Lesson:

Statistics

PW SKILLS

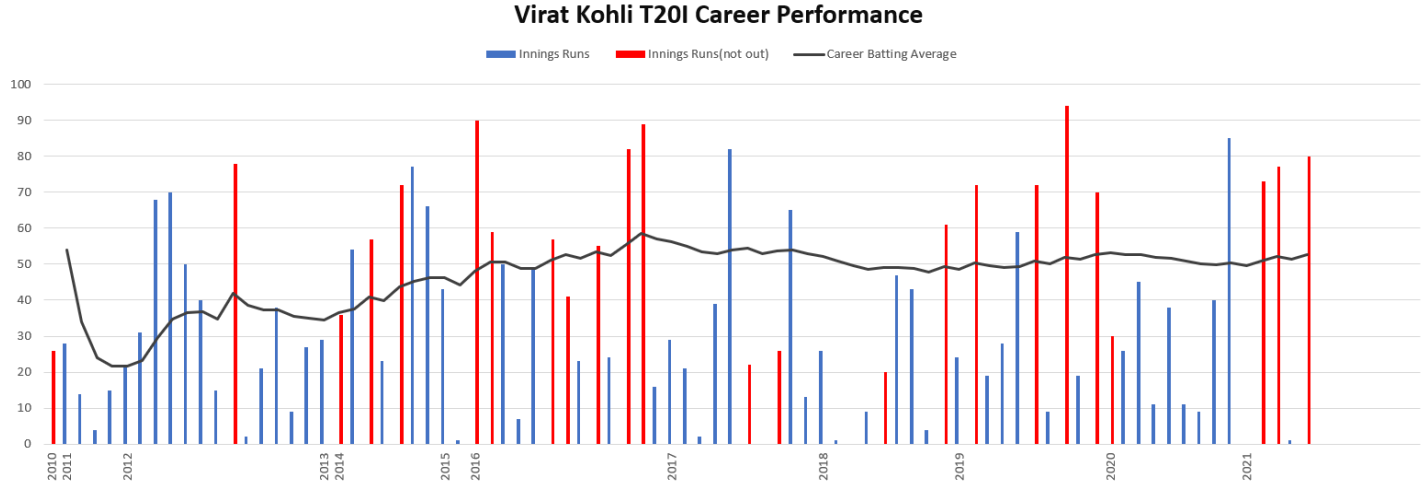
Topics To Be Covered

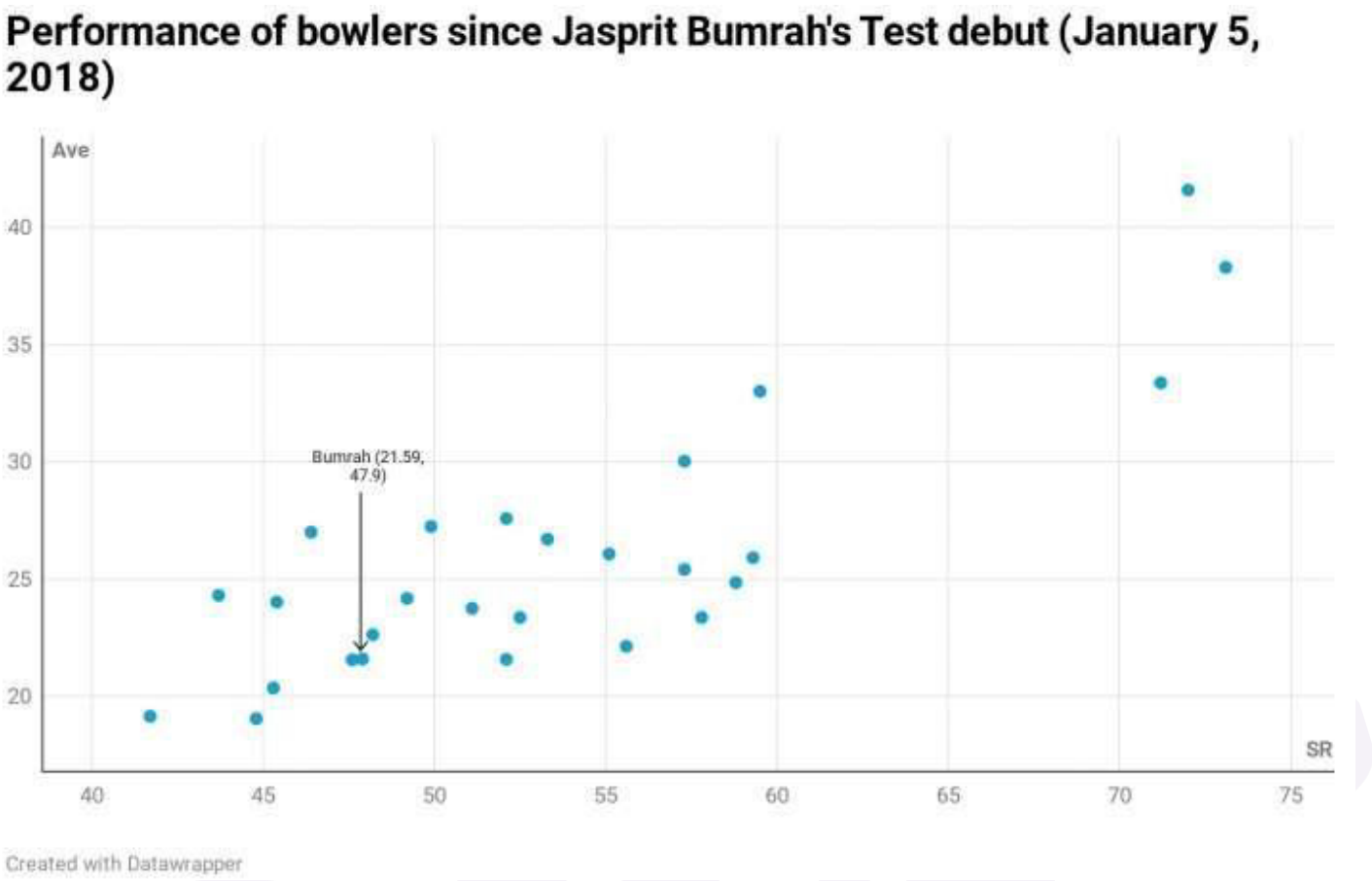
1. Types of Statistics
2. Descriptive
3. Inferential
4. Population and Sample
5. Parameter and Statistics (Mean, Median, Mode, Std, Variance)
6. Uses of variable
7. Dependent
8. Independent variable
9. Types of Variable
10. Continuous
11. Categorical variable
12. Distribution types and Skewness
13. Hypothesis testing
14. Type 1 Error
15. Type 2 Error
16. T-Test (one sample and sample)
17. ANOVA & CHI_SQUARE
18. Covariance and Correlation

TYPES OF STATISTICS

Statistics is the science that deals with methodologies to gather, review, analyse and draw conclusions from data. With specific Statistics tools in hand we can derive many key observations and make predictions from the data in hand. In the Real world, we deal with many cases where we use Statistics knowingly or unknowingly. Let's talk about one such classic use of statistics in the most famous sports in India, yes you guessed it right, Cricket. What makes Virat Kohli the best batsman in ODIs or Jaspreet Bumrah the best bowler in ODIs?

We all have heard about cricketing terms like batting average, bowler's economy, strike rate etc. We often see graphs like these





We see and talk about statistics all the time but very few of us know the science behind it.

Using different statistical methods, ICC compares players and teams and ranks them. So, if we learn the science behind it we can create our own rankings, compare players, teams or better if we debate with someone over who is the better player, we can debate now with facts and figures because we will understand the statistics behind it better. We can understand the above graphs better.

We will dive further into the various methods and terminologies which will help to answer the question above as well as see the vast uses of Statistics in much complex scenarios such as medical science, drug research, stock markets, Economics, Marketing etc.

DESCRIPTIVE STATISTICS

The type of statistics dealing with numbers (numerical facts, figures, or information) to describe phenomena. These numbers are descriptive statistics. They are used to describe, summarise the characteristics of a sample or dataset, such as variables mean, standard deviation, or frequency etc.

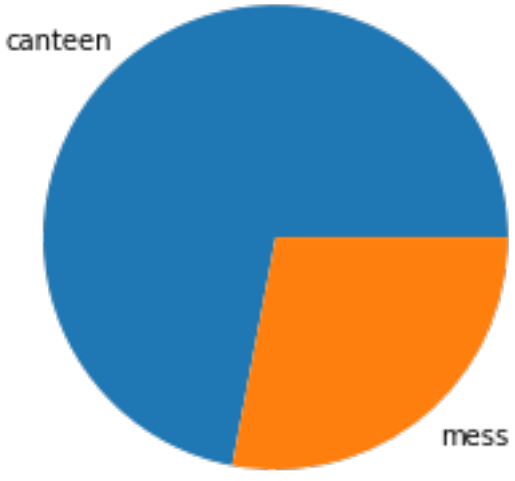
e.g. Reports of industry production, cricket batting averages, government deficits, Movie Ratings etc.

When we see an IMDB rating of any movie or the rating of any product on e-commerce websites, they are the average of all the ratings provided by many customers over the period. And that keeps on changing as new ratings are added to them regularly.

Suppose in your college there are 1000 students. You are interested in finding out how many students prefer eating in the college canteen to the college mess. A random group of 100 students is selected. Here our population is 1000 students, and the sample size is 100 students. You surveyed the sample group and got the following results:

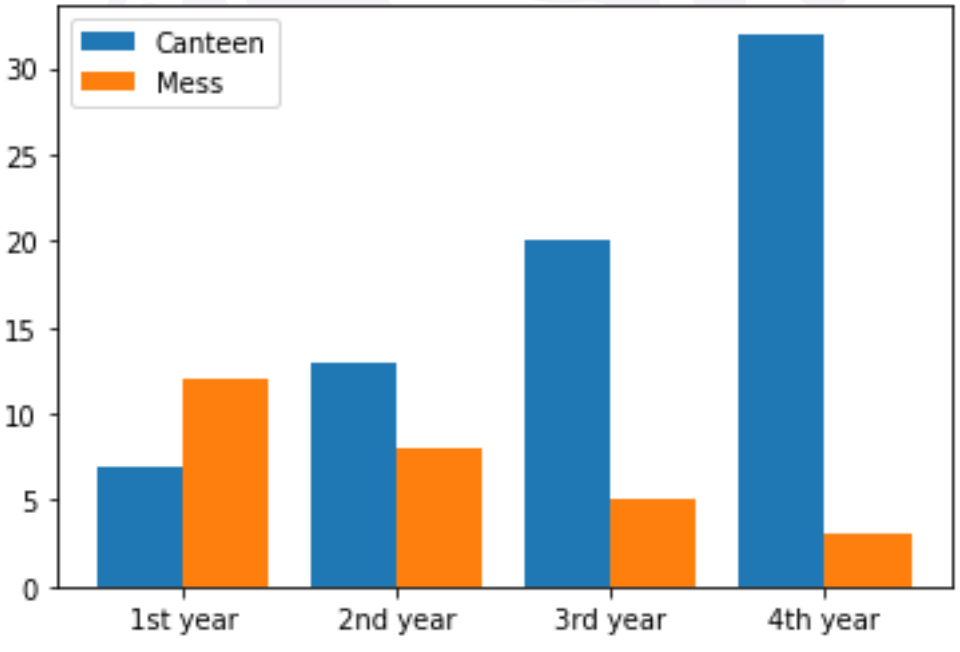
Year	1st	2nd	3rd	4th	Total
Canteen	7	13	20	32	72
Mess	7	8	5	3	28

Now, If we draw a pic chart for the same, It will look something like below.



Pie Chart

A Bar Chart will look like the below



Bar Graph

From the above visuals, we can see that

1. 72 % of the students prefer eating in the canteen.

2. Of the total students who prefer the canteen, 44.4 % are from the 4th year.
3. Of the total students who prefer the canteen, 72% are from the 3rd year and 4th year.
4. 1st-year students are more inclined towards eating in the mess.

The above statistics give us a variation trend among the students with their preference. We are using the numbers and figures to assess the data. This will be part of Descriptive statistics.

INFERENTIAL STATISTICS

Inferential statistics is a decision, estimate, prediction, or generalisation about a population based on a sample. Here we deal with the sample/samples and compare them and perform some tests to draw some conclusions on the population data based on the observations from the sample.

- A population is a collection of all possible individuals, objects, or measurements of interest.
- A sample is a portion, or part, of the population of interest.
- Inferential statistics is used to make inferences from data, whereas descriptive statistics describe what's happening in our data.

Suppose you got a contract to open a canteen in the College. Now with the above data, you can make the following assumptions:

1. 3rd and 4th-year students are the main target for restaurant sales.
2. You can discount the 1st year students to increase the number count.
3. Since most students prefer eating in the canteen, opening a canteen can be a profitable business

Based on the sample data, you made the above inferences/estimations for the whole college. This is part of Inferential statistics, where you make decisions based on the descriptive statistics of a sample data.

POPULATION AND SAMPLE

Population

In statistics, population refers to the entire set of items, individuals, or data points that are of interest for a particular study or analysis. This could include all people, objects, events, measurements, or any other entities that are being studied. The population is the complete collection that you want to conclude about.

For example, if you're conducting a study on the average height of all adult males in a country, the population would consist of the heights of every adult male in that country. Similarly, if you're studying the sales data for a specific product across all stores in a particular chain, the population would be the sales figures from all those stores.

It's important to note that analyzing an entire population is often not feasible due to factors like time, resources, and accessibility. Researchers often use a subset of the population, called a sample, to make inferences and draw conclusions about the entire population. This process is the basis of statistical inference, where you analyze the characteristics of a sample to make educated guesses about the characteristics of the larger population from which the sample was drawn.

Sample

In statistics, a sample refers to a subset of individuals, items, or data points that are selected from a larger population in order to gather information, make inferences, or draw conclusions about that population. Sampling is a practical way to study a population without having to examine every single element within it.

Sampling involves carefully selecting a representative subset of the population to ensure that the characteristics and attributes of the sample reflect those of the entire population as accurately as possible. If the sample is chosen correctly and is truly representative, the statistical analysis of the sample can provide valuable insights into the larger population.

There are different methods of sampling, including:

- **Random Sampling:** Every individual in the population has an equal chance of being selected. This helps minimize bias and increase the likelihood that the sample represents the population accurately.
- **Stratified Sampling:** The population is divided into distinct subgroups (strata) based on certain characteristics, and then a random sample is taken from each subgroup proportionate to its size. This ensures representation from various subgroups.
- **Systematic Sampling:** Every nth individual is selected from the population after an initial random start. For example, you might select every 10th person from a list.
- **Cluster Sampling:** The population is divided into clusters (e.g., geographic regions), and a random sample of clusters is selected. Then, all individuals within the selected clusters are included in the sample.
- **Convenience Sampling:** This involves selecting individuals who are most readily available or accessible. While it's convenient, it can introduce bias and might not be representative.
- **Purposive Sampling:** This is a non-random method where specific individuals are chosen intentionally because they possess certain characteristics of interest.

The choice of sampling method depends on the research objectives, the population being studied, available resources, and the desired level of accuracy. The goal is to ensure that the sample is as representative of the population as possible so that the conclusions drawn from analyzing the sample can be generalized to the larger population.

PARAMETER AND STATISTICS

Parameter and Statistic

To avoid verbal confusion with the statistical constants of the population, mean(μ), variance(σ^2) etc are called parameters, statistical measures computed from the sample observations alone eg. mean($\bar{x}$), variance (s^2), etc. have been termed as statistics.

Measure of central tendency

A measure of central tendency is a summary statistic that represents the center point or typical value of a dataset. These measures indicate where most values in a distribution fall and are also referred to as the central location of a distribution. You can think of it as the tendency of data to cluster around a middle value. In statistics, the three most common measures of central tendency are the mean, median, and mode. Each measure calculates the central point's location using a different method.

Mean:

The mean is the arithmetic average; for calculating the mean add up all of the values and divide by the number of observations in your dataset.

Let you have a dataset with n values as follows: D=X1, X2,X3, X4,.....Xn

$$Mean(\bar{x}) = \Sigma_{i=1}^{n} x_i/n$$

Here n is the size of the data set, x is the sample mean, and x_i the numbers in sequence.

∑ is the summation of the entire data set.

Similarly, for a data population of size N, the population mean is

$$\mu = \frac{1}{N}\sum_{i=1}^{N} x_i$$

Example:

The systolic blood pressure of seven middle aged men in 150,123,134,170,146,124, and 113.
The Mean is = (150+123+134+170+146+124+113)/7=137.14

Median:

The Median for the sample data arranged in increasing order is defined as:

1. if “n” is an odd number then median is middle value
2. if “n” is an even number then median is midway between the two middle values

Example:

1. if “n” is odd:
   The re-ordered systolic blood pressure data: 113,124,125,132,146,151,170
   The median here is 132.

2. if “n” is even:
   The new blood pressure data is : 125,176,155,149,163,178
   arrange the new blood pressure data in ascending order: 125,149,155,163,176,178
   The median is here is (155+163)/2=159

Mode:

The mode is the value that occurs the most frequently in your data set i.e. has the highest frequency. On a bar chart, the mode is the highest bar. If the data have multiple values that are tied for occurring the most frequently, you have a multimodal distribution. If no value repeats, the data do not have a mode.

Example:

Given a data set of height(cm) of students in a class is:
Height=180,167,154,142,147,154,162,154
Here 154 is repeated 3 times so mode is here 154

Mean, Median, Mode for Group Data:

Mean: Mean for grouped is calculated the same way as we do in ungrouped data, just the variable(x) becomes the midpoint of the interval.

Median:

$$Median = l + \frac{h}{f}\left(\frac{N}{2} - c\right)$$

Where:

l = lower class boundary of the median class

h = Size of the median class interval

f = Frequency corresponding to the median class

N = Total number of observations i.e. sum of the frequencies

c = Cumulative frequency preceding median class.

Mode:

$$Mode = l + h\left(\frac{f_m - f_1}{2f_m - f_1 - f_2}\right)$$

Where,

l = Lower Boundary of modal class

h = *size of model class*

f_m = *Frequency corresponding to modal class*

f_1 = *Frequency preceding to modal class*

f_2 = *Frequency proceeding to modal class*

Example:

Calculate the mean, median, mode for given below data:

Variable(x)	Frequency(f)	Cumulative Frequency(c.f)
0-10	3	3
10-20	5	8
20-30	7	15

Variable(x)	Frequency(f)	Cumulative Frequency(c.f)
30-40	9	24
40-50	4	28

Solution:

Group	Mid point(x)	Frequency(f)	Cumulative frequency(c.f)	f*x
0-10	5	3	3	15
10-20	15	5	8	75
20-30	25	7	15	175
30-40	35	9	24	315
40-50	45	4	28	180
Total=		**28**		**760**

Here for calculating the mean, we chose the midpoints of the groups as variable(x).

Mean = 760/28=27.14

Median:
Median class= Class with c.f value of (28/2)

Since 14 is not in the c.f column, the next closest value is 15.

Median class =[20-30]

Using the above formula for median:
l=20
h=10
f=7
N=28
c=8

Median = 20 +[(14-8) *10] / 7 = 28.57

Mode:
modal class =[30-40] (it is the group with the highest frequency 9)
l=30
h=10
f(m)=9
f1=7
f2=4
Mode=30+10*(9-7)/(2*9-7-4)=32.85

Measure of Dispersion

Averages give us an idea of the concentrations of the observations about the central part of the distribution. But only averages won't give us an complete idea about the distribution.
Consider these examples. Calculate the mean of these

= (7,8,9,10,11)

= (3,6,9,12,15)

= (1,5,9,13,17)

Mean of X : 9.0

Mean of Y : 9.0

Mean of Z : 9.0

Range

A range is the most common and easily understandable measure of dispersion. It is the difference between two extreme observations of the data set. If X max and X min are the two extreme observations then

$$Range = xmax - xmin$$

Since it is based on two extreme observations, it gets affected by fluctuations.

Thus, range is not a reliable measure of dispersion

Standard Deviation and Variance

In statistics, the standard deviation is a very common measure of dispersion. Standard deviation measures how spread out the values in a data set are around the mean. More precisely, it is a measure of the average distance between the values of the data in the set and the mean. If the data values are all similar, then the standard deviation will be low (closer to zero). If the data values are highly variable, then the standard variation is high (further from zero).
Let the population if n elements, $(x1, x2, ..., xn)$. The mean deviation of the data is

Where,
$\bar{x}$ = sample mean
The standard deviation is always a positive number and is always measured in the same units as the original data. Squaring the deviations overcomes the drawback of ignoring signs in mean deviations i.e. distance of points from mean must always be positive.
S^2
The Variance is defined as the average of the squared differences from the Mean.
Let the population if n elements, $(x1, x2, ..., xn)$ The mean deviation of the data is

So, Variance = Standard Deviation 22
Standard deviation is easier to interpret than variance because it is on same scale as of the given data. Mathematically working with squared values are preferred over absolute values especially for statistical models. Hence, we prefer standard deviation over mean absolute deviation.

Here are few use cases

A class of students took a test in Language Arts. The teacher determines that the mean grade on the exam is 65%. She is concerned that this is very low, so she determines the standard deviation to see if it seems that most students scored close to the mean, or not. The teacher finds that the standard deviation is high. After closely examining all of the tests, the teacher is able to determine that several students with very low scores were the outliers that pulled down the mean of the entire class's scores.

An employer wants to determine if the salaries in one department seem fair for all employees, or if there is a great disparity. He finds the average of the salaries in that department and then calculates the variance, and then the standard deviation. The employer finds that the standard deviation is slightly higher than he expected, so he examines the data further and finds that while most employees fall within a similar pay bracket, three loyal employees who have been in the department for 20 years or more, far longer than the others, are making far more due to their longevity with the company. Doing the analysis helped the employer to understand the range of salaries of the people in the department.

Coefficient of Variation(CV)

The coefficient of variation (CV), also known as relative standard deviation (RSD), is a standardized measure of dispersion of a probability distribution or frequency distribution. It is often expressed as a percentage, and is defined as the ratio of the standard deviation( ) to the mean( ). It gives the measure of variability

CV = Standard Deviation / Mean

Let's take one more example to try and understand how standard deviation and CV is helpful:

We are given batting score made by two batsmen in 10 matches:

Batsman	Match 1	Match 2	Match 3	Match 4	Match 5	Match 6	Match 7	Match 8	Match 9	Match 10	Sum	Mean
Batsman 1	54	35	68	12	13	120	6	0	18	184	510	51
Batsman 2	45	42	25	53	75	12	28	27	85	43	435	43.5

By seeing the above data, we can say that Batsman 1 is the better batsman than Batsman 2 and can be given preference since it's mean is greater. But is it really true? Let's check the variance of the data:

Batsman	Batsman 1	Batsman 2	diff_1	diff_2	var sqaure1 = (diff_1)^2	var sq_2 = (diff_2)^2
Match 1	54	45	-3	-1.5	9	2.25
Match 2	35	42	16	1.5	256	2.25
Match 3	68	25	-17	18.5	289	342.25
Match 4	12	53	39	-9.5	1521	90.25
Match 5	13	75	38	-31.5	1444	992.25
Match 6	120	12	-69	31.5	4761	992.25
Match 7	6	28	45	15.5	2025	240.25

Batsman	Batsman 1	Batsman 2	diff_1	diff_2	var sqaure1 = (diff_1)^2	var sq_2 = (diff_2)^2
Match 8	0	27	51	16.5	2601	272.25
Match 9	18	85	33	-41.5	1089	1722.25
Match 10	184	43	-133	0.5	17689	0.25
Sum	**510**	**435**			**31684**	**4656.5**

Variance (Batsman 1)= 31684/10 = 3168.4

Standard deviation(batsman1) = Variance (Batsman 1) ^ ½ = 56.288

Coefficient of Variation (batsman1) = 56.288/51 = 1.10

Variance (Batsman 2)= 4656.5/10 = 465.65

Standard deviation(batsman1) = Variance (Batsman 2) ^ ½ = 21.57

Coefficient of Variation (batsman2) = 21.57/43.5 = 0.50

We can clearly see that the standard deviation gives a different picture for both batsmen. Though batsman1 has a high average but his variance is very high. So, the batsman 1 is less reliable.

On the other hand, Batsman 2 has lower average but is much more consistent than Batsman 1.

Also, coefficient of variation for batsman 2 is lower than batsman 1 which ensures low variability and higher consistency.

If we only had taken mean into account, we wouldn't have gotten the true picture. The dispersion measures solve this problem.

Here we can see even though there are different values in the X , Y and Z. Mean of all the variables is same. Hence we need some other metrics also.

Here measure of dispersion helps us solve this issue.

The measure of dispersion shows the scatterings of the data. It tells the variation of the data from one another and gives a clear idea about the distribution of the data. The measure of dispersion shows the homogeneity or the heterogeneity of the distribution of the observations.

How is it useful?

Measures of dispersion show the variation in the data which provides information like how well the average of the sample represents the entire data. Less variation gives close representation while with larger variation the average may not closely represent all the values in the sample.

Measures of dispersion enable us to compare two or more series with regard to their variations. It helps to determine consistency.

With the checking for variation in the data, we can try to control the causes behind the variations.

USES OF A VARIABLE

In statistics and research, a variable is any property or feature you attempt to quantify, manage, or regulate. In every study, a variable is analysed. This variable may be a person, location, item, or idea. The value of a variable may vary among groups or over time.

Variables express traits or qualities that can differ amongst the people, things, or entities being investigated. They are essential for gathering, compiling, organising, and analysing data.

If a person's eye colour were the experiment's variable, it might range from brown to blue to green, depending on the individual.

Here are a few frequent statistical applications for variables:

- Variables make Data collection easier since they point out the precise traits that should be assessed or observed. For instance, the variable of interest in a poll regarding people's ages is "age".
- Data Organisation: Variables support data organisation by assigning names to various groups or properties. This makes it possible to organise data in a meaningful fashion, which facilitates management and analysis.
- Quantitative analysis is made possible by variables, which let researchers provide numerical values to many qualities. For instance, the variable "score" is given numerical values in research on exam results.
- Descriptive Statistics: Different descriptive statistics, such as mean, median, mode, range, and standard deviation, are calculated using variables. These statistics shed light on the distribution and central tendency of the data.
- Variables are crucial in inferential statistics, which uses samples to allow researchers to make generalisations about populations. Variables are used to generate hypotheses and estimates.
- Categorical Analysis: Categorical variables are qualitative characteristics that may be divided into several categories. They examine ratios, frequencies, and connections between various categories.
- Continuous variables are used for measurements with any value within a range, according to continuous analysis. They are essential for examining patterns, connections, and trends.
- Variables are utilised in correlation and regression studies to comprehend the relationships between variables and predict values based on these associations.
- Variables are changed in experimental research to see how they affect other variables. This enables scientists to establish cause-and-effect connections.
- Variables are inputs in predictive models, such as machine learning algorithms, impacting the model's capacity to predict outcomes accurately.
- Variables are employed in comparative analysis to contrast various circumstances or groups. This frequently occurs in research that includes control and experimental groups or several demographic groups.
- Time-dependent variables, such as stock prices, temperature fluctuations, or sales data, are used in time series analysis to examine patterns and trends.
- Multivariate Analysis: To better comprehend complicated links and interactions, multivariate analysis includes studying numerous variables simultaneously.
- Cluster Analysis: In cluster analysis, related data points are grouped together using variables to assist uncover hidden patterns in datasets.
- Anova (Analysis of Variance): Anova examines how a categorical variable varies across various circumstances or groups.
- Factor analysis is a method for figuring out the underlying causes of correlations between observable data.

DEPENDENT VARIABLE

What you wish to explain or forecast with the help of the model is known as the dependent variable (DV). This variable's values are dependent on other variables. The result is what you're looking at. It is also referred to as the left-hand variable, the response variable, and the result variable. Statisticians frequently represent them with a Y. Dependent variables are often plotted on the vertical, or Y, axis. For instance, a measure of plant growth is the dependent variable in the study of plant growth example. The experiment's result is that, and we want to know what factors contributed to it.

How to choose a dependent Variable?

Stability

Stability is frequently indicative of a more reliable dependent variable. The effects on the dependent variable should almost match those from the initial experiment if the experiment is repeated with the same subjects, surroundings, and experimental manipulations.

Complexity

A researcher may also select dependent variables based on the intricacy of their investigation. There can be more than one of each type of variable, even though some studies only contain one of each dependent and independent variable.

Additionally, researchers may be interested in discovering the effects of changing one independent variable on many dependent variables. Consider an experiment where the goal is to discover how a space's messiness influences people's creativity levels.

This study aims to examine how a person's mood may be affected by how messy space is. Two dependent variables—creativity level and mood—and an independent variable—the messiness of space—would make up the research.

Operationalisation Skill

The definition of operationalisation is "translating a construct into its manifestation." It simply refers to the method used to measure a variable. Therefore, a good dependent variable is one that you can quantify.

INDEPENDENT VARIABLE

The independent variables (IVs) you include in the model to explain or forecast changes in the dependent variable are known as IVs. You may understand their function in statistical analysis from the name. These elements are separate. Independent in this sense means that they are unaffected by other model variables and stand independently. The study's goal is not to determine why the independent variables vary.

Because they appear on the right side of the equals sign in a regression equation, independent variables are also called predictors, factors, treatment, explanatory, input, x-variables, and right-hand variables. Statisticians frequently use Xs to represent them in notation. Analysts plot independent variables down the horizontal axis of graphs.

Independent variables are referred to as features in machine learning.

For instance, the independent variables in a study on plant development can be soil moisture (continuous) and fertiliser type (categorical).

How to Include Independent Variables?

Controlled experiments: The values of the independent variables are carefully specified and controlled by the researchers. Relationships between independent and dependent variables are frequently causal in randomised trials. The independent variables bring about changes in the dependent variable.

In observational studies, the explanatory factors are not given values; the researchers watch the explanatory variables in their natural settings. The correlations between the independent and dependent variables may not be causal.

Simple regression is creating a regression model with only one independent variable. It is called multiple regression when there are numerous independent variables. Despite the title variations, the analysis is the same as the interpretations and presumptions.

Model specification is selecting which IVs to include in a statistical model. In-depth investigation and several subject-area, theoretical, and statistical issues are part of that process. Explicitly for observational studies, you'll want to include confounding factors that may skew your results if you don't include them as well as the predictors you are explicitly analysing in your study.

Some examples of Dependent and Independent Variables Variables
In a statistical study, the variable you're attempting to explain, predict, or comprehend is known as the dependent variable. In your opinion, it is the result or action that is affected by one or more independent factors. Here are some instances of dependent variables used in different situations:

1. **Economics**
   Analysing the effects of variations in advertising spending and product prices on sales income.
   Sales revenue is a dependent variable.
   Price of the product and advertising spending are independent variables.
2. **Medicine**
   Researching the effects of various pharmaceutical doses on patients' blood pressure levels.
   Blood pressure is a dependent variable.
   Independent Variables: A medication's dosage
3. **Education**
   Examining the effects of various teaching approaches and study time on students' test results.
   Dependent Test results are a variable.
   Independent Hours spent studying, teaching style
4. **Marketing**
   Analysis of the relationship between customer happiness and the responsiveness of customer service.
   Score of customer satisfaction, a dependent variable.
   Product quality and customer service responsiveness are independent variables.
5. **Finance**
   Analysing the effects of market sentiment and corporate profits on stock price.
   Stock price is a dependent variable.
   Market sentiment and company earnings are independent variables.

6. **Political Science**
   Examining how election campaigns and voting accessibility impact voter turnout rates.
   Voter turnout is a dependent variable.
   Election campaigns, voting accessibility, the independent variable
7. **Political Science**
   Investigating the connection between traffic volume, air quality, and industrial emissions.
   The air quality index is a dependent variable.
   Traffic density and industrial emissions are independent variables.

CONTINUOUS VARIABLE

A continuous variable is a numerical variable that may be measured to determine its value. The variables can be broken into meaningful smaller increments, including fractional and decimal values, and can accept nearly any form of numeric value. This specific type of quantitative variable is frequently used in statistical modelling and machine learning to represent data that may be measured in some way. Scales are commonly used to measure continuous variables, including height, weight, temperature, etc. Continuous variables can be used to calculate the mean, median, variance, or standard deviation. Continuous variables have endless possible values between any two numeric values. Various calculus methods are employed in continuous optimisation situations when the variables are Continuous.

Example
How many eggs does a hen lay? A chicken might or might not lay egg(s) every day, but two things are impossible. A part or fraction of an egg is also impossible, nor can there ever be an egg in a negative quantity.

Now that you are aware of how the two variables differ from one another. Therefore, there must be some significant distinctions between the two that allow for better data representation.

Temperature is an example of a continuous variable since it can be measured with decimals and can have any value within an interval. Up until a quantum level, practically all of the variables in nature are continuous.

Continuous Variable Types

There are two categories of continuous variables:

Instant variable: Instant variables are those variables that determine the static level or distance between each equal category.

Ratio variables are those variables that only differ in one way from an interval variable. The ratio of the scores reveals the link between the replies.

CATEGORICAL VARIABLE

Statistical information comprised of categorical variables—data divided into categories—is known as categorical data. A set of grouped data is one of the examples. More specifically, countable qualitative or quantitative data clustered within predetermined intervals might be used to create category data. The information is condensed into a probability table. However, "categorical data" refers to data sets when discussing data analysis. It should be noted that although the data set includes some category variables, it may also include non-categorical variables.

It's critical to understand the various data types when studying statistics. It's because statistical approaches can be carried out only with the aid of certain data types. Understanding various data types enables you to analyse the best technique. The actual pieces of information that are gathered via the investigation are called data. Most of the data are found to fit into one of two categories:

Quantitative data or numerical data

Qualitative or Categorical Data

Let's now examine categorical data in statistics in more depth.

Qualitative or Categorical Data

The categorical data is made up of categorical variables, which stand in for traits like a person's gender or birthplace. Natural language descriptions, not numerical values, are used to represent categorical measures. Categorical data can occasionally have numerical values, but such values are not mathematically meaningful. The following are some instances of categorical data:

Date of birth Favourite sport at school Postal code

how you go to school, etc.

The birthday and postcode in the aforementioned case both include digits. Although it includes numbers, it is still regarded as categorical data. Calculating the average is a simple approach to identify if the provided data is categorised or numerical. If you can figure out the average, it qualifies as numerical data. It is regarded as categorical data if you cannot determine the average. The average of the birthday and the postal code, like in the aforementioned example, has no significance and is regarded as categorical data.

Categorical Data Types

Typically, categorical data consists of values and observations that may be categorised or grouped. Pie charts and bar graphs are the ideal visual representations for these data. Additionally, categorical information is divided into two categories:

- Nominal Data
- Ordinal Data

Nominal Data

Without providing a numerical value, nominal data is a sort of data that is used to name the variables. The nominal scale is another name for it. Nominal data are not measurable or able to be arranged. However, nominal data can occasionally be both qualitative and quantitative. The few instances of nominal data that are often used are letters, words, symbols, gender, etc.

Ordinal Data

Data that has a natural order is referred to as ordinal data. The distinguishing characteristics of ordinal data include the inability to distinguish between data values. In surveys, questionnaires, finance, and economics, it is frequently used.

Tools for visual analysis can be used to analyse the data. Bar charts are frequently used to illustrate it. Tables may be used to represent data in some cases, with each row designating a different category.

Categorical Variables

A categorical variable in statistics has a finite, typically set number of potential values. They accept values,

often names or labels.

Examples include

- a wall's colour, such as red, blue, pink, green, etc.
- people's gender, including male, female, and transgender
- Blood type of an individual: A, B, O, AB, etc.

Based on some qualitative quality, these variables are used to categorise each human or another unit of observation into a particular group or nominal category. Typically, each of a categorical variable's possible values is referred to as a level. Categorical distribution refers to the probability distribution associated with a random categorical variable.

DISTRIBUTION TYPES AND SKEWNESS

Normal Distribution:

Normal Distribution is one of the most common continuous probability distributions. This type of distribution is important in statistics and is often used to represent random variables whose distribution is not known.
This type of distribution is symmetric, and its mean, median, and mode are equal.
Mathematically, Gaussian Distribution is represented as:

$$N\sim(\mu, \sigma 2\ )$$

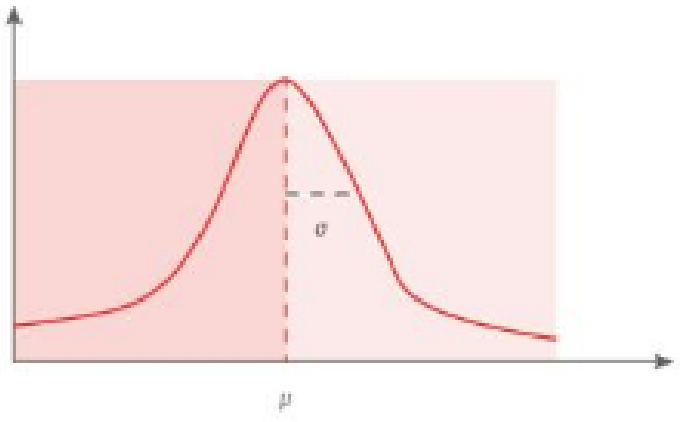
Where N stands for Normal, symbol ~ for distribution, whereas symbol μ stands for mean and σ2 stands for the variance.

Normal distribution, also known as the Gaussian distribution, is a continuous probability distribution that is symmetric about the mean, showing that data near the mean are more frequent in occurrence than data far from the mean. In graph form, normal distribution will appear as a bell curve (as in the figure below).

A random variable X is said to have a normal distribution with parameters mean μ and variance $\sigma 2$. if it's p.d.f is given by the probability law:

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

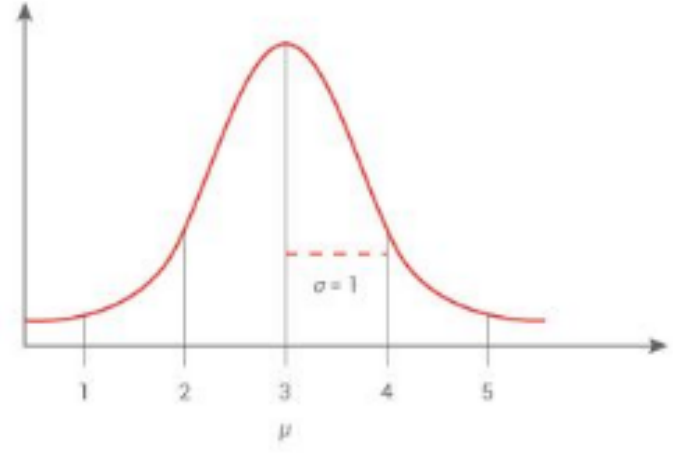
here π and e are mathematical constants 3.141 and 2.718 respectively.



In the above image, we can view the highest point is located at the mean μ, and the spread of the graph can be observed by the standard deviation 'σ'.

Let us understand this with the easiest example where we can have the random variable X with distribution:
X = {1, 2, 3, 4, 5}
When we take the mean and the standard deviation of the above data set, we get mean(μ) = 3 and standard deviation(σ) = 1.

When we plot it, we get a few distributions like this mentioned below:



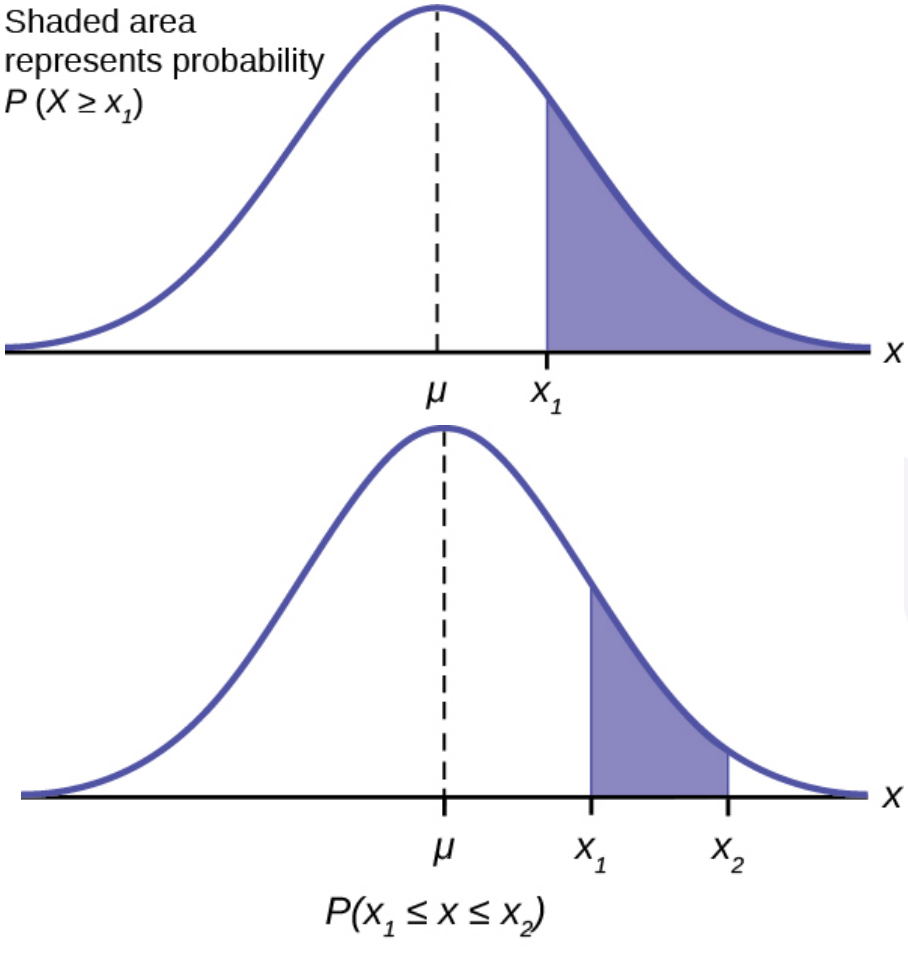
This Bell curve specifies the Gaussian distribution.

Properties of Normal Distribution:

- Mean=Median=Mode
- Symmetry about the center
- 50% of values less than the mean and 50% greater than the mean.
- Linear combination of independent normal variates is also a normal variate.
- X-axis is an asymptote to the curve.
- The point of inflection of the curve are:

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{1}{2}\right)$$

- The probability that X is greater than x1 is equal to the area under the normal curve as shown by the shaded area in the figure below.



Why Normal Distribution is essential:

1. Distribution of sample means with a large sample size can be approximated to normal distribution.
2. Decisions based on everyday distribution insights have proven to be of good value.
3. All computable statistics are elegant
4. It approximates a wide variety of random variables.

This frequency table will help us make better sense of the data given.
Skewed Distributions:

What is Skewness:
Skewness is asymmetry in a statistical distribution, in which the curve appears distorted or skewed to the left or right. Skewness can be quantified to define the extent to which a distribution differs from a normal distribution.

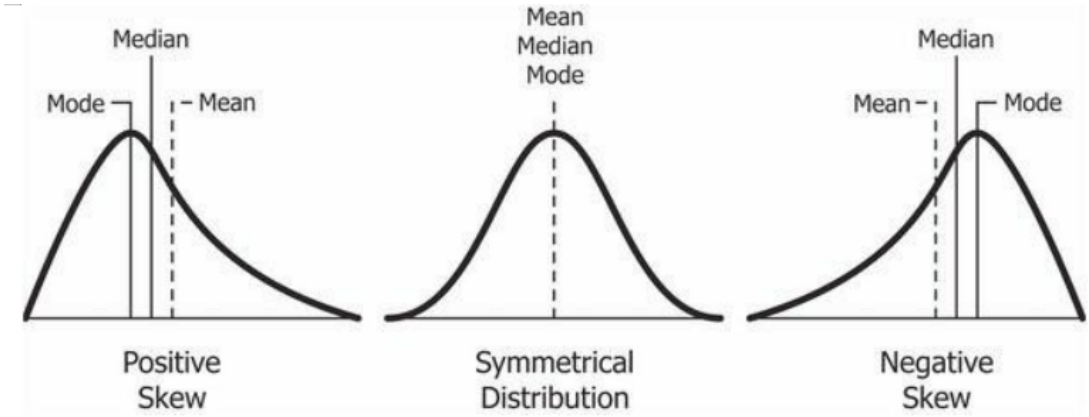
In a normal distribution, the graph appears as a classical, symmetrical "bell-shaped curve." The mean, or

average, and the mode, or maximum point on the curve, are equal.

In a perfect normal distribution, the tails on either side of the curve are exact mirror images of each other.

When a distribution is skewed to the left, the tail on the curve's left-hand side is longer than the tail on the right-hand side, and the mean is less than the mode. This situation is also called negative skewness.

When a distribution is skewed to the right, the tail on the curve's right-hand side is longer than the tail on the left-hand side, and the mean is greater than the mode. This situation is also called positive skewness.



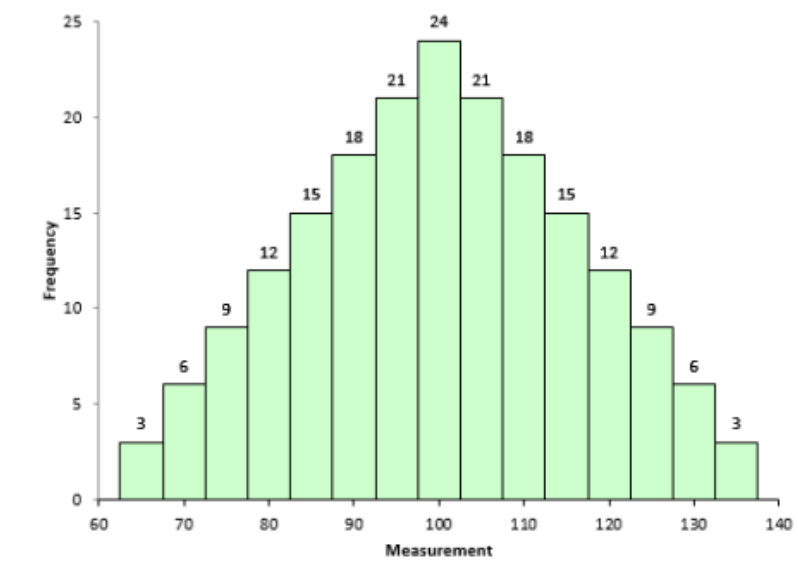
The skewness is defined as:

$$a_3 = \sum \frac{(X_i - \bar{X})^3}{ns^3}$$

Where Xi is the ith X value, n is the sample size, $\bar{x}$ is the average, and s is the sample standard deviation.

$$Skewness = \frac{n}{(n-1)(n-2)} \sum \frac{(X_i - \bar{X})^3}{s^3} = \frac{n}{s^3(n-1)(n-2)} (S_{above} - S_{below})$$

This sample size formula is used here:



The figure above is for a symmetrical data set. This data set was created by generating the data from 65 to 135 in 5 steps with the number of each value, as shown in Figure above.

The above figure shows Symmetrical Data set with Skewness equals to 0
For example, there are three 65's, six 70's, and nine 75's, etc.

The Set of symmetrical data has a skewness equal to 0, Where each X value is subtracted from the average. So if a collection of data is symmetrical for each point that is a distance (d) above the average, there will be a point that is a distance (-d) below the average.

Consider 65 values and 135 values. The average of the data in the above figure is 100.

When X=65 $$\frac{(X_i - \bar{X})^3}{s^3} = \frac{(65-100)^3}{s^3} = \frac{(-35)^3}{s^3} = \frac{-4278}{s^3}$$

For X=135 then: $$\frac{(X_i - \bar{X})^3}{s^3} = \frac{(135-100)^3}{s^3} = \frac{(35)^3}{s^3} = \frac{4278}{s^3}$$

So, the -4278 value and the value of+4278 even out at 0. So, a Symmetrical data set will have 0 skewness.

To explore +ve & -ve values of skewness, let's define the following terms:

$S_{above} = |\sum(X_i - \bar{X})^3|$ if X_i is above the average

$S_{below} = |\sum(X_i - \bar{X})^3|$ if X_i is below the average

So, when Xi is above the average, Sabove is the "size" of the deviations from average. Likewise, when Xi is below the average, Sbelow can be viewed as the "size" of the deviations from average.

Then the skewness becomes:

$$Skewness = \frac{n}{(n-1)(n-2)}\sum\frac{(X_i - \bar{X})^3}{s^3} = \frac{n}{s^3(n-1)(n-2)}(S_{above} - S_{below})$$

The skewness will be positive if Sabove is larger than Sbelow . This means that the right-hand tail will be longer than the left-hand tail. Figure 2 is an example of this. Skewness for this dataset is 0.514. Positive skewness indicates that the size of the right-handed tail is larger than the left-handed tail.

Fig 2: A dataset with Positive Skewness

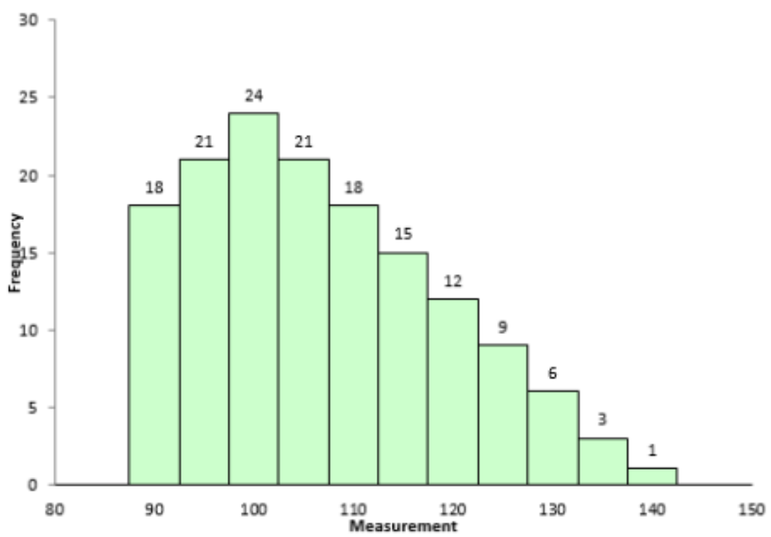
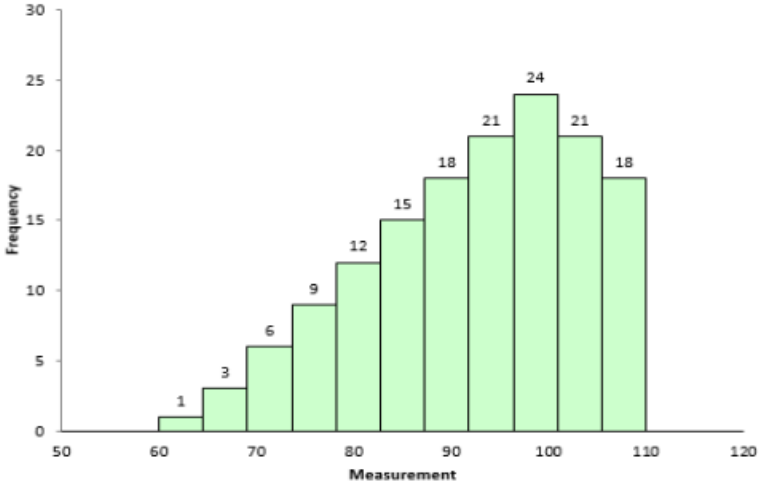


Figure 3 is an example of the datasets with negative skewness. it is the mirror image, necessarily of Figure 2. then the skewness is -0.514. In this case, Sabove .is smaller than Sbelow. The left-hand tail will typically be longer than the right hand tail.

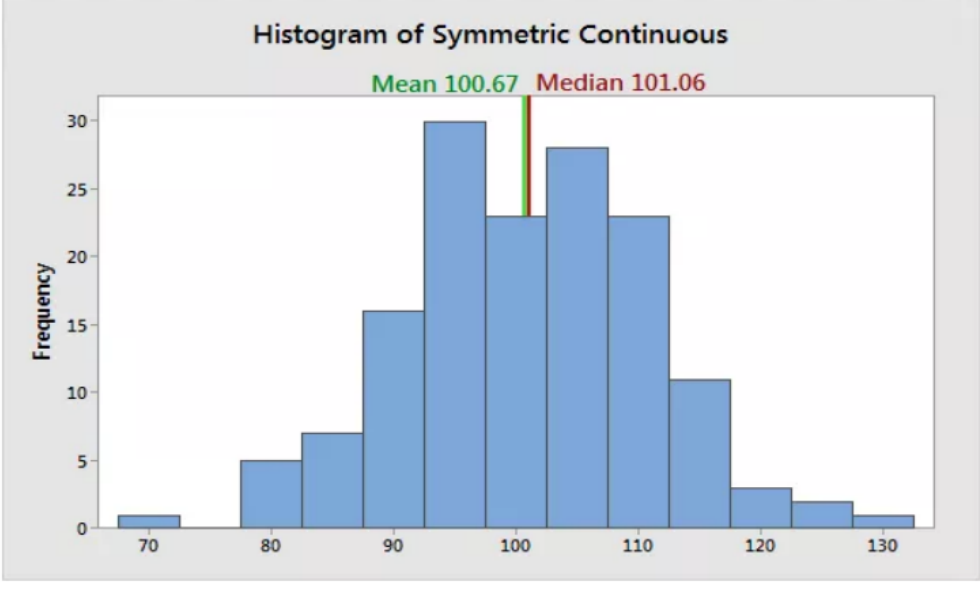
Fig 2: Negative Skewness with Dataset



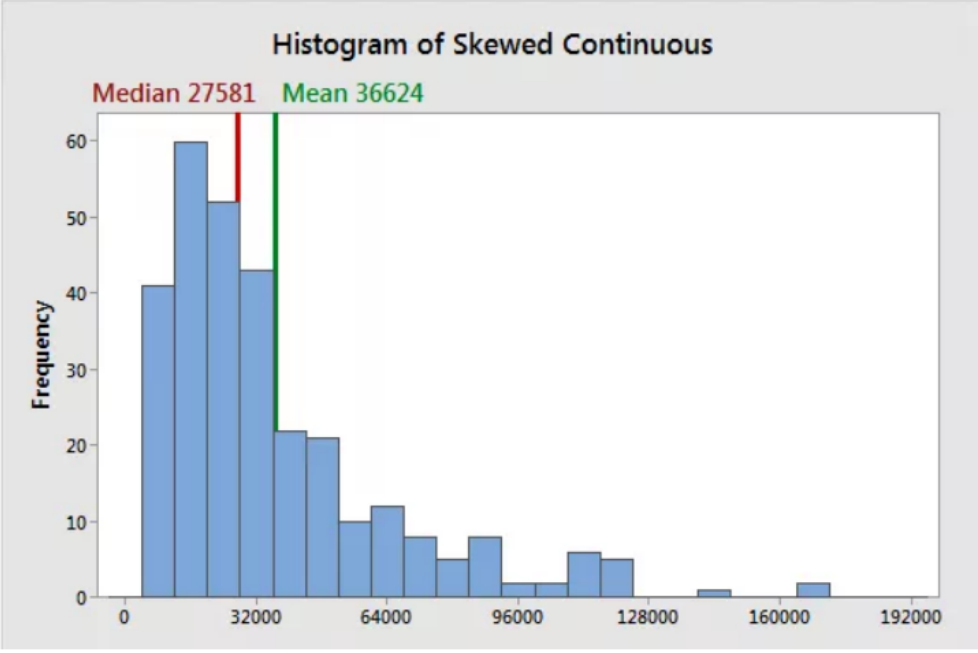
So When is the skewness too much?

If the skewness is between -0.5 and 0.5, the data are fairly symmetrical. if the skewness is between -1 and -0.5 or between 0.5 and 1, the data is moderately skewed.

If the skewness is greater than for 1 or less -1 the data is highly skewed.
In a symmetric distribution, the mean and median both find the center accurately. They are approximately equal.



However, in a skewed distribution, the mean can miss the mark. In the histogram below, it is starting to fall outside the central area. This problem occurs because outliers have a substantial impact on the mean. Extreme values in an extended tail pull the mean away from the center. As the distribution becomes more skewed, the mean is drawn further away from the center.
Here the median better represents the central tendency for the distribution.



Uses of Mean, Median, Mode:

When you have a symmetrical distribution for continuous data, the mean, median, and mode are equal. In this case, use the mean because it includes all of the data in the calculations. However, if you have a skewed distribution, the median is often the best measure of central tendency.
When you have ordinal data, the median or mode is usually the best choice. For categorical data, you have to use the mode.

Frequency Distribution:

Frequency distribution in statistics provides the information of the number of occurrences(frequency) of distribution within a given period of time or interval, in list , or graphical representation. Grouped and ungrouped are two types of frequency distribution.

Many times it is not easy or feasible to find the frequency of the data from a very large dataset. So to make sense of the data we make a frequency table and graphs. Let us take the example of the heights of the students in cms.

Example:

138,145,168,125,168,139,151,125,168,139

Height(cms)	Frequency
125	2
138	1
139	2
145	1

Height(cms)	Frequency
151	1
168	3

This frequency table will help us make better sense of the data given.

Standard Error

A statistic's standard error (SE) is the approximate standard deviation of a statistical sample population. The standard error is a statistical term that measures the accuracy with which a sample distribution represents a population by using standard deviation. In statistics, a sample mean deviates from the actual mean of a population—this deviation is the standard error of the mean.

The mean, or average, is generally calculated when a population is sampled. The standard error can include the variation between the population's calculated mean and one considered known, or accepted as accurate. This helps compensate for any incidental inaccuracies related to sample gathering.

In cases where multiple samples are collected, the mean of each sample may vary slightly from the others, creating a spread among the variables. This spread is often measured as the standard error, accounting for the differences between the means across the datasets.

The more data points involved in the mean calculations, the smaller the standard error tends to be. When the standard error is small, the data is said to be more representative of the true mean. In cases where the standard error is large, the data may have some notable irregularities.
The standard deviation is a representation of the spread of each of the data points. The standard deviation is used to help determine the validity of the data based on the number of data points displayed at each standard deviation level. Standard errors function more as a way to determine the accuracy of the sample or the accuracy of multiple samples by analyzing deviation within the means.
The following formula gives Standard Error:

Standard Error

$$Standard\ Error\ (\sigma_\mu) = \frac{\sigma}{\sqrt{n}}$$

σ = standard deviation

n = quantity of numbers in the group

Here σ is the population's standard deviation, whereas σ(u) is the standard deviation of the sample.
We can see that as the size of our sample increases, the Standard error decreases.
12.2 Standard Error of Mean:

The standard deviation of the mean, often known as the standard error of the mean, is a tool for determining discrepancies between multiple data samples. The calculation takes into consideration any data variations that may exist. If you take the weight of several men, for example, the measures can vary significantly amongst them; some may weigh 150 pounds, while others may weigh 300 pounds. On the other hand, the mean of these samples will differ by only a few pounds. The standard error of the mean shows how far the various weights diverge from the mean.

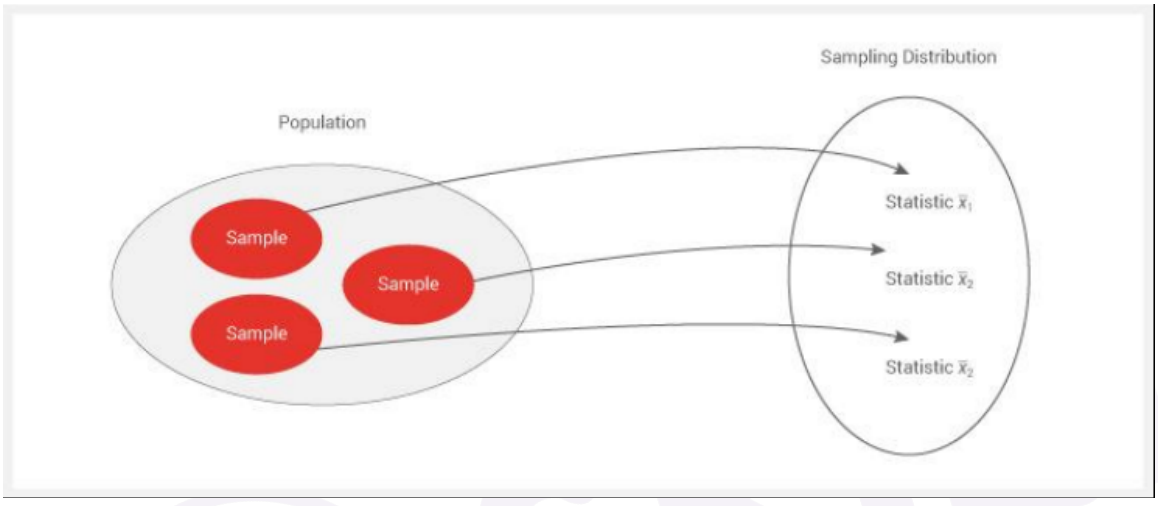
$$SE = \frac{\sigma}{\sqrt{n}}$$

← Standard deviation

← Number of samples

Central Limit Theorem:

The central limit theorem states that the sample mean follows approximately the normal distribution with mean(μ) and standard deviation (σ/√n), where μ and σ are the mean and standard deviation of the population from where the sample was selected. The sample size n has to be large (usually n≥30) if the population from where the sample is taken is non normal.



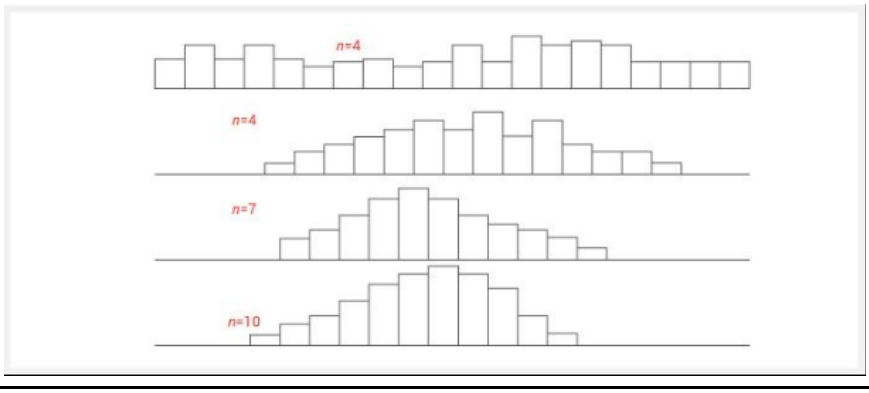
After fetching different samples which are enough in numbers, we can calculate each sample's mean and plot the various distributions.

Also, if we take the sample mean's average, the result will be equal to the actual population mean & the standard deviation equal σ/√n.
Where,

'σ' is the population of std deviation
n = the sample size

Important point while applying the Central Limit Theorem.

- The distribution of the original population datasets does not matter. It could be normal, uniform, binomial, etc.
- The distribution of the sample means would be a normal distribution.
- Larger the number of samples taken from the population the closer to a Normal Distribution the sample mean will be .



- The samples extracted should be more significant than 30 observations.
- The sample mean average extracted will be approximately equal to the mean of the population and its variance would be similar to the original variance, which is divided by the sample size i.e 'n'.

Mathematical Explanation of Central Limit Theorem:

The central limit theorem states that the mean (X) follows approximately the Normal distribution with mean μ and standard deviation /n, where μ and σ are the mean and standard deviation of the population.

$$\text{To summarize: } \bar{X} \sim N(\mu, \tfrac{\sigma}{\sqrt{n}})$$

$$\text{To transform } \bar{X} \text{ into } z \text{ we use: } z = \tfrac{\bar{x}-\mu}{\frac{\sigma}{\sqrt{n}}}$$

Example: Let X be a random variable along with =10 and =4. A sample of size 100 find the probability that the sample mean '' of 100 the number of observations is not more than (less than) 9.

Write: $P(\bar{X} < 9) = P(z < \frac{9-10}{\frac{4}{\sqrt{100}}}) = P(z < -2.5) = 0.0062$

Similarly the central limit theorem states that the sum T follows approximately a normal distribution,

$$T \sim N(n\mu, \sqrt{n}\sigma)$$

Where μ and σ are the mean and standard deviation of a population. To transform T into z, we use:

$$z = \frac{T-n\mu}{\sqrt{n}\sigma}$$

Example :
Let X be a random variable along with 'μ' = 10 and σ = 4 . A sample of size 100 is taken from the population. Find the probability that the sum of these 100 numbers of observations is less than 900.
Solution :
We write
P(T < 900)
= P(Z < 900-100(10)100(4))
= P(z < - 2.5)
= 0.0062 (from the table of standard normal probabilities)

Applications of Central Limit Theorem:

- This helps in the analysis of data in approaches such as the construction of confidence intervals.
- To more correctly estimate the population mean, we can increase the sample size obtained from the population, which will reduce the sample mean deviation.
- Central Limit Theorem is an approximation you can use when the population you're studying is so big, it would take a long time to gather data about each individual that's part of it.

Example :
A large freight elevator can transport a maximum of 9800 pounds. Suppose a load of cargo containing 49 boxes must be carried through the elevator. Experience has shown that the weight of some boxes of this type of cargo follows a distribution with mean μ = 205 pounds & the standard deviation 'σ' = 15 pounds. Based on this

information, what is the probability that all 49 boxes can be safely loaded onto the freight elevator and transported?

Solution :
We are given n = 49, μ = 205, σ = 15.
The elevator can transport up to 9,800 Pounds. Therefore, these 49 boxes will be safely transported if they weigh in total, not more than(less than) 9800 pounds.
The probability that the total weight of 49 boxes is not more than(less than) 9800 pounds is
P(T < 9800)
= P(Z < 9800 - 49(205)4915)
= P (z < -2.33)
= 1 - 0.9901
= 0.0099.

Example : It is known that the number of tickets purchased by a student standing in line at the ticket counter to buy the tickets for the Football match of U.C.L.A. against USC follows a distribution that has a mean μ = 2.4 & the standard deviation 'σ' = 2.0.
Suppose 100 eager students are standing in line to purchase the tickets. If only 250 tickets remain unbooked, what is a Probability that all 100 students will be able to buy the tickets they desire?

Solution :
We are given that μ = 2.4, σ = 2, n = 100.
There are 250 tickets available, so a hundred students will be able to purchase the tickets they want if all together ask for not more than 250 tickets. Probability for that is
P(T < 250)
= P(Z< 250 - 100(2.4)1002)
= P(z < 0.5)
= 0.6915.

Question : Suppose that you have a sample of 100 values from a population with mean μ = 500 and with standard deviation σ = 80.
a. What is the probability that the μ will be in the interval (490, 510)?
b. Give the interval that covers the average 95 percent of the distribution of the sample mean.
Solution :
We are given μ = 500, σ = 80, n = 100.

a. $P(490 < \bar{x} < 510)$

$= P(\frac{490-500}{\frac{80}{\sqrt{100}}} < z < \frac{490-500}{\frac{80}{\sqrt{100}}})$

= P(-1.25 < z < 1.25)
= 0.8944 - (1 - 0.8944)
= 0.7888.

b. $\pm 1.96 = \frac{\bar{x}-500}{\frac{80}{\sqrt{100}}}$

$\bar{x} = 484.32, \bar{x} = 515.68$

Therefore $P(484.32 < \bar{x} < 515.68)$

= 0.95

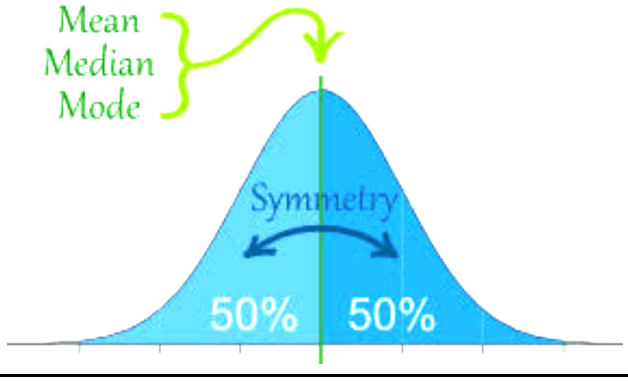
Question: The amount of regular unleaded gasoline purchased every week at a gas station near UCLA follows the normal distribution with mean 50,000 gallons and a standard deviation of 10,000 gallons. The starting supply of gasoline is 74,000 Gallons, and there is a scheduled weekly delivery of 47000 gallons.
a. Find the probability that, after 11 weeks, a supply of gasoline will be below 20000 gallons.
b. How much should the weekly delivery be so that after 11 weeks, the probability that the supply is only 0.5% below 20000 gallons is?

Solution :
Given: 'μ' = 50000, 'σ' = 10000, n = 11. The starting supply is 74000 gallons. 47000 gallons is the weekly delivery.
Therefore, the total supply for the eleven weeks is 74000 + 11 × 47000 = 591000 Gallons.
a. The supply will be below 20000 gallons if the gasoline purchased in these 11 weeks is more than 591000 - 20000 = 571000 Gallons
Therefore, we need to find
P(T > 571000)
P(z > 57100-11(50000)1110000)
= 1 - 0.7357
= 0.2643.

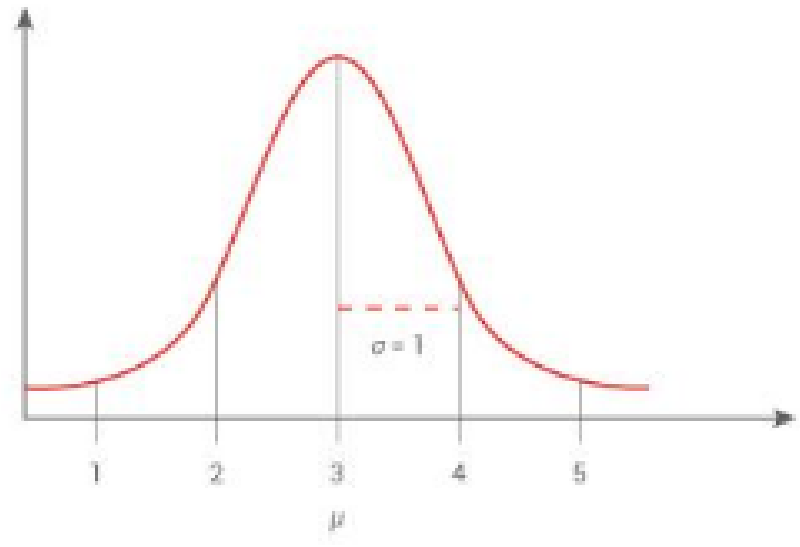
b. Let A be the unknown scheduled delivery.
The total gasoline purchased must be above 74000 + 11 × A - 20000.
We need this with a probability 0.5% or Probability P(T > 74000 + 11A - 20000) = 0.005.
The z value is 2.57566.
So, 2.575 = 74000 + 11 A -20000- 11(50000)1110000
A = 52854.88
The weekly delivery must be 52854.8 gallons.

Normal Distribution(Gaussian Distribution) :

Normal distribution, also known as the Gaussian distribution, is a continuous probability distribution that is symmetric about the mean, showing that data near the mean are more frequent in occurrence than data far from the mean.
Normal Distribution is one of the most common continuous probability distributions. This type of distribution is important in statistics and is often used to represent random variables whose distribution is not known.
This type of distribution is symmetric, and its mean, median, and mode are equal.
Mathematically, Gaussian Distribution is represented as:
N~(, 2)
Where N stands for Normal, symbol ~ for distribution, whereas symbol μ stands for mean and σ2 stands for the variance.



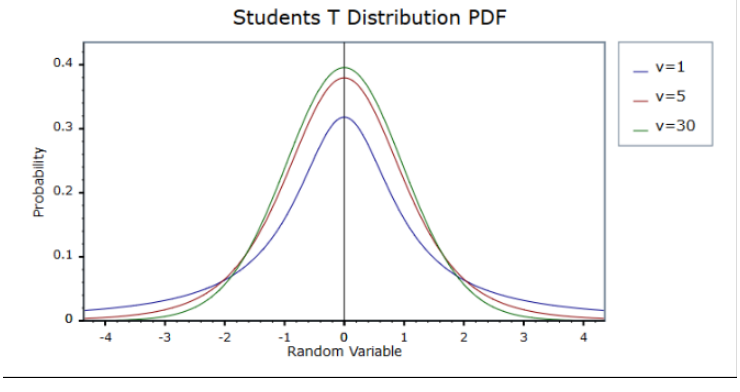
In the above image, we can view the highest point is located at the mean μ, and the spread of the graph can be observed by the standard deviation 'σ'.
Let us understand this with the easiest example where we can have the random variable X with distribution:
X = {1, 2, 3, 4, 5}
When we take the mean and the standard deviation of the above data set, we get mean(μ) = 3 and standard deviation(σ) = 1.
When we plot it, we get a few distributions like this mentioned below:



This Bell curve specifies the Gaussian distribution.

T- Distribution:

The t distribution (Student's t-distribution) is a probability distribution that is used to estimate population parameters when the sample size is small and/or when the population variance is unknown.
The t distribution is very similar to the normal distribution when the estimate of variance is based on many degrees of freedom, but has relatively more scores in its tails when there are fewer degrees of freedom.
The t distribution approaches the normal distribution as the degrees of freedom increase.



According to the central limit theorem, the sampling distribution of a statistic (like a sample mean) will follow a normal distribution as long as the sample size is sufficiently large. Therefore, when we know the population's standard deviation, we can compute a z-score and use the normal distribution to evaluate probabilities with the sample mean.

But sample sizes are sometimes small, and we often do not know the population's standard deviation. When either of these problems occurs, statisticians rely on the distribution of the t statistic (also known as the t score),

whose values are given by:

$t = (x- \mu) / (s/\sqrt{n})$
Where x is the sample mean,
μ is the population mean,
s is the standard deviation of the sample,
and n is the sample size.

The distribution of the t statistic is called the t distribution or the Student t distribution.

The t distribution allows us to conduct statistical analyses on specific data sets that are not appropriate for analysis using the normal distribution.

Degrees of Freedom :

There are many different t distributions. The particular form of t-distribution is determined by its degrees of freedom. The degrees of freedom refers to the number of independent observations in a data set.
When estimating a mean score or a proportion from a single sample, the number of independent observations equals the sample size minus one. Hence, the t statistic from samples of size 8 would be described by a t distribution having 8 - 1 or 7 degrees of freedom. Similarly, a t distribution having 15 degrees of freedom would be used with a sample of size 16.

T- Score Table:

Table C: T-distribution Chart

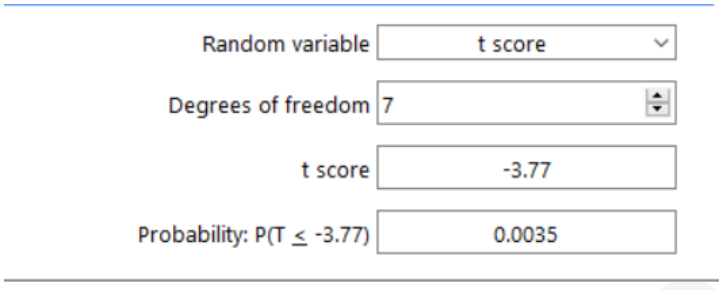
***t* Table**

cum. prob	$t_{.50}$	$t_{.75}$	$t_{.80}$	$t_{.85}$	$t_{.90}$	$t_{.95}$	$t_{.975}$	$t_{.99}$	$t_{.995}$	$t_{.999}$	$t_{.9995}$
one-tail	0.50	0.25	0.20	0.15	0.10	0.05	0.025	0.01	0.005	0.001	0.0005
two-tails	1.00	0.50	0.40	0.30	0.20	0.10	0.05	0.02	0.01	0.002	0.001
df											
1	0.000	1.000	1.376	1.963	3.078	6.314	12.71	31.82	63.66	318.31	636.62
2	0.000	0.816	1.061	1.386	1.886	2.920	4.303	6.965	9.925	22.327	31.599
3	0.000	0.765	0.978	1.250	1.638	2.353	3.182	4.541	5.841	10.215	12.924
4	0.000	0.741	0.941	1.190	1.533	2.132	2.776	3.747	4.604	7.173	8.610
5	0.000	0.727	0.920	1.156	1.476	2.015	2.571	3.365	4.032	5.893	6.869
6	0.000	0.718	0.906	1.134	1.440	1.943	2.447	3.143	3.707	5.208	5.959
7	0.000	0.711	0.896	1.119	1.415	1.895	2.365	2.998	3.499	4.785	5.408
8	0.000	0.706	0.889	1.108	1.397	1.860	2.306	2.896	3.355	4.501	5.041
9	0.000	0.703	0.883	1.100	1.383	1.833	2.262	2.821	3.250	4.297	4.781
10	0.000	0.700	0.879	1.093	1.372	1.812	2.228	2.764	3.169	4.144	4.587
11	0.000	0.697	0.876	1.088	1.363	1.796	2.201	2.718	3.106	4.025	4.437
12	0.000	0.695	0.873	1.083	1.356	1.782	2.179	2.681	3.055	3.930	4.318
13	0.000	0.694	0.870	1.079	1.350	1.771	2.160	2.650	3.012	3.852	4.221
14	0.000	0.692	0.868	1.076	1.345	1.761	2.145	2.624	2.977	3.787	4.140
15	0.000	0.691	0.866	1.074	1.341	1.753	2.131	2.602	2.947	3.733	4.073
16	0.000	0.690	0.865	1.071	1.337	1.746	2.120	2.583	2.921	3.686	4.015
17	0.000	0.689	0.863	1.069	1.333	1.740	2.110	2.567	2.898	3.646	3.965
18	0.000	0.688	0.862	1.067	1.330	1.734	2.101	2.552	2.878	3.610	3.922
19	0.000	0.688	0.861	1.066	1.328	1.729	2.093	2.539	2.861	3.579	3.883
20	0.000	0.687	0.860	1.064	1.325	1.725	2.086	2.528	2.845	3.552	3.850
21	0.000	0.686	0.859	1.063	1.323	1.721	2.080	2.518	2.831	3.527	3.819
22	0.000	0.686	0.858	1.061	1.321	1.717	2.074	2.508	2.819	3.505	3.792
23	0.000	0.685	0.858	1.060	1.319	1.714	2.069	2.500	2.807	3.485	3.768
24	0.000	0.685	0.857	1.059	1.318	1.711	2.064	2.492	2.797	3.467	3.745
25	0.000	0.684	0.856	1.058	1.316	1.708	2.060	2.485	2.787	3.450	3.725
26	0.000	0.684	0.856	1.058	1.315	1.706	2.056	2.479	2.779	3.435	3.707
27	0.000	0.684	0.855	1.057	1.314	1.703	2.052	2.473	2.771	3.421	3.690
28	0.000	0.683	0.855	1.056	1.313	1.701	2.048	2.467	2.763	3.408	3.674
29	0.000	0.683	0.854	1.055	1.311	1.699	2.045	2.462	2.756	3.396	3.659
30	0.000	0.683	0.854	1.055	1.310	1.697	2.042	2.457	2.750	3.385	3.646
40	0.000	0.681	0.851	1.050	1.303	1.684	2.021	2.423	2.704	3.307	3.551
60	0.000	0.679	0.848	1.045	1.296	1.671	2.000	2.390	2.660	3.232	3.460
80	0.000	0.678	0.846	1.043	1.292	1.664	1.990	2.374	2.639	3.195	3.416
100	0.000	0.677	0.845	1.042	1.290	1.660	1.984	2.364	2.626	3.174	3.390
1000	0.000	0.675	0.842	1.037	1.282	1.646	1.962	2.330	2.581	3.098	3.300
z	0.000	0.674	0.842	1.036	1.282	1.645	1.960	2.326	2.576	3.090	3.291
	0%	50%	60%	70%	80%	90%	95%	98%	99%	99.8%	99.9%
	Confidence Level										

Example :

A lock Company claims that their locks have an average breaking strength of 1000 Kg, with a standard deviation of 150 kg. Suppose a customer tests 8 randomly-selected locks. What is the probability that the average breaking strength in the test will be no more than 800 Kg?

Solution :
Population mean = 1000
Sample mean = 85
n=8
t-score = (800 - 1000) / $(150/(8)^{0.5})$ = -3.77
we will use t-distribution calculator for this exercise:



We get a cumulative probability for P(X<800) = 0.0035
Note: This was a very simple case of t-distribution. In real life, t-test is one of the most important and most used tests in Hypothesis testing. There are many terms associated such as critical value, significance, 1 tail test and 2 tailed test. We will study more in detail about t-tests and use of t-score in the upcoming topic "Hypothesis Testing".

Comparing z-score and t-values:

Z score is the standardisation from the population of raw data or more than 30 sample data to a standard score, while T score is the standardisation from the sample data of less than 30 data to a standard score.
Z-score from -3 to +3, While the T score ranges from 20 to 80.
As the data size increases, the distribution tends to be Z distribution. Z scores and T score distribution is part of the standard distribution, but based on the size, they differ.

The Z score is used in stock market data to check the chances of the company going into bankruptcy. In contrast, the t-score is extensively used in checking bone mineral density.

The use of standard error in inferential statistics :

The standard error is a type of inferential statistic that is used to make inferences. Within a dataset, it reflects the standard deviation of the mean. This provides a measurement for the spread and acts as a measure of variance for random variables. The more precise the dataset, the smaller the dispersion.

The mean, or average, is usually estimated when a population is sampled. The variation between the computed population mean and one that is considered known or acknowledged as accurate might be included in the standard error. This helps to adjust for any inaccuracies that may have occurred during the sample collection process.
When numerous samples are gathered, each sample's mean may differ slightly, resulting in a spread across the variables. The standard error, which accounts for the discrepancies in the means across datasets, is the

most used way to measure this dispersion.

The more data points involved in the mean calculations, the smaller the standard error tends to be. When the standard error is small, the data is said to be more representative of the true mean. In cases where the standard error is large, the data may have some notable irregularities.

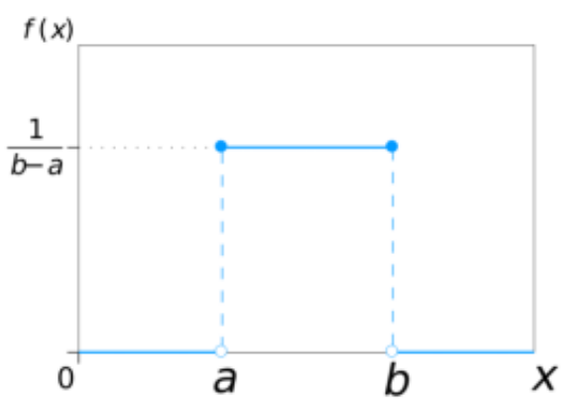
The standard deviation measures how far apart the data points are. Based on the number of data points displayed at each standard deviation level, the standard deviation is used to assist establish the validity of the data. Standard errors can be used to estimate the accuracy of a single sample or numerous samples by evaluating deviation within the means.

Uniform Distribution:

The probability distribution function of the continuous uniform distribution:

$$f(x) = \begin{cases} \frac{1}{b-a} & \text{for } a \leq x \leq b, \\ 0 & \text{for } x < a \text{ or } x > b \end{cases}$$

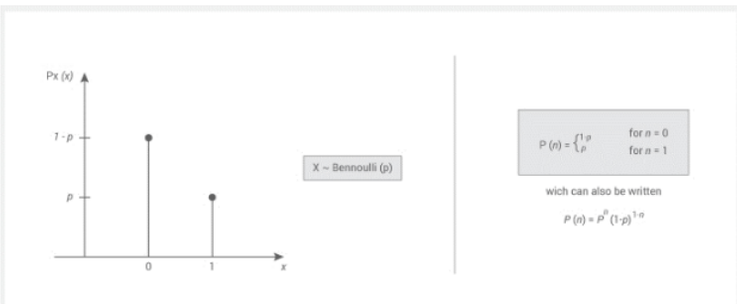
Since the area under the curve must be equal to 1, and the length of the interval determines the height of the curve, the following figure shows a uniform distribution (a,b).



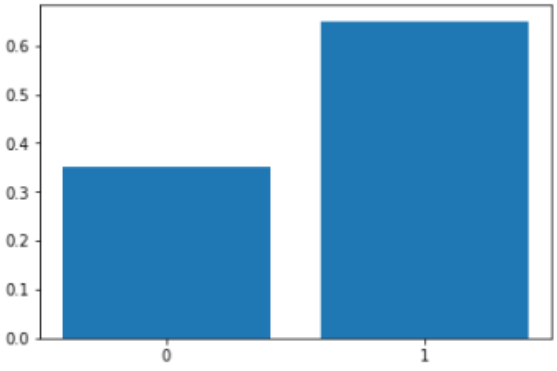
If we roll a die(numbered 1 to 8), then the probability of getting 1 is one out of 8.
Similarly, the probability of getting 2 to 6 is 1/6 . There is an equal chance to get each of 8 results(outcomes).

Bernoulli Distribution:

Bernoulli distribution is a discrete probability distribution of a random variable that has only two outcomes. namely 1 (success) and 0 (failure). Where n=1 occurs with probability p and n=0 (usually called a “failure”) occurs with probability q=1-p .
Where 0<p<1.
Therefore the probability density function(PDF) & the graph for Bernoulli’s distribution is shown in the figure below:



In the above diagram, 1 refers to 'success' & 0 refers to failure. The head and tail distribution in tossing a coin is an example of Bernoulli's with p=q=½ .
For example, probability (p) of scoring a goal in the last 10 minutes is 0.35 (success); the probability of not scoring a goal in the previous 10 minutes (failure) is 1 - p = 0.65.



Poisson Distribution:

Poisson Distribution can be used to find the probability of several events in a time period.

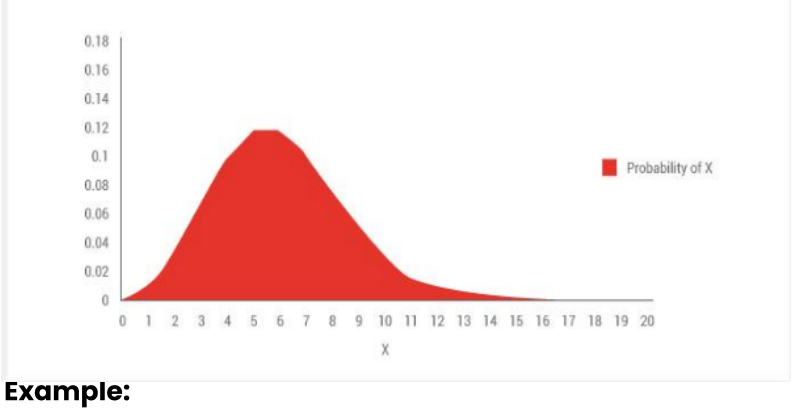
Condition for Poisson Distribution:
Here the events can occur independently.
An event can occur any number of times.
The rate of occurrence is constant i.e the rate does not change based on time.

Example:

A specific fast-food restaurant gets an average of three visitors to the drive per minute. This is just an average . However, the actual amount can vary.
Suppose we conduct the Poisson experiment, where the average of success in a given region is .
Then the Poisson probability is:

$$P(x;\ \mu) = (e^{-\mu})\ (\mu^{x})\ /\ x!$$

The graph of Poisson Distribution:



Example:

Question : A life insurance salesman sells on the average 3 life insurance policies per week. Use Poisson's law to calculate the probability.

In a given week he will sell how many policies.
In a given week, he will sell 2 or more policies but not more than five policies.
Assuming that per week, there are 5 working days, what is the probability that on a given day, he will sell one policy?

Solution:
1."Some policies" means "1 or more policies". We can work this out by computing to find 1- "zero policies" probability:
P(X > 0) = 1 - P(x0)

$$\text{Now } P(X) = \frac{e^{-\mu}\mu^{x}}{x!} \text{ so } P(x_0) = \frac{e^{-3}3^{0}}{0!} = 4.9787 \times 10^{-2}$$

Therefore, the probability of 1 or more than 1 policies:
P (X ≥ 0)
=1-4.9787×10-2 because of 1-P(x0)
=0.95021
2. The probability of selling 2 or more policies but less than 5 policies are:

P(2≤X<5) = P(x2)+P(x3)+P(x4)

$$= \frac{e^{-3}3^{2}}{2!} + \frac{e^{-3}3^{3}}{3!} + \frac{e^{-3}3^{4}}{4!}$$

= 0.61611

3. The average number of policies sold per day is 3/5 = 0.6. on given day:

$$P(X) = \frac{e^{-0.6}(0.6)^{1}}{1!} = 0.32929$$

Question: Aluminium alloy sheets(20 sheets) were examined for surface flaws. The frequency of sheets, along with the number of flaws, is as follows:

Number of flaws	Frequency
0	4
1	3
2	6
3	2
4	3
5	1
6	1

When chosen at random, what is the probability of finding a sheet that contains three or more surface flaws?

Solution:

The average number of flaws for twenty sheets is given by:
(0×4)+ (1×3)+ (2×6)+ (3×2)+ (4×3) + (5×1)+ (6×1) = 44
An average number of flaws for twenty sheets is given by:
μ = 44/20 = 2.3
The required probability is:
Probability=P(X≥3)
=1-(P(x0)+P(x1)+P(x2))

$$= 1 - \left(\frac{e^{-2.3}2.3^0}{0!} + \frac{e^{-2.3}2.3^1}{1!} + \frac{e^{-2.3}2.3^2}{2!} \right)$$

=0.40396
7.4 Exponential Distribution:

A continuous probability distribution that times the occurrence of events is known as the exponential distribution. These occurrences are unrelated and occur at a consistent rate. In other words, it calculates the time a person must wait until an event occurs.

The exponential distribution is a continuous probability distribution in probability theory and statistics that often concerns the time until a given event occurs. It is a process in which events occur continuously and independently at a constant average rate. The key attribute of the exponential distribution is that it is memoryless. More little values or fewer more extensive variables can make up the exponential random variable. The sum of money a customer spends on a single trip to the grocery, for example, follows an exponential distribution.

$$f(x;\lambda) = \begin{cases} \lambda e^{-\lambda x}, & \text{if } x \geq 0 \\ 0, & \text{if } x < 0 \end{cases}$$

Where, = Rate parameter
x = Random variable

HYPOTHESIS TESTING

What is hypothesis testing?

In our daily life, we often hear statements like Dhoni is the better captain than his contemporaries, Or a Motorcycle company claiming that a particular model gives an average mileage of 100Km per litre, or a Toothpaste company claiming to be the number one brand suggested by dentists.

Suppose you have to purchase a motorcycle and heard about the above claim made by the Motorcycle company. Would you go and buy it or instead look for the proof of it? There must be a parameter based on which one would judge the correctness of the statement made. In this case, our parameter will be the Average mileage, which you will use to check if the statement made is true or just a hoax.

The hypothesis is a statement, assumption or claim about the value of the parameter(mean, variance, median etc)

A hypothesis is an educated guess about something in the world around you. It should be testable, either by experiment or observation.

For example, if we say that "Dhoni is the best Indian Captain ever." We are making this assumption based on the average wins and losses the team had under his captaincy. We can test this statement based on all the match data.

Simple and Composite Hypothesis:

When a hypothesis specifies an exact value of the parameter, it is a simple hypothesis and if it specifies a range of values then it is called a composite hypothesis.
e.g. Motorcycle company claiming that a certain model gives an average mileage of 100Km per liter, this is a case of a simple hypothesis.

The average age of students in a class is greater than 20. This statement is a composite hypothesis.

Null Hypothesis:

The null hypothesis is the hypothesis to be tested for possible rejection under the assumption that it is true. The concept of the null is similar to innocent until proven guilty We assume innocence until we have enough evidence to prove that a suspect is guilty.
H0 denotes it.

Alternate Hypothesis:
The alternative hypothesis complements the Null hypothesis. It is the opposite of the null hypothesis, such that both the Alternate and null hypotheses cover all the possible population parameter values.
H1 denotes it.

Let's understand this with an example:
A soap company claims that its product kills on average 99% of the germs. To test this company's claim, we will formulate the null and alternative hypotheses.
Null Hypothesis(H0): Average =99%
Alternate Hypothesis(H1): Average is not equal to 99%.
Note : The thumb rule is that a statement containing equality is the null hypothesis.

When we test a hypothesis, we assume the null hypothesis to be true until there is sufficient evidence in the sample to prove it false. In that case, we reject the null hypothesis and support the alternate hypothesis.
If the sample fails to provide sufficient evidence to reject the null hypothesis, we cannot say that the null hypothesis is true because it is based on just the sample data. To say the null hypothesis is true we will have to study the whole population data

Steps involved in Hypothesis testing:

- Setup the null hypothesis and the alternate hypothesis.
- Decide a level of significance i.e. alpha = 5% or 1%
- Choose the type of test you want to perform as per the sample data (z test, t-test, chi-squared, etc.) (we will study all the tests in next section)

- Calculate the test statistics (z-score, t-score, etc.) using the respective formula of test chosen
- Obtain the critical value in the sampling distribution to construct the rejection region of size alpha using z-table, t-table, chi table, etc.
- Compare the test statistics with the critical value and locate the position of the calculated test statistics i.e. is it in the rejection region or non-rejection region.
  - If the critical value lies in the rejection region, we will reject the hypothesis i.e. sample data provide sufficient evidence against the null hypothesis and there is a significant difference between hypothesized value and the observed value of the parameter.
  - If the critical value lies in the non-rejection region, we will not reject the hypothesis i.e. sample data does not provide sufficient evidence against the null hypothesis, and the difference between hypothesised value and the observed value of the parameter is due to fluctuation of the sample.

One-Tail and Two-Tail testing :

If the alternate hypothesis gives the alternate in both directions (less than and greater than) of the value of the parameter specified in the null hypothesis, it is called a two-tailed test.

Suppose the alternate hypothesis gives the alternate in only one direction (either less than or greater than) of the value of the parameter specified in the null hypothesis. In that case, it is called the One-tailed test.

For example:
if H0: mean= 100 H1: mean is not equal to 100
Here, the mean is less than 100, it is called a One-tailed test.

Critical Region :
The critical region is that region in the sample space in which if the calculated value lies then we reject the null hypothesis.

Let's understand this with an example:
Suppose you are looking to rent an apartment. You listed out all the available apartments from different real estate websites. You have a budget of Rs. 15000/ month. You cannot spend more than that. The list of apartments you have made has prices ranging from 7000/month to 30,000/month.

You select a random apartment from the list and assume the below hypothesis:
H0: You will rent the apartment.
H1: You won't rent the apartment.
Since your budget is 15000, you must reject all the apartments above that price.

Here all the Prices greater than 15000 become your critical region. If the random apartment's price lies in this region, you have to reject your null hypothesis and if the random apartment's price doesn't lie in this region, you do not reject your null hypothesis.

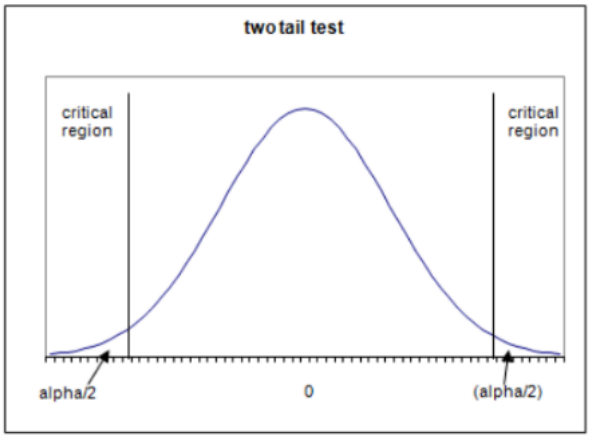
According to the alternative hypothesis, the critical region lies in one tail or two tails on the probability distribution curve. The critical region is a pre-defined area corresponding to a cut-off value in the probability distribution curve. It is denoted by α.
Critical values are values separating the values that support or reject the null hypothesis and are calculated on the basis of alpha.

We will see more examples later on and it will be clear how we choose α.

Based on the alternative hypothesis, three cases of critical region arise :

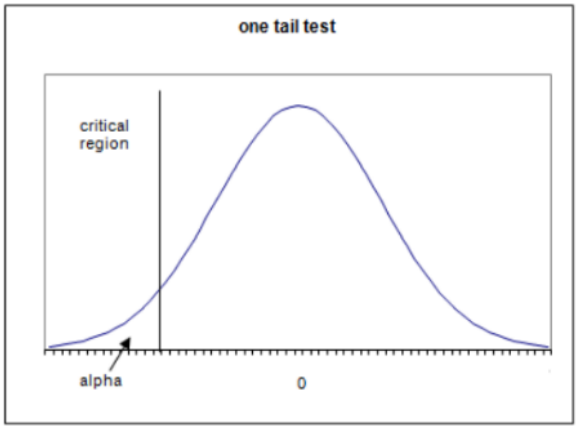
Case (1) This is a double tailed test.

$H_0 : \mu = \mu_0$ $H_1 : \mu \neq \mu_0;$



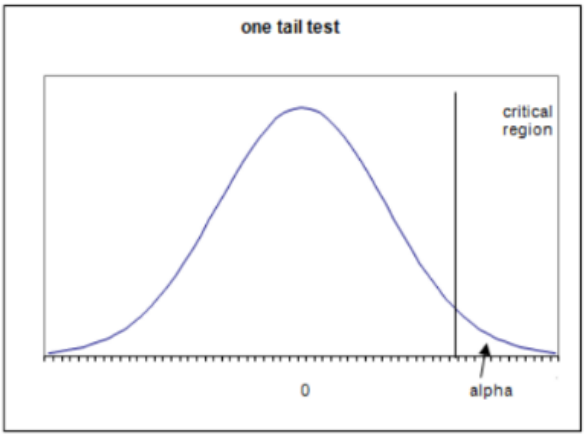
Case (2) This scenario is also called the Left-tailed test.

$H_0 : \mu = \mu_0$ $H_1 : \mu < \mu_0;$



Case (3) This scenario is also called the Right-tailed test.

$H_0 : \mu = \mu_0$ $H_1 : \mu > \mu_0;$



TYPE I ERROR AND TYPE II ERRORS

Decision	H0 True	H0 False
Reject H0	Type 1 error	Correct Decision
Do not reject H0	Correct Decision	Type 2 error

A false positive (type I error) — when you reject a true null hypothesis.
A false negative (type II error) — is when you accept a false null hypothesis.

The probability of committing a Type I error (False positive) is equal to the significance level or size of critical region α.
α= P [rejecting H0 when H0 is true]

The probability of committing a Type II error (False negative) is equal to the beta β and is called the 'power of the test'.
β = P [not rejecting H0 when H1 is true]

Example : A person is arrested on the charge of being guilty of burglary. A jury of judges has to decide guilty or not guilty.

H0: Person is innocent
H1: Person is guilty

Type I error will be if the Jury convicts the person [rejects H0] although the person was innocent [H0 is true].
Type II error will be the case when the Jury releases the person [Do not reject H0] although the person is guilty [H1 is true].

Level of Significance(α):

It is the probability of type 1 error. It is also the size of the critical region.
Generally, a strong control on α is desired and in tests, it is pre-fixed at very low levels like 0.05(5%) or 01(1%).

If H0 is not rejected at a significance level of 5%, then one can say that our null hypothesis is true with 95% assurance.

P-Values:

Let's suppose we are conducting a hypothesis test at a significance level of 1%.
Where, H0: mean<X (we are just assuming a scenario of 1 tail test.)
We obtain our critical value (based on the type of test we are using) and find that our test statistic is greater than the critical value. So, we have to reject the null hypothesis here since it lies in the rejection region. Now if the null hypothesis is getting rejected at 1%, then for sure it will get rejected at the higher values of significance level, say 5% or 10%.

What if we take a significance level lower than 1%, would we have to reject our hypothesis then also?

Yes, there might be a chance that the above scenario can happen, and here comes "p-value" into play.

The p-value is the smallest level of significance at which a null hypothesis can be rejected.

That's why many tests nowadays give a p-value and it is more preferred since it gives out more information than the critical value.

For right-tailed test:
p-value = P[Test statistics >= observed value of the test statistic]

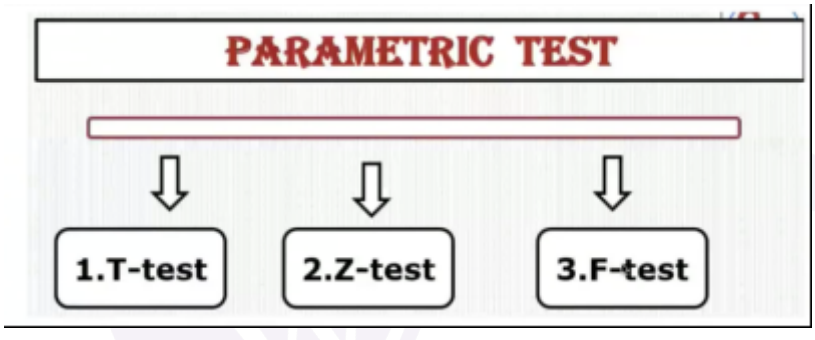
For left tailed test:
p-value = P[Test statistics <= observed value of the test statistic]

For two-tailed test:
p-value = 2 * P[Test statistics >= |observed value of the test statistic|]

Decision making with p-value

The p-value is compared to the significance level(alpha) for decision making on null hypothesis.
If the p-value is greater than alpha, we do not reject the null hypothesis.
If the p-value is smaller than alpha, we reject the null hypothesis.

T TEST
What is a T-test :
A t-test is a statistical method used to see if two sets of data are significantly different. T-tests are used to involve data analysis that has applications in business, science, and many other disciplines.



The T distribution is bell-shaped & symmetric, like the normal distribution, but has more massive tails. The T distribution is the family of distributions that looks identical to the normal distribution curve.

Only a bit shorter and fatter.

It is used in place of the normal distribution when we have small samples. (n<30)
The T distribution is similar to the normal distribution if the sample size increases

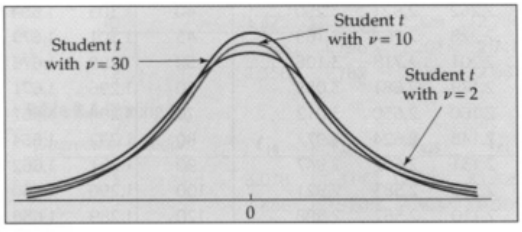
Here the letter t is used to represent the random variable, hence the name. The density function for the Student t distribution is as follows,

$$f(t) = \frac{\Gamma[(\nu+1)/2]}{\sqrt{\nu\pi}\,\Gamma(\nu/2)}\left[1+\frac{t^2}{\nu}\right]^{-(\nu+1)/2}$$

ν (nu) is called the degrees of freedom, and

Γ (Gamma function) is $\Gamma(k)=(k-1)(k-2)\ldots(2)(1)$

In much the same way that μ and σ define the normal distribution, ν, the degrees of freedom, defines the student t distribution:

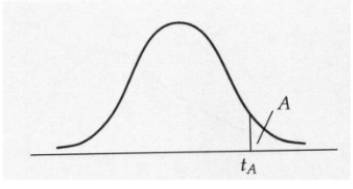


As the number of degrees of freedom increases, the t distribution approaches the standard normal distribution.

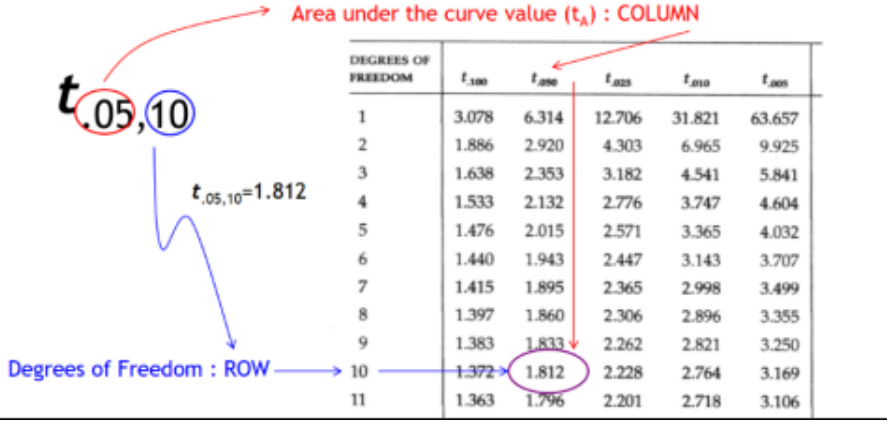
Determining Student values :

The Student 't' distribution is used extensively in statistical inferences.
The value of a Student 't' random variable with V degrees of freedom is:

$$P(t > t_{A,\nu}) = A$$



The values for 'A' are pre-determined 'critical' values typically in the 10%,5%,2.5%,1% and 0.5%.
Example : The value of t with degrees of freedom 10, so that the area under the Student 't' curve is 0.05:



DEGREES OF FREEDOM	$t_{.100}$	$t_{.050}$	$t_{.025}$	$t_{.010}$	$t_{.005}$
1	3.078	6.314	12.706	31.821	63.657
2	1.886	2.920	4.303	6.965	9.925
3	1.638	2.353	3.182	4.541	5.841
4	1.533	2.132	2.776	3.747	4.604
5	1.476	2.015	2.571	3.365	4.032
6	1.440	1.943	2.447	3.143	3.707
7	1.415	1.895	2.365	2.998	3.499
8	1.397	1.860	2.306	2.896	3.355
9	1.383	1.833	2.262	2.821	3.250
10	1.372	1.812	2.228	2.764	3.169
11	1.363	1.796	2.201	2.718	3.106

The t distribution has the following properties:

- The distribution's mean = 0 .
- v / (v - 2) is equal to variance, where v is the degrees of freedom.
- The variance is although close to 1, but it is always greater than 1 when degrees of freedom is greater.
- The t distribution is similar to the standard normal distribution, with infinite degrees of freedom.
-

T-Distribution:

When 'n' - the sample size, is drawn from a population having a normal distribution, using the equation of t-distribution, the sample mean is transformed into a t statistic. Following is the equation:

t = [X - μ] / [S / sqrt(n)]

where, μ is the population mean, X is the sample mean, and the sample size is n, and n – 1 is degrees of freedom here and s is the standard deviation of the sample.

Example : The Company by the name Acme Corporation manufactures light bulbs. The company's CEO claims that an average Acme light bulb lasts 300 days. So for testing, 15 bulbs are selected by a researcher randomly. An average sample bulb lasts for 290 days, with 50 days of standard deviation. What is the probability that 15 randomly selected bulbs do not have an average life of more than 290 days?

Solution :
First we need to calculate the t statistic :
t-distribution 't' = [X - μ] / [S / sqrt(n)]
t = (290 - 300) / [50 / sqrt(14)]
t = -10 / 12.909945 = - 0.7745966
where, μ is the population mean, X is the sample mean, the standard deviation of the sample is s and the sample size is 'n.'
As we are aware of the t statistic, we can select the "T score" from the Random Variable dropdown box and enter the following data:
14 - 1 = 13 (The degrees of freedom)
= - 0.7745966. is t statistics

Example : IQ test's scores are normally distributed, with 100 as the population mean. Suppose 20 people are selected randomly and then tested.15 is the standard deviation in the sample group. What is the probability that in the sample group, an average test score will be at-most 110?.

Solution :
20 - 1 = 19 is equal to degrees of freedom.
100 is the population mean.
110 is the sample mean
15 is the standard deviation
0.996 is the Cumulative Probability
So the chance that the sample average will be no higher than 110 is 99.6%

Independent Sample T-test:

Sample that is unrelated to the other The t-test is a statistical tool for analysing the mean difference between two independent groups. When two samples from the same population are used in an independent sample t-test, the mean of the two samples may be the same. However, if samples are gathered from two different populations, the sample mean may differ. It is used to derive judgments about the means of two populations in this situation, and to determine whether they are similar.

Assumptions in independent samples t-test :

- Assume that the dependent variable is normally distributed.
- Assume that the two samples are independent of each other.
- Assume that the two samples are independent of each other.
- Samples are drawn from the population at random.
- In independent samples t-test, all observations must be independent of each other.
- In independent samples t-test, dependent variables must be measured on an interval or ratio scale.

Procedures for independent samples t-test :

Null Hypothesis : It is assumed when the means of the two groups are not significantly different.

Alternative Hypothesis : Assume that the means of the two groups are significantly different.

Calculate the standard deviation for the independent sample t-test by using this formula:

$$S = \sqrt{\frac{\sum (X_1 - \overline{X}_2)^2 + \sum (X_2 - \overline{X}_2)^2}{n_1 + n_2 - 2}}$$

Calculate the value of the independent sample t-test by using this formula:

$$t = \frac{\overline{X}_1 - \overline{X}_2}{S} \cdot \sqrt{\frac{n_1 n_2}{n_1 + n_2}}$$

Degree of freedom for independent sample t-test :
V= n1+n2 -2

Dependent Sample T-test:

Dependent samples t-test is conducted when the observations of one sample group is known to be related in some way to the other observations of the sample group. So typically, the dependent sample t-test compares the means of those two related sample groups.

This type of statistical t-test method is mostly used when you know you are studying similar specimens or relatable units or even when there are repeated measurements taken for a sample group.

$$t = \frac{\Sigma d}{\sqrt{\frac{n(\Sigma d^2) - (\Sigma d)^2}{n-1}}}$$

where d: difference per paired value

n: number of samples

ANOVA & CHI_SQUARE

Analysis of Variance(ANOVA) :

Analysis of variance (ANOVA) is a statistical technique used to check if the means of two or more groups are significantly different by analyzing comparisons of variance estimates. ANOVA checks the impact of one or more factors by comparing the means of different samples.

The t-test and ANOVA give the same results when we have only two samples. However, using a t-test would not be reliable in cases where there are more than 2 samples. If we conduct multiple t-tests for comparing more than two samples, it will compound the type 1 error.

Assumption in ANOVA.

Assumption of Randomness: The samples should be selected in a random way such that there is no

dependence among the samples.
The experimental errors of the data are normally distributed.

Assumption of equality of variance(Homoscedasticity) and zero correlation:
The variance should be constant in all the groups and all the covariance among them are zero although means vary from group to group.

One-Way ANOVA:

When we are comparing groups based on only one factor variable then it is said to be a one-way analysis of variance(ANOVA).
For example, if we want to compare whether or not the mean output of three workers is the same based on the working hours of the three workers.

The ANOVA model :

Mathematically, ANOVA can be written as: Xij = i + ij Where x are the individual data points(i and j denote the group and the individual observation), is the unexplained variation and the parameters of the model () are the population means of the population means of each group. Thus, each data point(Xij) is its group mean plus error.

Let's understand the working procedure of One-way Anova with an example:

Sample (k)	1	2	3	Mean
1	X11	X12	X13	Xm1
2	X21	X22	X23	Xm2
3	X31	X32	X33	Xm3
4	X41	X42	X43	X34

Suppose we are given with the above data set we have an independent variable x and 3 samples with different values of x and each sample has its respective mean as shown in the last column.

Grand Mean :

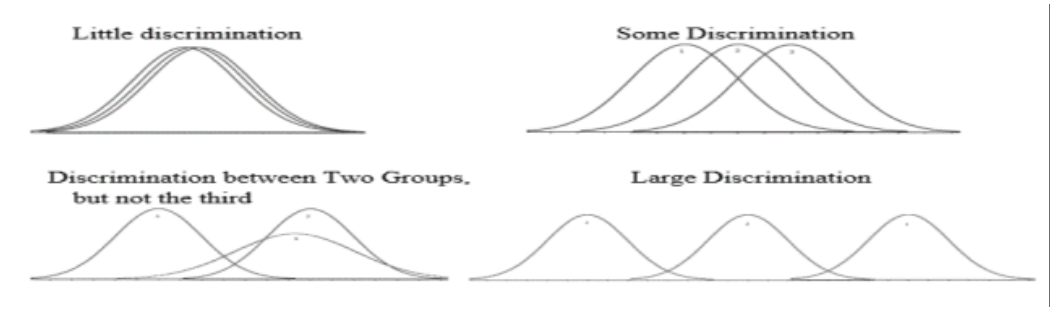
Mean is a simple or arithmetic average of a range of v alues. There are two kinds of means that we use in ANOVA calculations. Which are separate sample means and the grand mean.

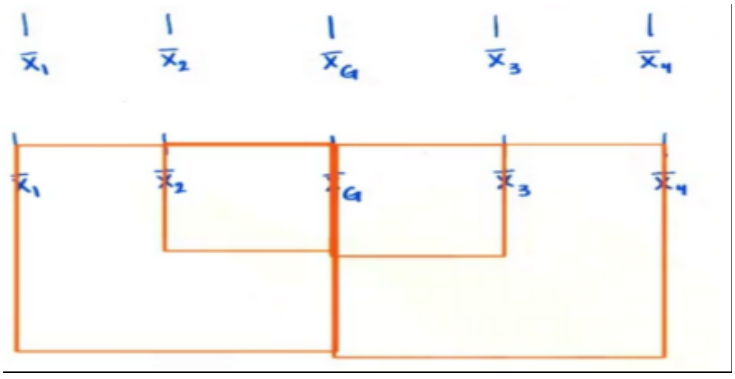
The grand mean (Xgm) is the mean of sample means or the mean of all observations combined, irrespective of the sample. Xgm = (Xm1 + Xm2 + Xm3 + Xm4 +..... Xmk)/k where, k is the number of samples
For our dataset, k = 4
Xgm = (Xm1 + Xm2 + Xm3 + Xm4)/4

Between Group Variability(SST):

It refers to variations between the distribution of individual groups(or levels) as the values within each group are different.

Each sample is looked at and the difference between its mean and grand mean is calculated to calculate the variability. If the distributions overlap or are close, the grand mean will be similar to the individual means whereas if the distributions are far apart, the difference between means and grand mean would be large.





Let's calculate Sum of Squares for between-group variability:

SSbetween = n1 * (Xm1 - Xgm)2 + n2 * (Xm2 - Xgm)2 + n3 * (Xm3 - Xgm)2 + ... + nk * (Xmk - Xgm)2
where, n1, n2,....,nk are the number of observations in each sample

Degree of freedom for between-group variability = number of samples – 1 = k-1 MeanSSbetween = SSbetween/ k-1
In our dataset example we have k =4 and nk = 3, so for our dataset: SSbetween = 3 * (Xm1 - Xgm)2 + 3 * (Xm2 - Xgm)2 + 3 * (Xm3 - Xgm)2 + 3 * (Xm4 - Xgm)2
MeanSSbetween(MSST) = SSbetween/ (4-1) = SSbetween/3

Within Group Variability(SSE) :

It refers to variations caused by differences within individual groups (or levels) as not all the values within each group are the same. Each sample is looked at on its own and variability between the individual points in the sample is calculated. In other words, no interactions between samples are considered.

We can measure Within-group variability by looking at how much each value in each sample differs from its respective sample mean. So, first, we'll take the squared deviation of each value from its respective sample mean and add them up. This is the sum of squares for within-group variability.

$$SS_{within} = \Sigma(x_{i1} - \bar{x}_1)^2 + \Sigma(x_{i2} - \bar{x}_2)^2 + \ldots + \Sigma(x_{ik} - \bar{x}_k)^2 = \Sigma(x_{ij} - \bar{x}_j)^2$$

Note: x_{i1} is the ith value from the first sample, x_{i2} is the ith value from the second sample, and so on all the way to x_{ik}, the ith value from the kth sample. x_{ij} is therefore the ith value from the jth sample.

$$df_{within} = (n_1 - 1) + (n_2 - 1) + \ldots (n_k - 1) = n_1 + n_2 + n_3 + \ldots. + n_k - k(1) = N - k$$

$$MS_{within} = \sum(x_{ij} - \overline{x_j})^2 / (N - k)$$

Where N is the total number of observations.

In our dataset example we have k=4 an N=12, so for our dataset:

SSwithin = (X11 - Xm1)2 + (X12 - Xm1)2 + (X13 - Xm1)2 + (X21 - Xm2)2 + (X22 - Xm2)2 + (X23 - Xm2)2 + (X31 - Xm3)2 + (X32 – Xm3)2 + (X33 – Xm3)2 + (X41 - Xm4)2 + (X42 - Xm4)2 + (X43 - Xm4)2

Degree of freedom = N-k=12-4=8
Mean(sswithin) MSSE = SSwithin/8.

Total sum of square (TSS) :
TSS = SSbetween + SSwithin = SST + SSE

Hypothesis in ANOVA :

The Null hypothesis in ANOVA is valid when all the sample means are equal, or they don't have any significant difference. Thus, they can be considered as a part of a larger set of the population. On the other hand, the alternate hypothesis is valid when at least one of the sample means differs from the rest. In mathematical form, they can be represented as:

$$H_o: \ \mu_1 = \mu_2 = \cdots = \mu_L \qquad \textit{Null hypothesis}$$

$$H_1: \ \mu_l \neq \mu_m \qquad \textit{Alternate hypothesis}$$

where μ1 and μm belong to any two samples out of all the samples considered for the test. In other words, the null hypothesis states that all the sample means are equal or the factor did not have any significant effect on the results. Whereas, the alternate hypothesis states that at least one of the sample means is different from another. To test the null hypothesis, test statistics are given by the F-statistic.

Two-Way ANOVA:

Mathematically, ANOVA can be written as:

xij = μij + εij

where x are the individual data points (i and j denote the group and the individual observation), ε is the unexplained variation and the parameters of the model (μ) are the population means of each group. Thus, each data point (xij) is its group means plus error.

Just like a one-way model, we will calculate the sum of squares between, in this case, there will be two SSTs for both the categories and the sum of squares of errors (within).
We calculate F-statistics for both the MSST and see which one is a greater value than F-critical and compare them to find the effect of both categories on our assumption.

Example:
Below given is the data of the yield of crops based on temperature and salinity.
Calculate the ANOVA for the table.

Temperature (in F)	Categorical variable salinity				
	700	1400	2100	Total	Mean(temp)
60	3	5	4	12	4
70	22	10	12	33	11
80	16	21	17	54	18
Total	30	36	33	99	11
Mean(salanity)	10	12	11	11	

Solution:
Hypothesis for temperature
H0 : Yield is the same for all temperatures.
H1 : Yield varies with temperature with significant differences.
Hypothesis for Salinity:
H0 : Yield is same for all salinity.
H1 : Yield varies with temperature with significant salinity
Grand mean = 11
N = 9, K=3, n=3
N = 9, K =3, nt= 3, ns = 3
SSbetween_temp = 3 *(4-11)2 + 3*(11-11)2 + 3*(18-11)2 = 294
MSSTtemp = 294 / 2 = 147
SSbetween_salanity = 3 *(10-11)2 + 3*(12-11)2 + 3*(11-11)2 = 6
MSSTsalainity = 6 /2 = 3

In such a question calculating SSE can be tricky, so instead of calculating SSE let's calculate TSS then we can subtract SST values from it and get SSE. To calculate the total sum of squares, we need to find the sum of the squares of the difference of each value from the grand mean.

TSS = (3-11)2 + (5-11)2 + (3-11)2 +(4-11)2 +(11-11)2 +(10-11)2 +(12-11)2 +(16-11)2 +(21-11)2 +(17-11)2
TSS = 312

SSE = TSS - SSbetween_temp - SSbetween_salanity = 312 – 294-6 = 12
Degree of freedom for SSE = (nt-1)(ns-1) =(3-1)(3-1) = 4
MSSE = SSE/4 = 3

F-Test For temperature
Ftemp = MSSTtemp/ MSSE = 14/3 = 49
F-Test For Salinity
Fsalinity = MSSTsalinity/MSSE = 3/3 = 1

F-critical for 5% significance and degree of freedom (k-1, (p-1) (q-1)) i.e. (2,4):
F-critical = 10.649

Clearly, we can see that Ftemp is greater than F-critical, so we reject the null hypothesis and support that temperature has a significant effect on yield. On the other hand, Fsalinity is less than the F-critical value, so we do not reject the null hypothesis and support that salinity doesn't affect the yield.

CHI SQUARE TEST

Chi-square is one of the most important non-parametric distributions. It can be defined as Sum of the square of independent normally distributed variables with zero means and unit variances.
Hence, if

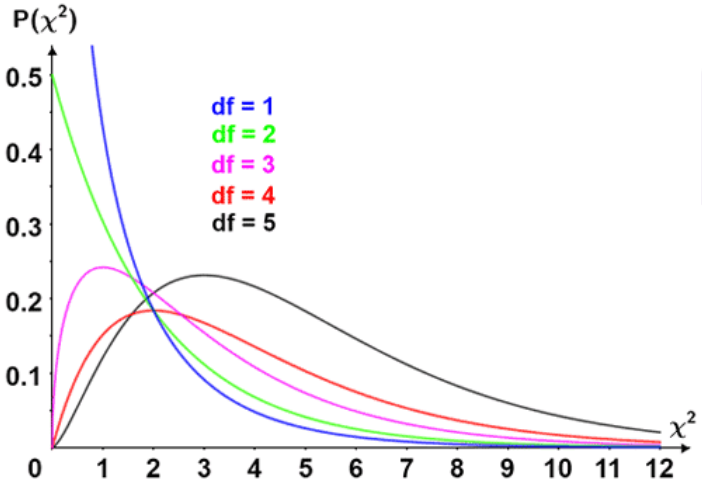
$$X \sim N(\mu, \sigma^2), \text{ then } Z = \frac{X-\mu}{\sigma} \sim N(0,1)$$

$$Z^2 = \left(\frac{X-\mu}{\sigma}\right)^2 \sim \chi^2_{1\ d.f}$$

with one degree of freedom.
In general if (=1,2,...,) and independent normal variates with means and variances 2 (=1,2,...,), then

$$\chi^2 = \sum_{i=1}^{n} \left(\frac{X_i - \mu_i}{\sigma_i}\right)^2 \sim \chi^2_{n\ d.f}$$

Graph of probability density function of χ^2 distribution for different degree of freedom.



As we can see in the graph that f $n>2, f(x)$ is monotonically increasing $0<x<(n-2)$ d monotonically decreasing for $(n-2)<x<\infty$ white a $x=n-2$, it attains the maximum value.

$\chi 2 \to N$ as $n \to \infty$
In general for $n \geq 30$, the x^2 approximation to normal is fairly good. So whenever $n \geq 30$, we use the normal probability tables for testing the significance of value of x^2 . Hence in x^2 tables the significant values have been given till $n=30$.

Chi-Square test for goodness of fit:

A very powerful test for testing the significance of the discrepancy between theory and experiment is known as chi-square goodness of fit. It enables us to find if the deviation of the experiment from theory is just by chance or is it really due to the inadequacy of the theory to fit the observed data.

if *fi(i=1,2,...,n)* is a set of observed frequencies and *ei(i=1,2,...,n)* is the corresponding set of expected (theoretical or hypothetical) frequencies, then chi-square given by

$$\chi^2 = \sum_{i=1}^{2} \left[\frac{(f_i - e_i)^2}{e_i} \right] \sim \chi^2_{n-1 \ d.f}$$

Chi-Square test independence:

The Chi-square test of independence is a statistical hypothesis test that is used to see if two categorical or nominal variables are likely to be connected.

When you have counts of values for two categorical variables, you can apply this test.

The chi-square test is widely used to estimate how closely the distribution of a categorical variable matches an expected distribution (the goodness-of-fit test), or to estimate whether two categorical variables are independent of one another (the test of independence).

In mathematical terms, the $\chi 2$ variable is the sum of the squares of a set of normally distributed variables. Suppose that a particular value $Z1$ is randomly selected from a standardized normal distribution. Then suppose another value $Z2$ is selected from the same standardized normal distribution. If there are d degrees of freedom, then let this process continue until d different Z values are selected from this distribution. The $\chi 2$ variable is defined as the sum of the squares of these Z values.

This sum of squares of d normally distributed variables has a distribution which is called the $\chi 2$ distribution with d degrees of freedom.

Covariance and Correlation

Covariance
It is a kind of variance between two variables.
It is the relationship between a pair of random variables where change in one variable causes change in another variable.

It can take any value between -infinity to +infinity, where the negative value represents the negative relationship whereas a positive value represents the positive relationship.

It is used for the linear relationship between variables.
It gives the direction of the relationship between variables.
It has unit.

$$cov(x, y) = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{n}$$

Since the covariance has units hence we cannot compare the data having different units Hence to normalize them we divide the covariance by individual standard deviations. Now the coefficient is unit free and can be used for compare different quantities. This is called correlation coefficient.

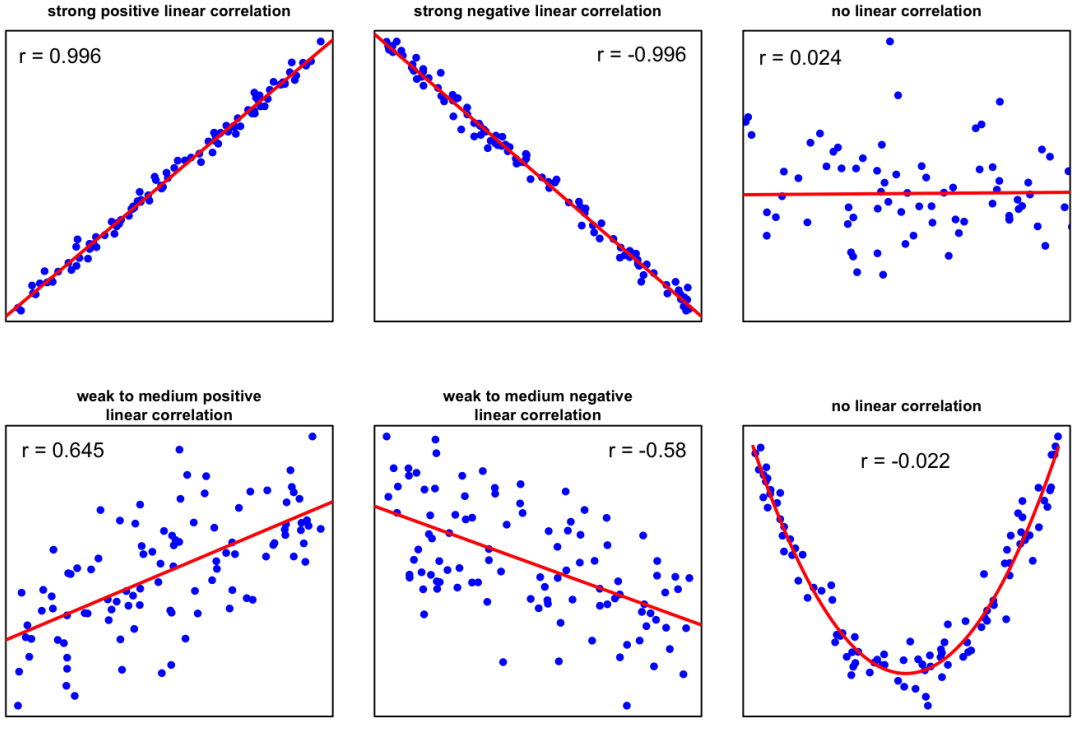
Correlation

For bivariate distribution correlation represents the extent of linear relationship between the variables. But it doesn't tell us about the non-linear relationship. It shows whether and how strongly pairs of variables are related to each other.
Correlation takes values between -1 to +1, wherein values close to +1 represents strong positive

correlation and values close to -1 represents strong negative correlation.
These variables are indirectly related to each other.
It gives the direction and strength of the relationship between variables.
It is the scaled version of Covariance.
It is unit less.

$$Correlation(x, y) = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=n}^{n}(x_i - \bar{x})\sum_{i=1}^{n}(y_i - \bar{y})}}$$

Here,
$\bar{x}$ and $\bar{y}$ = mean of given sample set
n = total number of sample
xi and yi = individual sample of set



Example
x = [1,2,3,4,5,6,7,8,9]
y = [9,8,7,6,5,4,3,2,1]

Find the correlation between x and y.

Solution We can clearly see in the dataset that as x increase y decreases and vice versa.

Let's prove this with the formula we have studied above.

x	y	$x-\bar{x}$	$y-\bar{y}$	$(x-\bar{x})2$	$(y-\bar{y})2$	$(x-\bar{x})*(y-\bar{y}$
1	9	-4	4	16	16	-16
2	8	-3	3	9	9	-9
3	7	-2	2	4	4	-4
4	6	-1	1	1	1	-1
5	5	0	0	0	0	0
6	4	1	-1	1	1	-1
7	3	2	-2	4	4	-4
8	2	3	-3	9	9	-9
9	1	4	-4	16	16	-16
45	45			60	60	-60

*Corr(x,y)=-60/[(60*60)1/2]=-1*

When to use correlation and what it tells us ?

Use correlation when you want to quantify the linear relationship between two variables and neither of the variables represents a response or "outcome" variable.
Correlation analysis is used to quantify the degree to which two variables are related. Through the correlation analysis, you evaluate the correlation coefficient that tells you how much one variable changes when the other one does.
Correlation analysis provides you with a linear relationship between two variables.

Pearson correlation:

The Pearson correlation coefficient measures how strong a linear link between two variables is. Its value ranges from -1 to 1, with -1 indicating a total negative linear correlation, 0 indicating no correlation, and + 1 indicating a total positive linear correlation. The Spearman correlation assesses the strength of a monotonic association between two variables scaled similarly to the Pearson correlation.

Suppose that a group of n individuals is arranged in order of merit in possession of two characteristics A and B. These ranks in the two characteristics will, in general, be different. For example, consider the relation between intelligence and beauty. It is not necessary that beautiful individuals are intelligent also. Now we are using the rank of the data instead of the data itself and that is given by pearson's correlation coefficient.

Formula:

r = (xi-x)(yi-y)(xi-x)2 (yi-y) 2

Where,
r = correlation coefficient
xi = values of the x-variable in sample
x = mean of the values of the x-variable
yi = values of the y-variable in sample
y = mean of the values of the y-variable

Types of correlation coefficients:

The types of correlation coefficients:

1. Covariance correlation coefficient
2. Pearson correlation coefficient
3. Spearman's correlation coefficient

Covariance correlation coefficient:

It is a kind of variance between two variables.

1. It is the relationship between a pair of random variables when change in one variable causes changes in another variable.
2. It can take my value between - to + where the negative value represents the negative relationship whereas a positive value represents the positive relationship.
3. It is used for the linear relationship between two variables.
4. It gives the direction of the relationship between variables
5. It has a unit.

$$cov(x, y) = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{n}$$

Since the variance has units hence we can not compare the data having different units Hence to normalise them we divide the covariance by individual standard deviation. Now the coefficient is unit free and can be used to compare different quantities. This is called correlation coefficient.

Pearson correlation coefficient:

For Pearson's correlation, there is also a need for a linear relationship between a pair of variables. In the literature, it can be called "Pearson product-moment correlation"

Pearson's correlation formulation can be generated from the covariance correlation mathematical formula.

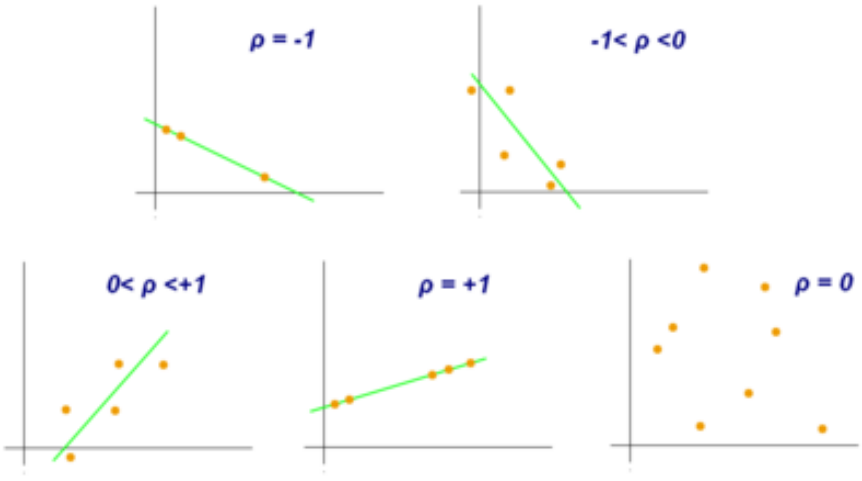
Pearson(X, Y) = Covariance(X, Y) / (stdDev(X) * stdDev(Y))

Pearson correlation coefficient parameters may be observed in five different ranges according to the variables' current location lie on the x and y-axis, correlation's range may be subject to change.

When the data points follow a downward trend, it can be accepted as a Negative Correlation. On the other hand, when a trend is observed to be upward, it can be labeled as a Positive Correlation between the two

compared variables.

In the below depiction, p can be accepted as the correlation coefficient. You may examine the alignments of the correlation according to the distributions of the data points.



Spearman's correlation coefficient:

Suppose that a group of n individuals is arranged in order of merit in possession of two characteristics A and B. These ranks in the two characteristics will, in general, be different. For example, consider the relationship between intelligence and beauty. It is not necessary that beautiful individuals are intelligent also. Now we use the rank of the data instead of the data itself and that is given by pearson's correlation coefficient.

$$\rho = 1 - \frac{6\sum_{i=1}^{n} d_i^2}{n(n^2 - 1)}$$

Where,
p= Spearman's correlation coefficient
di = Different is ranks