	a^2	b^2
14.2	$215	-4.5	-$187	842	20.3	34,969
16.4	$325	-2.3	-$77	177	5.3	5,929
11.9	$185	-6.8	-$217	1,476	46.2	47,089
15.2	$332	-3.5	-$70	245	12.3	4,900
18.5	$406	-0.2	$4	-1	0.0	16
22.1	$522	3.4	$120	408	11.6	14,400
19.4	$412	0.7	$10	7	0.5	100
25.1	$614	6.4	$212	1,357	41.0	44,944
23.4	$544	4.7	$142	667	22.1	20,164
18.1	$421	-0.6	$19	-11	0.4	361
22.6	$445	3.9	$43	168	15.2	1,849
17.2	$408	-1.5	$6	-9	2.3	36
18.7	**$402**			**5,325**	**177.0**	**174,757**

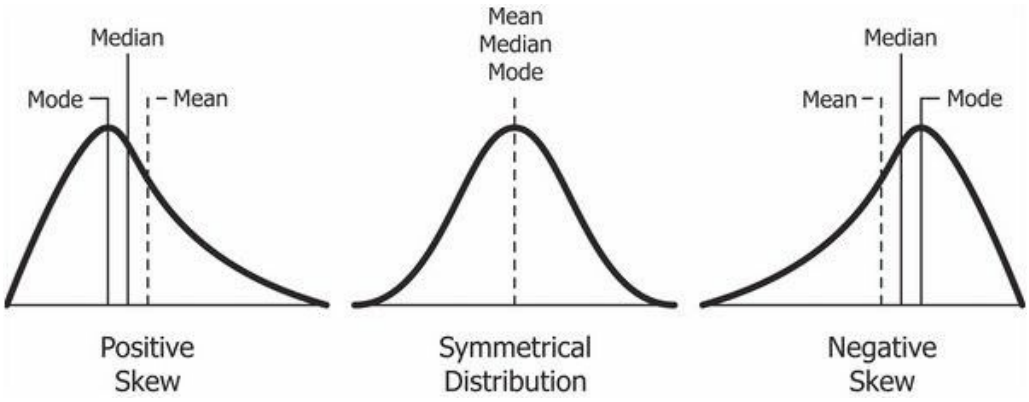
1 *Calculate Means*

4 *Sum Up*

5 $\frac{5{,}325}{\sqrt{177.0 \times 174{,}757}} = 0.9575$

Skewness

It is the degree of distortion from the symmetrical bell curve or the normal distribution. It measures the lack of symmetry in data distribution . A symmetrical distribution will have a skewness of 0 . When the skewness is 0 i.e when distribution is not skewed then the centrality measure used is mean. Usually in this case mean and median are equal.



Positive Skewness means when the tail on the right side of the distribution is longer or fatter. In this case mean is larger than median.

Example : Distribution of Income- If the distribution of the household incomes of a region is studied, from values ranging between $5,000 to $250,000, most of the citizens fall in the group between $5,000 and $100,000, which forms the bulk of the distribution towards the left side of the distribution, which is the lower side. However, a couple of individuals may have a very high income, in millions. This makes the tail of extreme values (high income) extend longer towards the positive, or right side. Thus, it is a positively skewed distribution.

Negative Skewness is when the tail of the left side of the distribution is longer or fatter than the tail on the right side. In this case mean is smaller than median. In both positive and negative skewed cases median will be preferred over mean.

Example : Retirement Age - When the retirement age of employees is compared, it is found that most retire in their mid-sixties, or older. Thus, the distribution of most people will be near the higher extreme, or the right side. However, there is an increasingly new trend in which very few people are retiring early, and that too at very young ages. This will make the tail of the distribution longer towards the left side or the lower side, and the less values (low ages) will shift the mean towards the left, making it a negatively skewed distribution.

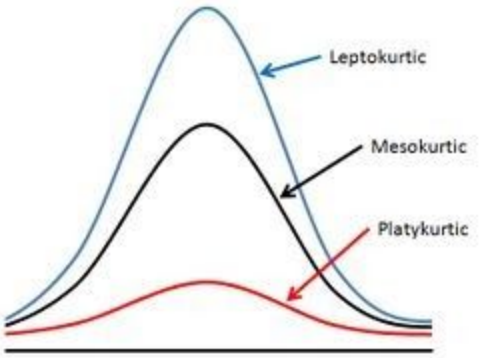
- If the skewness is between -0.5 and 0.5, the data are fairly symmetrical.
- If the skewness is between -1 and -0.5(negatively skewed) or between 0.5 and 1(positively skewed), the data are moderately skewed.
- If the skewness is less than -1(negatively skewed) or greater than 1(positively skewed), the data are highly skewed.

Kurtosis

It is the sharpness of the peak of a frequency-distribution curve.It is actually the measure of outliers present in the distribution.

High kurtosis in a data set is an indicator that data has heavy outliers.

Low kurtosis in a data set is an indicator that data has lack of outliers.



Mesokurtic: This distribution has kurtosis statistic similar to that of the normal distribution.

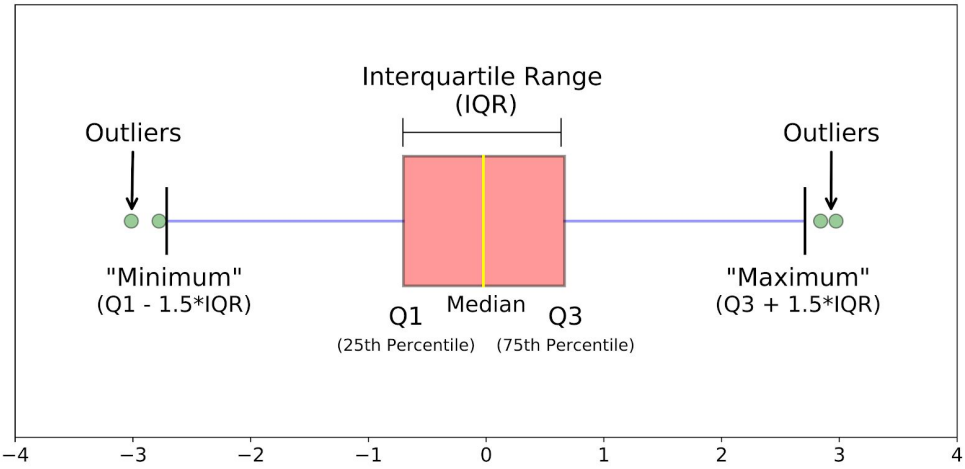
Leptokurtic (Kurtosis > 3) : Peak is higher and sharper than Mesokurtic, which means that data has heavy outliers.

Platykurtic (Kurtosis < 3): The peak is lower and broader than Mesokurtic, which means that data has a lack of outliers.

Box plots

Box plots (also called box-and-whisker plots or box-whisker plots) give a good graphical image of the concentration of the data. They also show how far the extreme values are from most of the data.
A box plot is constructed from five values: the minimum value, the first quartile, the median, the third quartile, and the maximum value. We use these values to compare how close other data values are to them.The very purpose of this diagram is to identify outliers.

Outlier is a value that lies in a data series on its extremes, which is either very small or large and thus can affect the overall observation made from the data series.


Interquartile Range
(IQR)
Outliers
Outliers
"Minimum"
(Q1 - 1.5*IQR)
"Maximum"
(Q3 + 1.5*IQR)
Q1
Median
Q3
(25th Percentile)
(75th Percentile)
−4
−3
−2
−1
0
1
2
3
4

DATA FOLKZ

CATAPULT DATA LEADERS

Thank You

www.datafolkz.co.in