**Module 1 Overview of Statistics**

**Introduction:-**The science of collecting,describing and analyzing data is popularly known as statistical leveraging in Data Science.

standar

Many of our day to day assumptions and decisions already depends on it.Take a simple example:suppose 2 of my adult friends are sitting in the next room while one is 5 feet tall and other is 6 feet tall.

What would be your best guess as to each one’s gender,based on that information alone??

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

I expect you felt fairly confident in assuming that friend 5 feet tall was female and 6 feet tall was male. You could have been wrong also , of course experience tells you that five foot men and six foot women are somewhat rare.You have noticed that,by and large men tend to betaller than women.

Again the assumption can be wrong as we have not seen all the men and women.

The above is a simple,everyday example of statistical thinking.

Let’s explore more of them

1)I cycle about 100 miles a week

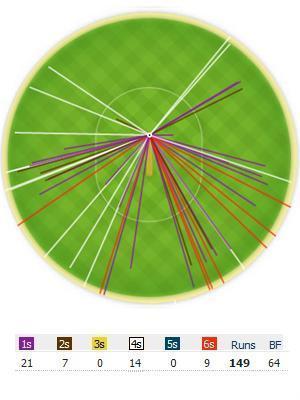
2)We can expect a lot of rain at this time of year.

Above 2 examples past experience is being summarized in a rough and ready way.

In 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 above

Here by using different statistical methods ICC compares players, 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.

**Types of Statistics**

**1)Descriptive Statistics**:

The type of statistics dealing with numbers (numerical facts, figures, or information) to describe any phenomena. These numbers are descriptive statistics.

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

**2)Inferential statistics**

Inferential statistics is a decision, estimate, prediction, or generalization about a population, based on a sample.

A population is a collection of all possible individuals, objects, or measurements of interest.

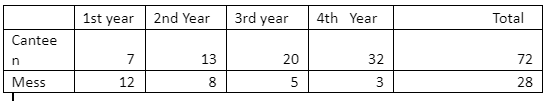
A sample is a portion, or part, of the population of interest.

Inferential statistics is used to make inferences from data whereas descriptive statistics simply describes what’s going on in our data.

Let’s clear our understanding about the above types with a basic scenario:

Suppose in your college there are 1000 students. You are interested in finding out of how many students prefer eating in the college canteen than college mess.

A random group of 100 students is selected. Here our population size is of 1000 students and the sample size is of 100 students. You surveyed the sample group and got the following results:



Let’s analyze the data:

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

2)Of the total students who prefer canteen, 44.4 % are from the 4th year.

3)Of the total students who prefer canteen, 72% are from the 3rd year and 4th year.

5)1st year students are more inclined towards eating in the mess.

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

Now, suppose you got a contract to open a canteen in the College.

Now with the above data, you can make following assumptions:

1)3rd and 4th year students are the main target for restaurant sales.

2)You can give discounts to the 1st year students to increase the number count.

Since most students prefer eating in canteen, opening a canteen can be profitable business

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

Though the above example is very basic and the real scenarios are much more complex, this would help in getting the underlying difference. We will see more complex examples ahead.

Q) The average salary of employees of a company in 2017 is greater than the average salary of teachers of school in 2017.

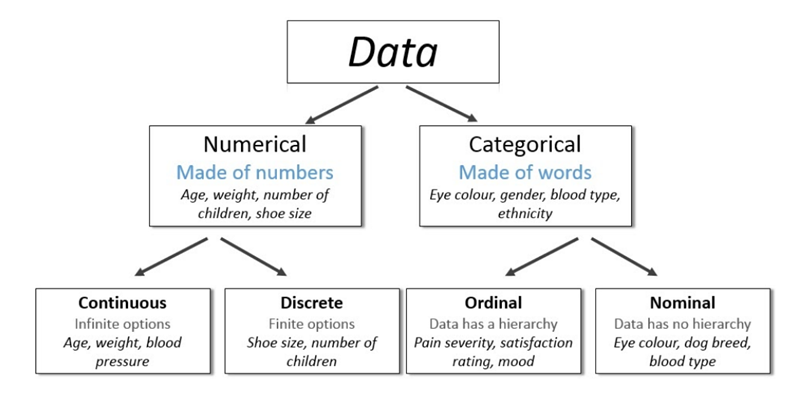
Is the above statement an example of descriptive or inferential statistics?

.

Q) By 2030 , World will face the shortage of water.

**Inferential statistics is used to make inferences from data whereas descriptive statistics simply describes what’s going on in our data**.

**Types of Data**



*Example*: Identify each of the following as examples of qualitative or numerical variables:

1. The temperature in Barrow, Alaska at 12:00 pm on any

given day.

2. The make of automobile driven by each faculty member.

3. Whether or not a 6 volt lantern battery is defective.

4. The weight of a lead pencil.

5. The length of time billed for a long distance telephone call.

6. The brand of cereal children eat for breakfast.

7. The type of book taken out of the library by an adult.

*Example*: Identify each of the following as examples of

(1)nominal, (2) ordinal, (3) discrete, or (4) continuous variables:

1. The length of time until a pain reliever begins to work.
2. The number of chocolate chips in a cookie.
3. The number of colors used in a statistics textbook.
4. The brand of refrigerator in a home.
5. The overall satisfaction rating of a new car.
6. The number of files on a computer’s hard disk.
7. The pH level of the water in a swimming pool.
8. The number of staples in a stapler.

**Population:** A collection, or set, of individuals or objects or events whose properties are to be analyzed.

Two kinds of populations: finite or infinite.

**Sample:** A subset of the population.

**Variable:** A characteristic about each individual element of a population or sample.

**Data (singular):** The value of the variable associated with one element of a population or sample. This value may be a number, a word, or a symbol.

**Data (plural):** The set of values collected for the variable from each of the elements belonging to the sample.

**Experiment:** A planned activity whose results yield a set of data.

**Parameter:** A numerical value summarizing all the data of an entire population.

**Statistic:** A numerical value summarizing the sample data.

**Module 2- Harnessing Data**

In the last module we have seen 2 types of statistics and that 2 we will focus now.The descriptive statistics has following stages

1)Collecting the data

2)Presenting the data-->Visualization using matplotlib and seaborn

3)Summarizing the data-->Module 3

**Collecting Data:-**

The most important point to keep while collecting the data is that whatever sample that has been collected should represent the population.In order to be representative ,members for the sample must be chosen at random from population.That is each member of the population should have equal chance of being chosen for the sample.This is always not easy to achieve.

* **Process Of Data Collection**

**Step1:-**Define the object or aim of the experiment.

I.e Estimate the average life of electronic component

**Step2:-**Define the variable and population of interest.

I.e usage,power rating,battery life etc

**Step3:-**Defining the data collection scheme and data measuring scheme.

I.e sampling procedure,sample size,data measuring device.

**Step4:-**Defining the appropriate descriptive and inferential analysis techniques

* **Methods used to collect data**

1. **Experiment:** The investigator controls or modifies the environment and observes the effect on the variable under study.
2. **Survey**: Data are obtained by sampling some of the population of interest. The investigator does not modify the environment.
3. **Census:** A 100% survey. Every element of the population is listed. Seldom used: difficult and time-consuming to compile, and expensive.
4. **Judgment Samples:** It is a non-probability sampling technique in which the sample members are chosen only on the basis of the researcher's knowledge and judgment.

Consider a scenario where a panel decides to understand what are the factors which lead a person to select ethical hacking as a profession. Ethical hacking is a skill which has been recently attracting youth. More and more people are selecting it as a profession. The researchers who understand what ethical hacking is will be able to decide who should form the sample to learn about it as a profession. That is when judgmental sampling is implemented. Researchers can easily filter out those participants who can be eligible to be a part of the research sample.

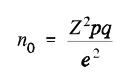
1. **Probability Samples:** Samples in which the elements to be selected are drawn on the basis of probability. Each element in a population has a certain probability of being selected as part of the sample.

* **Probability Sampling types**

Many datasets are samples from an infinite population.We are most interested in measures on the population, but we have access only to sample of it.Lets see the types of sampling(Random) available

1. **simple random sample** :-each sample of the same size has an equal chance of being selected
2. **stratified sample** :-divide the population into groups called strata and then take a sample from each stratum
3. **cluster sample** :-divide the population into strata and then randomly select some of the strata. All the members from these strata are in he cluster sample.
4. **systematic sample** :-randomly select a starting point and take every n-th piece of data from a listing of the population.

**Cochran’s Formula:-**It calculates an ideal sample size given a desired level of precision,described confidence intervals and estimated proportion of the attribute present in the population.



Where:

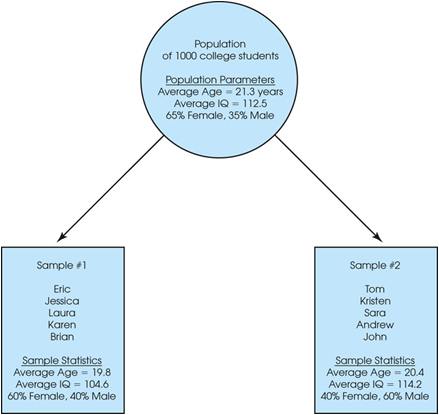
e is the desired level of precision (i.e. the margin of error),

p is the (estimated) proportion of the population which has the attribute in question,

q is 1 – p.

* **Sampling Error**

The actual process of sampling causes sampling errors. The discrepancy between a sample statistic and its population parameter is called sampling error. Defining and measuring sampling error is a large part of inferential statistics



A demonstration of sampling error. Two samples are selected from the same population. Notice that the sample statistics are different from one sample to another, and all of the sample statistics are different from the corresponding population parameters. The natural differences that exist, by chance, between a sample statistic and a population parameter are called sampling error.

++++++++++++++++++++++++++++++++++++++++++++++++++++++++

**Module 3**

**Exploratory Analysis(Summarizing the data)**

**1)Measures of Central Tendencies**

By central tendency we mean the tendency of the observations to pile up around a particular value or in a particular group rather than spread themselves evenly across the range or among the available categories.

The measure of central tendency is average and there are 3 types of averages

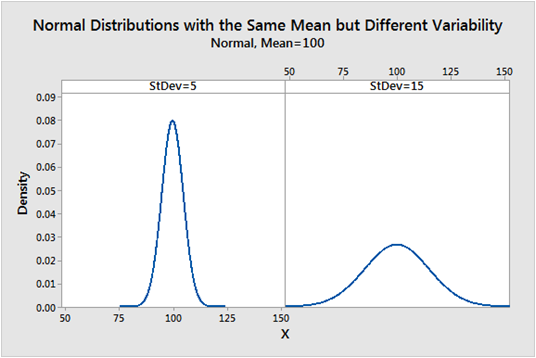
**1)Mean:-**The mean is the average of all numbers and is sometimes called the arithmetic mean

**2)Median:**-The statistical median is the middle number in a sequence of numbers. To find the median, organize each number in order by size; the number in the middle is the median

**3)Mode:-**The mode is the number that occurs most often within a set of numbers.

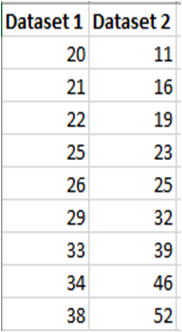
**2)Measures of Dispersion**

It is also called data variability or spread of the data.I.e how much the data is spread out.



The various measures of dispersion are

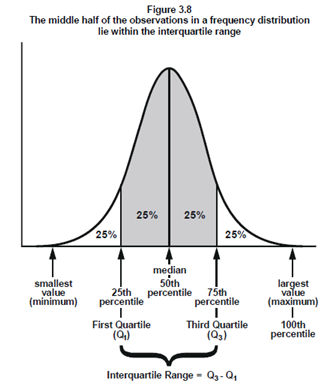
**1)Range:-**The range is the difference between the highest and lowest values within a set of numbers.

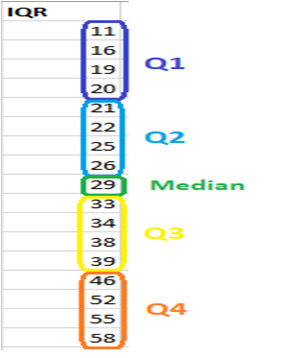


2)**Interquartile Range (IQR)Range:-**

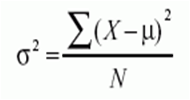
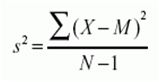
The interquartile range is the middle half of the data.

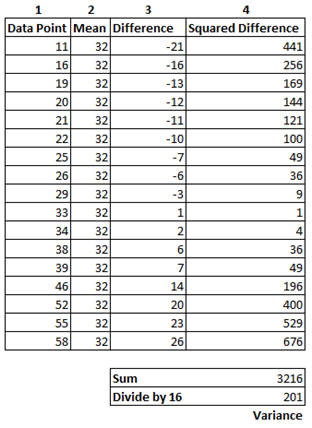
Mathematically the interquartile range includes the 50% of data points that fall between Q1 and Q3.

****

****

**3)Variance:-**Variance is the average squared difference of the values from the mean. Unlike the previous measures of variability, the variance includes all values in the calculation by comparing each value to the mean.

Ex:-

4)**Standard Deviation (σ):-**Standard Deviation (SD) is a measure that is used to quantify the amount of variation or dispersion of a set of data values.

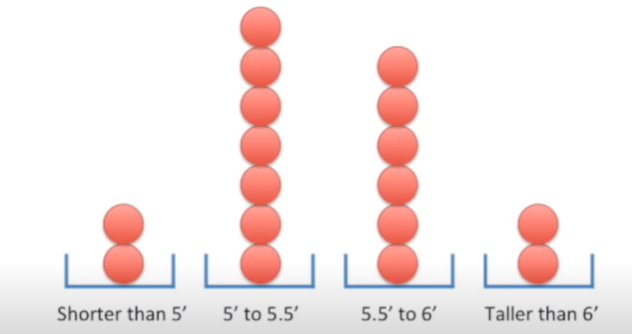
\*\*If the standard deviation is small, the data has little spread(i.e., the majority of points fall very near the mean).

\*\*If standard deviation = 0, there is no spread. This only happens when all data items are the same value.

\*\*The standard deviation is significantly affected by outliers and skewed distributions.

**3)The Shape of Distribution**

The graphical representation of all observations is known as distribution.Consider you are collecting the data of height of people and we have used histogram to represent the data or observation,(as histogram creates bins on the data and plot bars by counting the number of observations falling in the bins.)

****

\*\*Most of the measurements come from people between 5 and 6 feet tall.

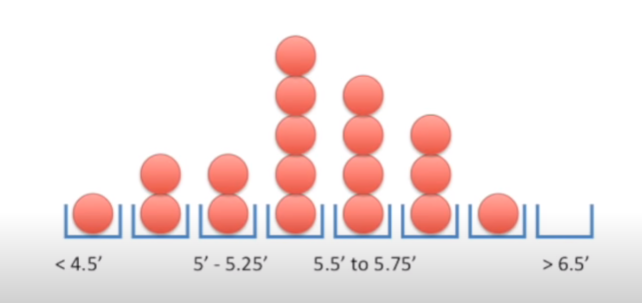
\*\*People smaller than 5 feet tall were relatively rare.

\*\*People taller than 6 feet were relatively rare.

\*\*In other words if you pick a random person,it is a good chance it would be between 5 and 6 feet tall.

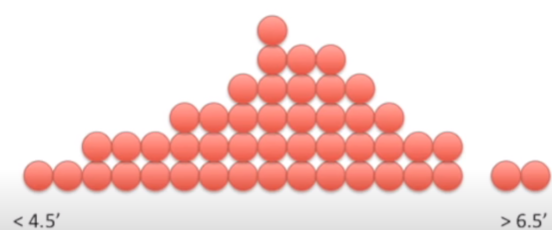
**The histogram gives us sense how likely we will measure someone really tall or really short or closer to average height.**

What if we used smaller bin sizes for our measurements?

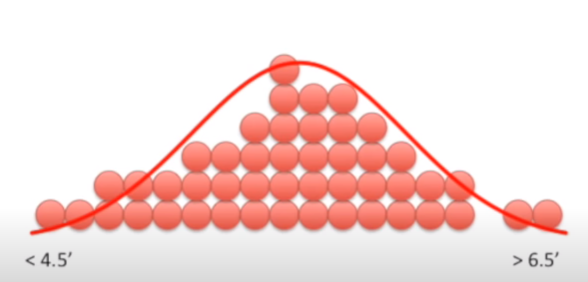


Again most of measurements are between 5 and 6 feet tall,but we can be more precise and say half of the people are between 5.25 to 5.75

By measuring more people and using smaller bins,we get a more accurate and precise estimate of how heights are distributed.



We can use curve to estimate the histogram.



The curve tells us the same thing as the histogram,i.e low probability at the tail ends.We can calculate the probability of measuring someone between 5.021 and 5.317 by using calculus.

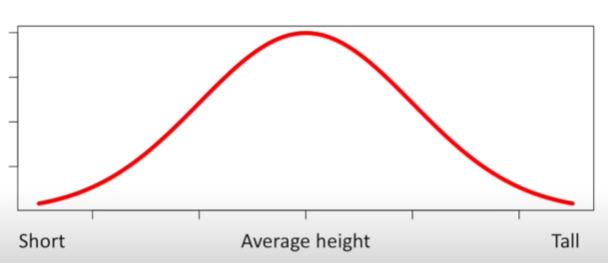
If we don't have enough time and money to get tons of measurements,the approximate curve (based out on mean and std deviation,),is usually good rough.



Histogram and curve are distributions.they show us how the probabilities of measurements are distributed.

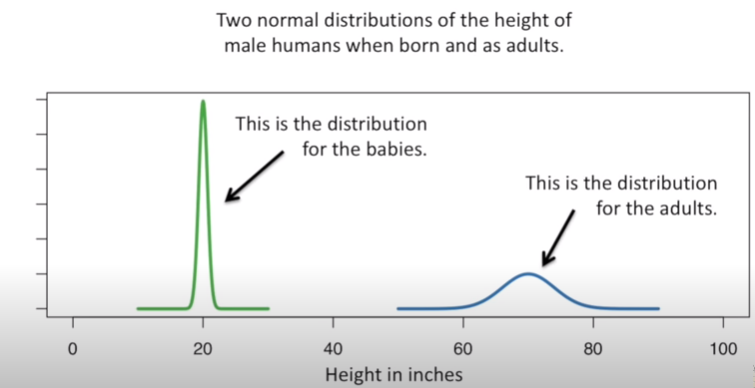
There are a lot many distributions present in statistics but we will focus on Normal distribution only.

**4)The Normal distribution**

****

It is also called a Gaussian distribution or bell curve.The y axis of the distribution represents the probability of observing someone who is really tall,short or average.

Lets see some more type of normal distribution



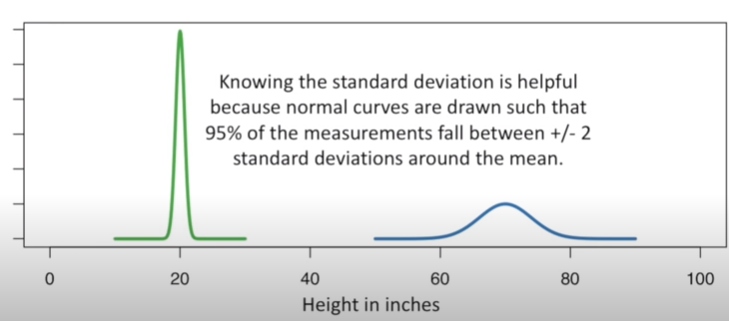
The average baby height is 20inche and adult human heights are 70inches.

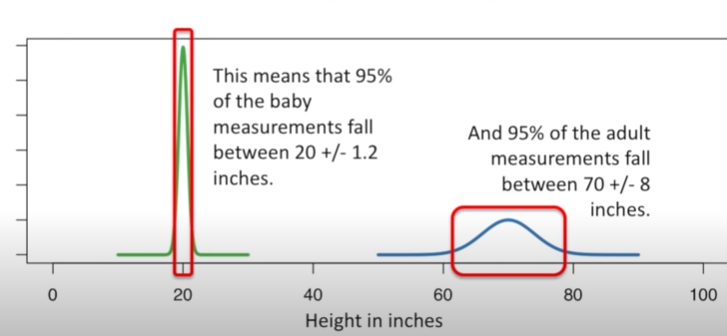
Just by looking at the graph we can tell that there is high probability that a newborn baby will be between 19 to 20inches.In contrast,adults are between 60 to 80 inches tall.

The curve of a baby is more taller than the curve of an adult human.This is because there are many more possibilities for adult height than for babies.

The more option s there are for height,the less likely any specific measurement will be one of them.

Just by looking at the curve we can say that babies have little standard deviation as compared to adult humans.

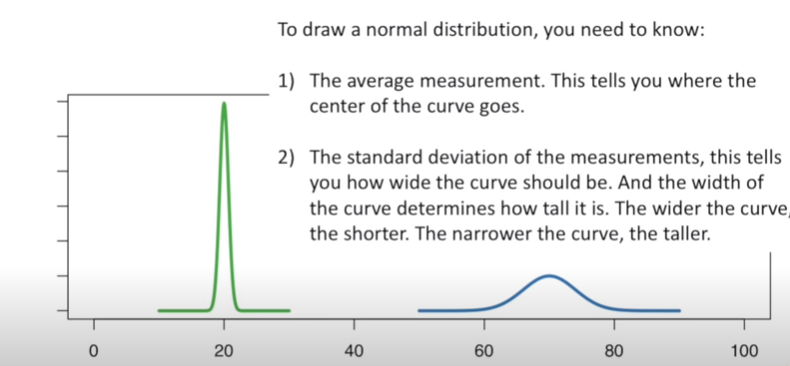




\*\*Normal distributions are always centered on the average value.

\*\*The width of normal distribution is defined by standard deviation.

\*\*95% of measurements falls between +/-2 standard deviations.



There are many more variables which are normally distributed like age,height,salary and many more.

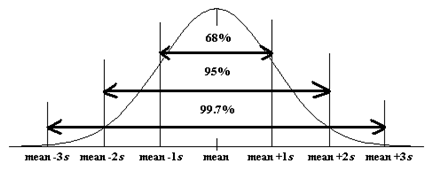
**TASK:-Get all practical examples of normally distributed variables.**

The normal distribution is kind of magical in that we see it a lot in nature,but there are other reasons too to use it widely.

Lets see some more important property of normal distribution

**5)The Empirical Rule**

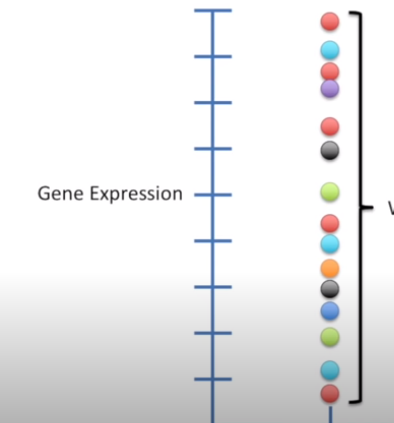
The empirical rule states that for a normal distribution, nearly all of the data will fall within three standard deviations of the mean. The empirical rule can be broken down into three parts:



1. 68% of data falls within the first standard deviation from the mean. (1 Sigma)
2. 95% fall within two standard deviations. (2 Sigma)
3. 99.7% fall within three standard deviations. (3 Sigma)

**6)Quantiles and Percentiles**

Consider we have collected gene expression(information stored in DNA),as shown below

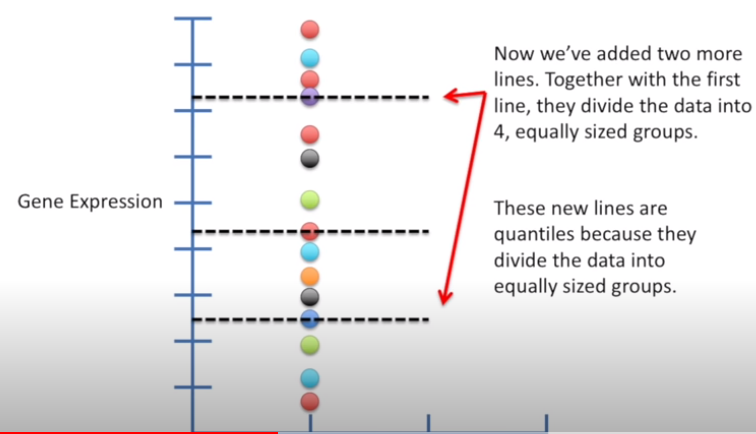


We measured in total 15 genes,the median is let say x,so we can say 50%

Of data is less than median and 50% of data is more than median.

So we can say a quantile is datapoint that divides entire dataset into 2 equal parts.Technically median is quantile because it splits the data into groups that contains same number of datapoints.

Since here median is x and there are equal numbers of datapoints above and below x,so median is called as 0.5 or 50 quantile.



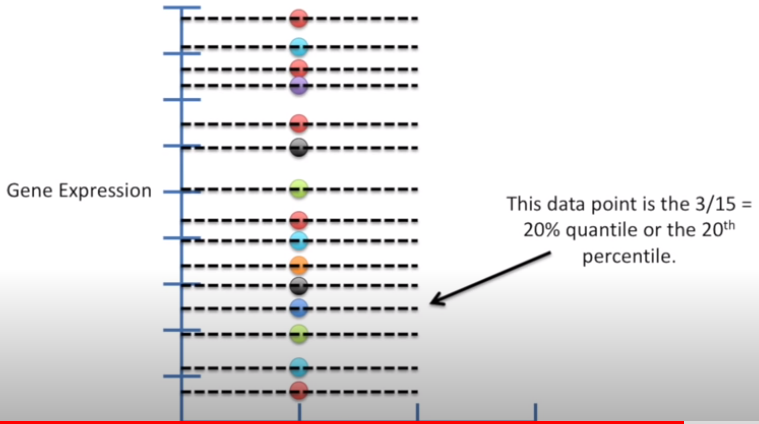
The first line from the bottom is called as 0.25 or 25% because a quarter of points are less than it.

The last one from top is called as 0.75 or 75% because 3 quarters points are less than it.

In general,quantiles are just the lines that divides the data into equally sized groups.

Now lets understand percentiles,and technically they are just quantiles that divides the data into 100 equally sized groups.

Even though this dataset is not large enough to be divided into 100 groups we still call the median is 50th percentile.Often the term percentile and quantile are used when we divide data into its own group.



**7)Skewed Distribution**

#### *What is Skewness*

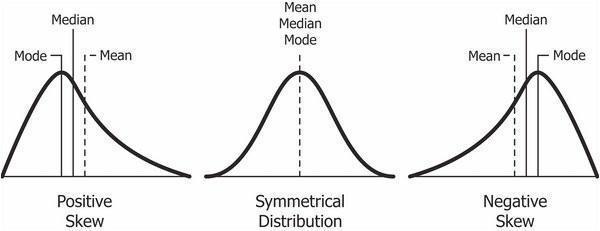
Skewness is asymmetry in a statistical distribution, in which the curve appears distorted or skewed either to the left or to the right. Skewness can be quantified to define the extent to which a distribution differs from a normal distribution.

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.

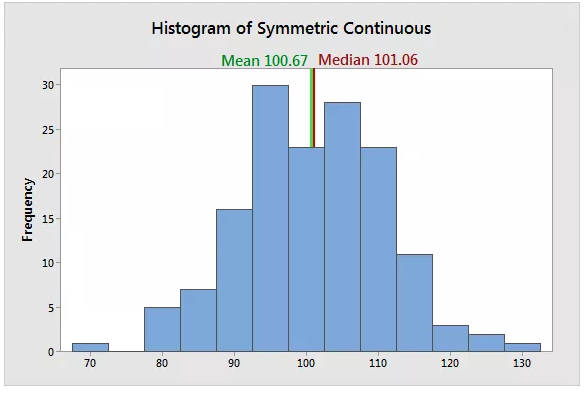
In a perfect normal distribution, the tails on either side of the curve are exact mirror images of each other.

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**.

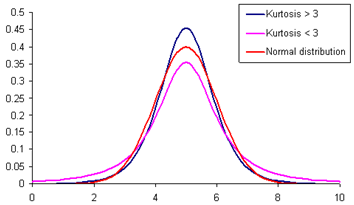


\*\*In a symmetric distribution, the mean and median both find the center accurately. They are approximately equal.



**8)Kurtosis**

In probability theory and statistics, kurtosis is a measure of the "tailedness" of the probability distribution of a real-valued random variable. In a similar way to the concept of skewness, kurtosis is a descriptor of the shape of a probability distribution and, just as for skewness, there are different ways of quantifying it for a theoretical distribution and corresponding ways of estimating it from a sample from a population.



**How much Skewness and Kurtosis**

* 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 1or less than -1, the data is highly skewed.
* A standard normal distribution has kurtosis of 3 and is recognized as mesokurtic. An increased kurtosis (>3) can be visualized as a thin “bell” with a high peak whereas a decreased kurtosis corresponds to a broadening of the peak and “thickening” of the tails.

**Which is Best—the Mean, Median, or Mode?**

* When you have a symmetrical distribution for continuous data, the mean, median, and mode are equal. In this case, analysts tend to 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 ,categorical,count(discrete), the median or mode is usually the best choice. For categorical data, you have to use the mode.

**9)Outlier**

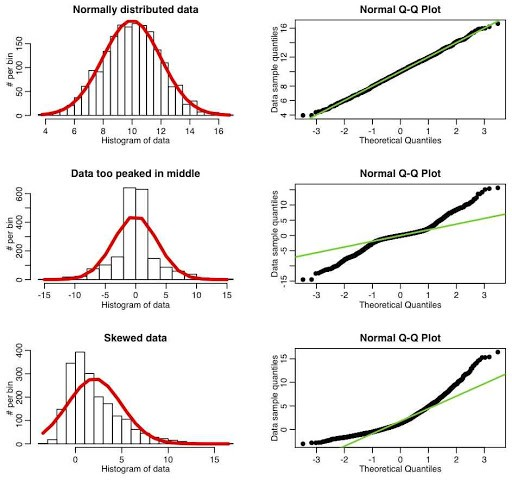
An outlier is an observation point that is distant from other observations. An outlier may be due to variability in the measurement or it may indicate experimental error; the latter are sometimes excluded from the data set.

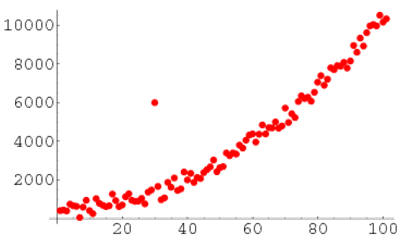
**10)Q-Q Plot**

In Statistics, Q-Q(quantile-quantile) plots play a very vital role to graphically analyze and compare two probability distributions by plotting their quantiles against each other. If the two distributions which we are comparing are exactly equal then the points on the Q-Q plot will perfectly lie on a straight line y = x.

Normally distributed, but why?

Q-Q plots are used to find the type of distribution for a random variable whether it be a Gaussian Distribution, Uniform Distribution, Exponential Distribution or even Pareto Distribution, etc. You can tell the type of distribution using the power of the Q-Q plot just by looking at the plot. In general, we are talking about Normal distributions only because we have a very beautiful concept of 68–95–99.7 rule which perfectly fits into the normal distribution So we know how much of the data lies in the range of first standard deviation, second standard deviation and third standard deviation from the mean. So knowing if a distribution is Normal opens up new doors for us to experiment with the data easily. Secondly, Normal Distributions occur very frequently in most of the natural events which have a vast scope.

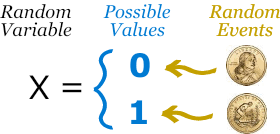


****

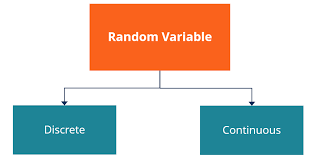
**https://tribe.datamites.com/posts/outliers**

**10)Random Variable**

Variable are placeholder where you can store anything.It can number,or string,sentences .Consider we want to store age in some variable so can define like this x=25, or you want to store a x=’Hey’.Now when we start collecting data or when we have data set,in that we have variables,now when different values of variable let's say age has been recorded,such variables are known as random variable.



**A *random variable*, usually written *X*, is a variable whose possible values are numerical outcomes of a random phenomenon. There are two types of random variables, *discrete* and *continuous*.**

****

**11) Central Limit Theorem**

The central limit theorem states that the distribution of sample means approximates a normal distribution as the sample size gets larger (assuming that all samples are identical in size), regardless of population distribution shape.

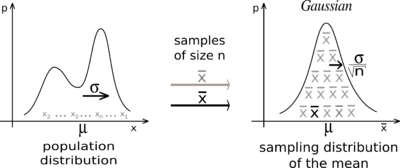
But what are practical implications of knowing that means are normally distributed?

When we do the experiment we don't know what distribution our data comes from,to this The central limit theorem says “who cares” as we know that means from these distributions will be normally distributed.So we can use it to design confidence intervals,(It tells you how confident you can be that the results from a poll or survey reflect what you would expect to find if it were possible to survey the entire population)

It helps in comparing samples 2 or more than 2 samples based on statistical tests which used means to compare the samples.

When collecting means of the samples from any distribution,the no of samples taken for calculating the mean should be greater or equal to 30.

**CLT in one sentence "Even if I'm not normal, the average is normal"**



**12)Standard Normal Distribution**

The standard normal distribution is a special case of the normal distribution. It is the distribution that occurs when a normal random variable has a mean of zero and a standard deviation of one.

The normal random variable of a standard normal distribution is called a standard score or a z score.

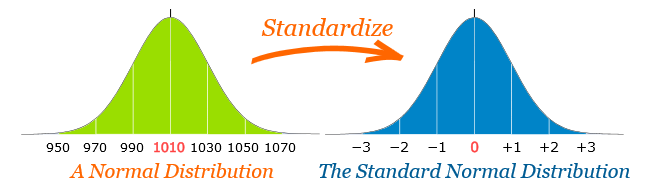
Every normal random variable X can be transformed into a z score via the following equation:

**z = (X - μ) / σ**

where, X is a normal random variable,

μ is the mean,

and σ is the standard deviation.

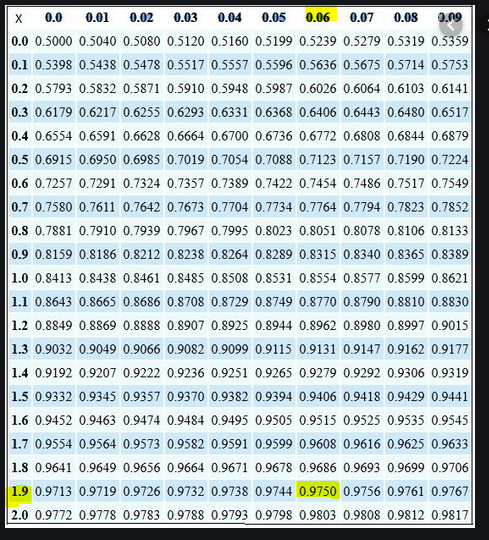


**z score**

*z*-score is a measure of position that indicates the number of standard deviations a data value lies from the mean. z-scores may be positive or negative, with a positive value indicating the score is above the mean and a negative score indicating it is below the mean.

z score is very powerful tool to find the probability distribution using the z-score table.

We do not need the normal distribution calculator and will simply use z-score table.

let’s see the application of z-score table and see how to use it:

We will use the same example from above where we used the Normal Distribution calculator, let’s try to calculate the same using z-score table and compare the results.

Q) The Light Bulb Company has found that an average light bulb lasts 900 hours with a standard deviation of 80 hours. Assuming that bulb life is normally distributed. What is the probability that a randomly selected light bulb will burn out in 1000 hours or less?

Answer: let’s convert our data in standard normal form.

Mean = 900

Std. deviation = 80

x(a) = 1000

standardized x(a) = (1000-900)/80 = 1.25

z-score = 1.25 = 1.2 + 0.05 (in the table we will match the value corresponding to 1.2 and 0.05)

Let’s use the z-score table for this:

We found that Probability coming as 89.44 %.

So it tells that the probability of that bulb to burn in 1000hrs 89.44%

Thus, if we standardize a normal distribution, z-score becomes a very important tool in helping finding the probability distribution.

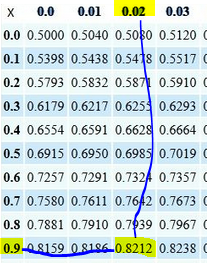
**Q) Ravi scored 980 in a Physics Olympiad. The mean test score was 870 with a standard deviation of 120. How many students scored more than Ravi? (Assume that test scores are normally distributed.)**

**Answer**: Let’s standardize the test score

Mean = 870

St. deviation= 120

z-score = (980-870)/120 = 0.917 (we will approximate it to 0.92 )



So, P(x<=980) = 0.8212

We need to find the probability of scoring more than 980,

P(x>980) = 1 – 0.8212 = 0.1788

Thus, we can estimate 17.88% students scored more than Ravi in the test.

**13)Probability Density Function**

Probability density is the relationship between observations and their probability.

A random variable x has a probability distribution p(x).

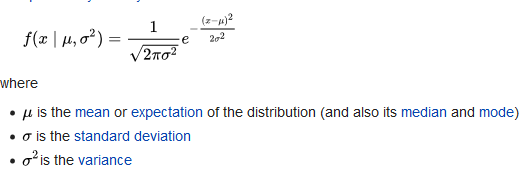
The relationship between the outcomes of a random variable and its probability is referred to as the probability density, or simply the “density.”

Some outcomes of a random variable will have low probability density and other outcomes will have a high probability density.

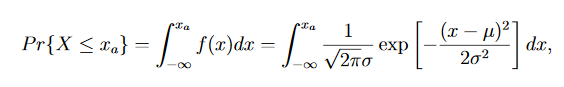
The overall shape of the probability density is referred to as a probability distribution, and the calculation of probabilities for specific outcomes of a random variable is performed by a probability density function, or PDF for short.

It is useful to know the probability density function for a sample of data in order to know whether a given observation is unlikely, or so unlikely as to be considered an outlier or anomaly and whether it should be removed. It is also helpful in order to choose appropriate learning methods that require input data to have a specific probability distribution.

**The probability density function for Normal distribution is given as:**



Given a variable x(a), the probability of the random variable X, which follows a normal distribution, is less than or equal to x(a) is given as:





The Normal Distribution has following properties:

* 1. mean = median = mode
  2. symmetry about the center
  3. 50% of values less than the mean and 50% greater than the mean
  4. The probability that X is greater than ‘a’ is equal to the area under the normal curve as shown by the non-shaded area in the figure below.
  5. The probability that X is less than ‘a’ is equal to the area under the normal curve as shown by the shaded area in the figure below.

**14)Cumulative Density Function**

This function is again related to the probabilities of the random variable equaling specific values. It provides a shortcut for calculating many probabilities at once. We integrate the **pdf** function to get the cumulative probability.

15)**Standard Error**

The standard error (SE) of a statistic 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

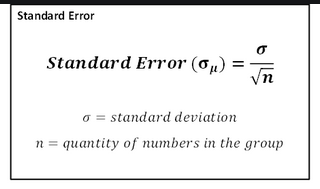
When a population is sampled, the mean, or average, is generally calculated. **The standard error can include the variation between the calculated mean of the population and one which is considered known, or accepted as accurate. This helps compensate for any incidental inaccuracies related to the gathering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

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 most often measured as the standard error, accounting for the differences between the means across the datasets.

The more data points involved in the calculations of the mean, 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 level of standard deviation. 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.

Standard Error is given by the following formula:



Here σ is the standard deviation of the population, whereas σ(u) is the standard deviation of the sample.

We can see that as the size of our sample increases the Standard error decrease.

Now, that we know the Standard error, let’s rephrase our Central Limit theorem as:

**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.**

So, when we transform our sample data, we will use following formula for the z-score:

z = (X - μ) / (σ/√n)

where, X is the sample mean,

μ is the mean of the population,

and σ is the standard deviation of the population.

Let’s see an example based on the above explaination.

**Q) Let X be a random variable with μ= 10 and σ= 4. A sample of size 100 is taken from this population. Find the probability that the sample mean of these 100 observations is less than 9.**

**Ans:** population mean = 10 population std. deviation = 4 sample size(n) = 100

Sample mean = 9

z= (9-10)/ (4/ (100) ^0.5) = -2.5

We will use the z-score table and find the value to be 0.0062

P(X<9) = 0.0062

**Q) A large freight elevator can transport a maximum of 9800 pounds. Suppose a load of cargo containing 49 boxes must be transported via the elevator. Experience has shown that the weight of boxes of this type of cargo follows a distribution with mean= 205 pounds and standard deviation = 15 pounds. Based on this information, what is the probability that all 49 boxes can be safely loaded onto the freight elevator and transported?**

**Ans:** For all the boxes to be loaded the total weight must be at most 9800.

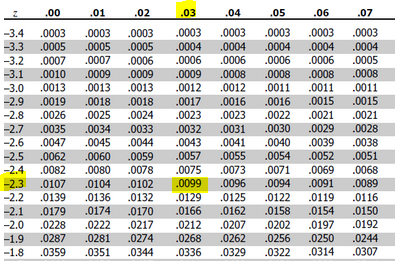
So, the sample mean should be = 9800/49 = 200 sample size(n) = 49

Population mean = 205

Std. deviation = 15

z-score = (200-205)/(15/(49)^0.5) = -2.33

using z-score table:

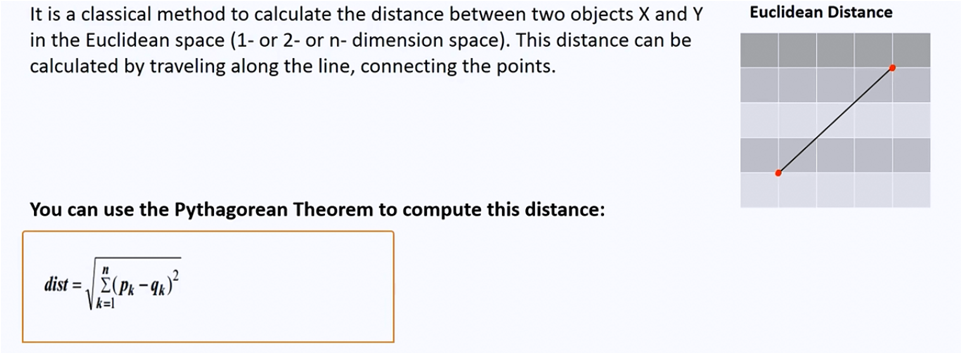


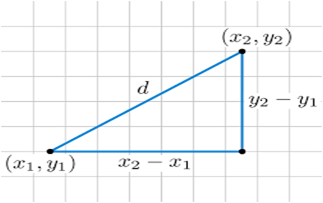
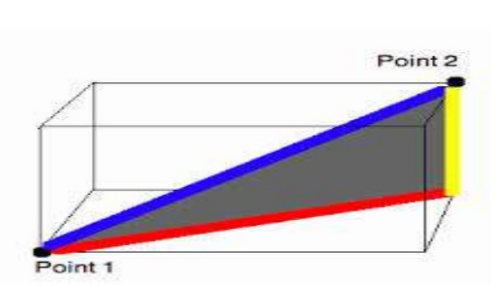
P(X<200) = 0.0099

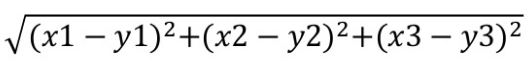
**16)Measure of distance**

Statistical Distances are used to quantify the distance between two distributions and are extremely useful in ML observability

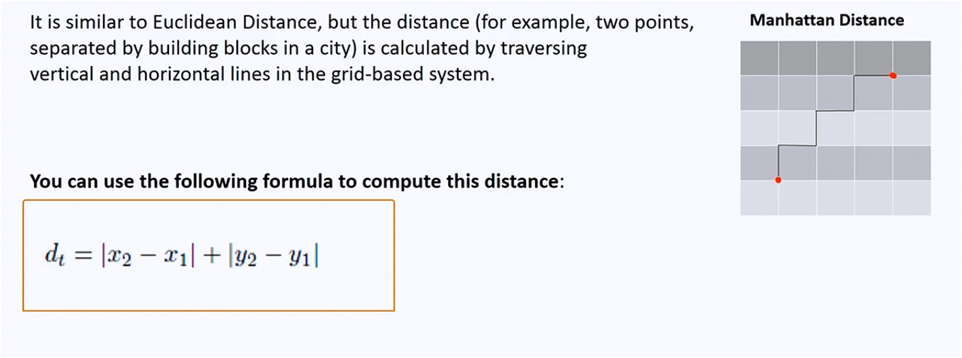
**1)Euclidean Distance**

****

** **

****

**2)Manhattan Distance**

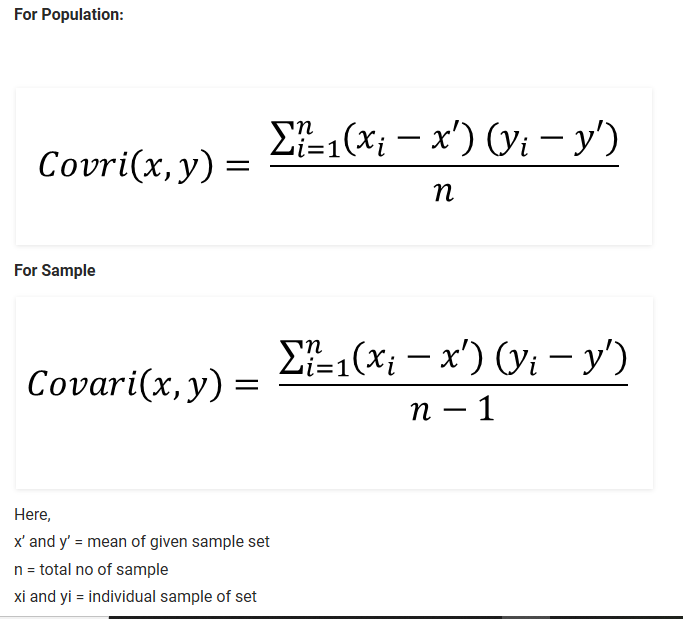
****

**17)Covariance and Correlation**

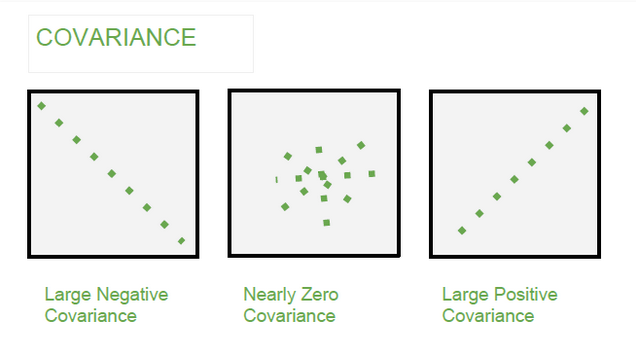
### **Covariance**

It is method to find the variance between two variables.

1. It is the relationship between a pair of random variables where change in one variable causes change in another variable.
2. 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.
3. It is used for the linear relationship between variables.
4. It gives the direction of relationship between variables.
5. It has dimensions.



**Covariance Relationship**

****

### **Correlation**

\* 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.

\* In this variable are indirectly related to each other.

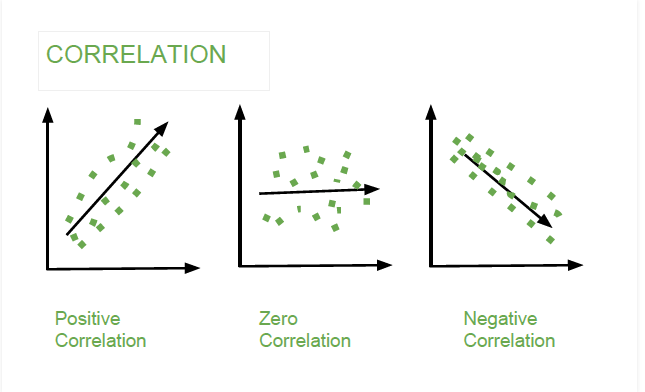
\* It gives the direction and strength of relationship between variables.

\* It is the scaled version of Covariance.

\* It is dimensionless.



**Correlation Relationship**

****

**Positive Correlation**

When the values of variables deviate in the same direction i.e. when value of one variable increases(decreases) then value of other variable also increases(decreases).

Examples:

1. Height and weight of persons
2. Amount of rainfall and crops yield
3. Income and Expenditure of Households
4. speed of a wind turbine, the amount of electricity that is generated
5. The more years of education you complete, the higher your earning potential will be
6. As the temperature goes up, ice cream sales also go up
7. The more it rains, the more sales for umbrellas go up

**Negative Correlation**

When the values of variables deviate in the opposite direction i.e. when value of one variable increases(decreases) then value of other variable also decreases(increases).

Examples:

1. Price and demand of goods
2. Poverty and literacy
3. Sine function and cosine function
4. If a train increases speed, the length of time to get to the final point decreases
5. The more one works out at the gym, the less body fat one may have
6. As the temperature decreases, sale of heaters increases

**Zero Correlation**

When two variables are independent of each other, they will have a zero correlation.

**Note: - When data is scaled covariance and correlation will give the same value. Also, correlation and Causality are not the same thing.**

**Module 4 Comparing Samples and Analyzing Relationships.(Hypothesis Testing)**

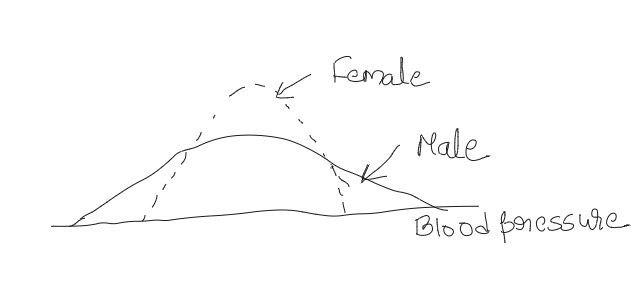
In the last module we have seen samples and done the inferences.We can consider another highly important area of statistical inference:looking at the 2 different samples and asking whether or not imply a real difference in populations.For example

\*Does the treatment with new drug help more patients than the standard treatment?

\*Which of these four methods is the most efficient way of teaching machine learning?

FROM THE SAME OR DIFFERENT POPULATIONS?

Let’s consider we want to measure blood pressure of 2 random samples of students,50 male and 50 females.What can be these 2 random sample tells us about the difference between the blood pressures of male and female student in general.?Are the samples so similar that we’d be justified in lumping them together and saying they are from same population; or are they so different that they signify different populations?

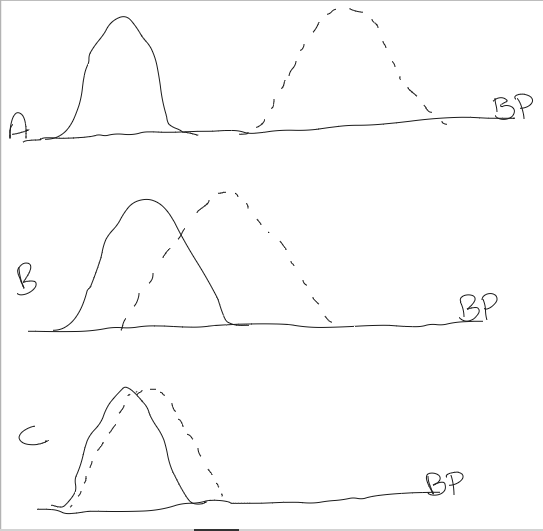


If this is the distribution of the samples which we have collected,what would think??

Would we be justified in thinking that both samples came from the same population?On the basis of mean and standard deviation of both samples of male and female are if same then we can say both belong to the same population.

Although means are much same in the above 2 samples ,but the standard deviation are very different.

COnclusion:-



Let’s try to do more difficult comparison.In each pair standard deviations are same but mean differ between 2 samples in each pair.In which case A<B<C would you feel

1)more confident

2)least confident

That the pair of samples came from different population.?

In each of the 3 samples,the means are different .In (A),they are very different,in (B)they are much less different and in (C) they differ scarcely.So we can say C is the sample that came from populations with different means.

So when we have to compare samples that time significance testing comes into picture.

Significance testing is the method where we ask whether the difference between samples is big enough to specify a real difference between the populations.

Now to identify two or more samples are different we will understand hypothesis testing.Let’s first understand what does hypothesis means.

**What is Hypothesis?**

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

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

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.

Like, if we make a statement that “Dhoni is the best Indian Captain ever.” This is an assumption that we are making based on the average wins and looses 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. Motor cycle company claiming that a certain model gives an average mileage of 100Km per liter, this is a case of simple hypothesis.

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.

It is denoted by H0.

**Alternate Hypothesis**

The alternative hypothesis complements the Null hypothesis. It is opposite of the null hypothesis such that both Alternate and null hypothesis together cover all the possible values of the population parameter.

It is denoted by H1.

Let’s understand this with an example:

A soap company claims that it’s product kills on an average 99% of the germs. To test the claim of this company we will formulate the null and alternate hypothesis.

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.

**Task:-Create H0 and H1 for the below statement**

**1)A researcher thinks that if knee surgery patients go to physical therapy twice a week (instead of 3 times), their recovery period will be longer. Average recovery times for knee surgery patients is 8.2 weeks.**

**2)Is it true that vitamin C has the ability to cure or prevent the common cold?**

**3)Young boys are prone to more behavioral problems than young girls.**

**Hypothesis Testing**

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.

**One Tailed and Two Tailed Tests**

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

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

e.g. if H0: mean= 100 H1: mean not equal to 100

here according to H1, mean can be greater than or less than 100. This is an example of Two tailed test

Similarly, if H0: mean>=100 then H1: mean< 100

Here, mean is less than 100, it is called 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 state websites. You have budget of Rs. 15000/ month. You cannot spend more than that. The list of apartments you have made have price ranging from 7000/month to 30,000/month.

You select a random apartment from the list and assume below hypothesis:

H0: You will rent the apartment.

H1: You won’t rent the apartment.

Now, since your budget is 15000, you have to reject all the apartments above that price.

Here all the Prices greater than 15000 becomes 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.

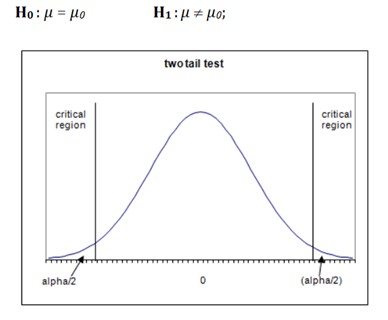
The critical region lies in one tail or two tails on the probability distribution curve according to the alternative hypothesis. Critical region is a pre-defined area corresponding to a cut off value in the probability distribution curve. It is denoted by α.

Critical values are values separating the values that support or reject the null hypothesis and are calculated on the basis of alpha.

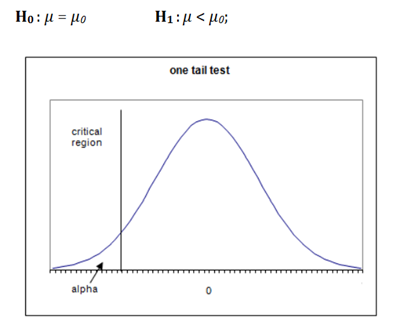
We will see more examples later on and it will be clear how we choose α.

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

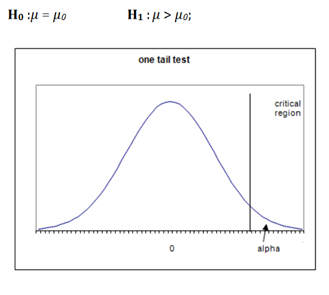
Case 1) This is double tailed test.



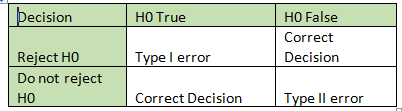
Case 2) This scenario is also called Left-tailed test.



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



**Type I and Type II Error**

****

**A false positive (type I error) — when you reject a true null hypothesis.**

**A false negative (type II error) — when you accept a false null hypothesis.**

The probability of committing 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 Type II error (False negative) is equal to the beta β and is called ‘power of the test’.

β = P [not rejecting H0 when h1 is true]

Example:

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 Jury released 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.

**Steps involved in Hypothesis testing**

1)Setup the null hypothesis and the alternate hypothesis.

2)Decide a level of significance i.e. alpha = 5% or 1%

3)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)

4)Calculate the test statistics (z-score, t-score etc.) using the respective formula of test chosen

5)Obtain the critical value for in the sampling distribution to construct the rejection region of size alpha using z-table, t-table, chi table etc.

6)Compare the test statistics with the critical value and locate the position of the calculated test statistics i.e. is it in rejection region or non-rejection region.

7)I)If the critical value lies in the rejection region, we will reject the hypothesis i.e. sample data provides sufficient evidence against the null hypothesis and there is significant difference between hypothesized value and observed value of the parameter.

II) 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 hypothesized value and observed value of the parameter is due to fluctuation of the sample.

**p-value**

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 statistics 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 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” in play.

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

That’s why many tests now a days gives 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 p-value is greater than alpha, we do not reject the null hypothesis.

If p-value is smaller than alpha, we reject the null hypothesis.

**Confidence Intervals**

A confidence interval, in statistics, refers to the probability that a population parameter will fall between two set values. Confidence intervals measure the degree of uncertainty or certainty in a sampling method. A confidence interval can take any number of probabilities, with the most common being a 95% or 99% confidence level.

Calculating a Confidence Interval (Theory)

Suppose a group of researchers is studying the heights of high school basketball players. The researchers take a random sample from the population and establish a mean height of 74 inches. The mean of 74 inches is a point estimate of the population mean. A point estimate by itself is of limited usefulness because it does not reveal the uncertainty associated with the estimate; you do not have a good sense of how far away this 74-inch sample mean might be from the population mean. What's missing is the degree of uncertainty in this single sample.

Confidence intervals provide more information than point estimates. By establishing a 95% confidence interval using the sample's mean and standard deviation, and assuming a normal distribution as represented by the bell curve, the researchers arrive at an upper and lower bound that contains the true mean 95% of the time. Assume the interval is between 72 inches and 76 inches. If the researchers take 100 random samples from the population of high school basketball players as a whole, the mean should fall between 72 and 76 inches in 95 of those samples.

If the researchers want even greater confidence, they can expand the interval to 99% confidence. Doing so invariably creates a broader range, as it makes room for a greater number of sample means. If they establish the 99% confidence interval as being between 70 inches and 78 inches, they can expect 99 of 100 samples evaluated to contain a mean value between these numbers. A 90% confidence level means that we would expect 90% of the interval estimates to include the population parameter. Likewise, a 99% confidence level means that 95% of the intervals would include the parameter.

The Confidence Interval is based on Mean and Standard Deviation and is given as:

For n>30

Confidence interval = X ± (z \* s/√n)

where z critical value is derived from the z score table based on the confidence level.

X is the sample mean.

s is sample standard deviation.

n is the sample size

We obtain these values from the z-score table only, but since the confidence levels are most of the times fixed as the above values, so we can use this table.

For n<30

Confidence interval = X ± (t \* s/√n)

where t critical value is derived from the t score table based on the confidence level.

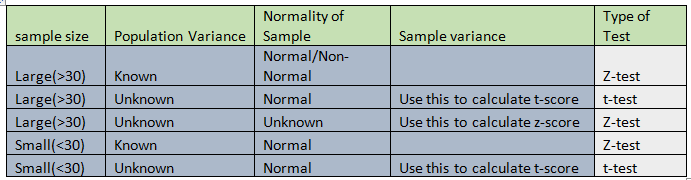
X is the sample mean.

s is sample standard deviation.

n is the sample size.

We will see how to create confidence intervals in the examples to follow.

**Now that we have got all the theory behind Hypothesis testing, let’s see different types of tests that are used for testing. We have already seen examples on finding z-score and t-score, we will see how they are used in the testing scenario.**

****

**Hypothesis Testing for Large Size Samples**

**Thumb rule: A sample of size greater than 30 is considered a large sample and as per central limit theorem we will assume that all sampling distributions follows a normal distribution.**

We are familiar with the steps of hypothesis testing as shown earlier. We also know, from the above table, when to use which type of test.

Let’s start with few practical examples to help our understanding more.

Note: We have learned in previous section how to use the z-score table to calculate probabilities, in this section we have some standard Significance level for which we need to find the critical value(z-score). So instead of going through the whole table, we will just use the below standardized critical value table for calculation purposes.

Q) A manufacturer of printer cartridge clams that a certain cartridge manufactured by him has a mean printing capacity of at least 500 pages. A wholesale purchaser selects a sample of 100 printers and tests them. The mean printing capacity of the sample came out to be 490 pages with a standard deviation of 30 printing pages.

Should the purchaser reject the claim of the manufacturer at a significance level of 5%?

Ans. population mean = 500

Sample mean = 490

Sample standard deviation = 30

Significance level(alpha) = 5% = 0.05

Sample size = 100

H0: Mean printing capacity >=500

H1: Mean printing capacity < 500

We can clearly see it is one tailed test (left tail).

Here, the sample is large with an unknown population variance. Since, we don’t know about the normality of the data, we will use the Z-test (from the table above).

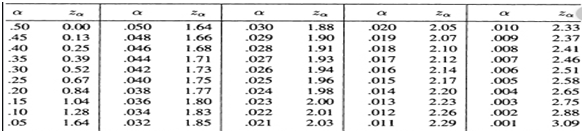
We will use the sample variance to calculate the critical value.

Standard error (SE) = Sample standard deviation/(sample size)\*0.5

= 30 / root(100)= 3

Z(test) = (Sample mean - population mean)/ (SE)

= (490-500)/3 = -3.33



Let’s find out the critical value at 5% significance level using the above Critical value table.

Z(5%) = - 1.645 (since it is left tailed test).(Given in problem statement)

We can clearly see that Z(test) < Z(5%), that means our test value lies in the rejection region.

Thus, we can reject the null hypothesis i.e. the manufacturer’s claim at 5% significance level.

Using p-value to test the above hypothesis:

p-value = P[Z<=-3.33] (we know p(-x) = 1 -p(x) also, remember that the p(x) represents the cumulative probability from 0 to x)

let’s use z-table to find the p-value:

p-value = 1 – 0.9996 = 0.0004

Here, p-value is less than the significance level of 5%. So, we are right to reject the null hypothesis.

Q) A company used a specific brand of Tube lights in the past which has an average life of 1000 hours. A new brand has approached the company with new Tube lights with same power at a lower price. A sample of 120 light bulbs were taken for testing which yielded an average of 1010 hours with standard deviation of 90 hours. Should the company give the contract to this new company at a 1% significance level.

Also, find the confidence interval.

Ans. Population mean = 1000

Sample mean = 1010

Significance level = 1% = 0.01

Sample size = 120

Sample standard deviation = 90

H0: average life of tube lights >= 1000

H1: average life of tube lights < 1000

Here, the sample is large with an unknown population variance. Since, we don’t know about the normality of the data, we will use the Z-test (from the table above).

Standard error (SE) = Sample standard deviation/(sample size)\*\*0.5

= 90 / (120) \*0.5 = 8.22

Z(test) = (Sample mean - population mean)/ (SE)

= (1010-1000)/8.22 = 1.22

Let’s find out the critical value at 1% significance level using the above Critical value table.

Z (0.01%) = -2.33(since it is left tailed test).

We can clearly see that Z(test) >Z (0.01%), that means our test value doesn’t lie in the rejection region.

Thus, we cannot reject the null hypothesis i.e. the company can give the contract at 1% significance level.

Using p-value to test the above hypothesis:

p-value = P[T<1.22]

p-value = 0.88

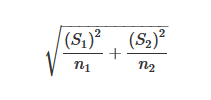
Here, p-value is greater than the significance level of 1%. So, we do not reject the null hypothesis.

**Comparing two population samples mean using Z test**

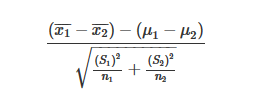
The comparison of two population means is very common. A difference between the two samples depends on both the means and the standard deviations. Very different means can occur by chance if there is great variation among the individual samples. In order to account for the variation, we take the difference of the sample means, X1(mean) - X2(mean) , and divide by the standard error (shown below) in order to standardize the difference.

Because we do not know the population standard deviations, we estimate them using the two sample standard deviations from our independent samples. For the hypothesis test, we calculate the estimated standard deviation i.e. standard error.

The standard error (SE) is:



Z is given as :



In this comparison case, our null assumption is that µ(1) = µ(2)

So, Z becomes = X1(mean)- X2(mean)/ (SE)

Q) In two samples of men from two different states A and B , the height of 1000 men and 2000 men respectively are 76.5 and 77 inches. If population standard deviation for both states is the same and is 7 inches, can we assume that mean heights of both states can be regarde same at 5% level of significance?

Ans. n1 = 1000

n2 = 2000

X1(mean) = 76.5

X2(mean) = 77

S1=S2= 7

Let’s µ(1) = µ(2) be the mean heights of men from states A and B

H0: µ(1) = µ(2)

H1: µ(1) is not equal to µ(2)

Standard error(SE) = [((S1)^2/n1 )+((S2)^2/n2)]^0.5 = 0.27

Z(test) = X1(mean)- X2(mean)/ (SE) = (76.5-77)/0.27 = -1.85

Since, it is a two tailed test, we need to find critical value for 2.5% on each tail.

Z(2.5%) = 1.96 and Z(-2.5%) = -1.96

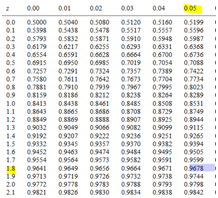
We can clearly see, Z(-2.5%) < Z(test) <Z(2.5%)

Thus, we cannot reject the null hypothesis.

Using p-value

p-value = 2\* P[Z>=|-1.85|] = 2 \* P[Z>=-1.85]

p-value = 2 \* (1- 0.9678) (since we want z> 1.85) = 0.0644



We can clearly see, p-value is greater than 0.05% ,thus we cannot reject the null hypothesis.

**Hypothesis Testing forSmall Size Samples**

In real world scenarios, large sample sizes are not possible most of times because of the limited resources such as money. We generally do hypothesis testing based on small samples, only assumption being the normality of the sample data.

We will see how to use t- tests in this section and how to use the t-score table (continued from the topic of student t’s distribution).

All the steps involved are similar to the z-test, only we will calculate t-score instead of z-score.

Let’s start with an example:

Q) A tyre manufacturer claims that the average life of a particular category of its tyre is 18000km when used under normal driving conditions. A random sample of 16 tyres was tested. The mean and SD of life of the tyres in the sample were 20000 km and 6000 km respectively.

Assuming that the life of the tyres is normally distributed, test the claim of the manufacture at 1% level of significance.

Construct the confidence interval also.

Ans: population mean = 18000 km

Sample mean = 20000 km

Standard deviation = 6000 km

Sample size = 16

H0: population mean = 18000km

H1: population mean is not equal to 18000km (It will be a two tailed test.)

Since sample size is small, population variance is unknown and the sample is normally distributed, we will used t-test for this.

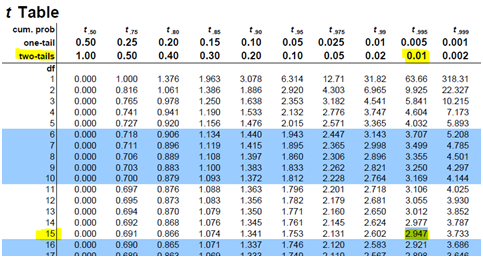
Standard error = [6000/(16)^0.5] = 1500

t-score(test) = (20000 - 18000)/1500 = 1.33

Let’s find out the critical t- value, for significance level 1% (two tailed) and degree of freedom = 16-1 = 15

It is possible times where we can vary a sample value in sample.

For degree of freedom:-https://www.statisticshowto.com/probability-and-statistics/hypothesis-testing/degrees-of-freedom/



t(0.005) = 2.947 and t(-0.005) = -2.947

We can see that,t (- 0.005) < t-score(test) = 1.33 <t (0.005)

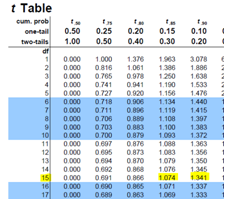
So, the value lies in non-rejection region and we cannot reject our null hypothesis.

**Using the p-value**

p-value = P[t>|1.33|]

degree of freedom = 15

let’s see the p-value from the table for the above values:



from the table we can see: 0.20 < p < 0.30

Here, p > significance level(1%), thus we cannot reject the null hypothesis.

**Confidence interval** = [20000 – 2.47\*1500 , 20000 +2.47\*1500]

= [ 16295, 23705]

## **Comparing two population samples mean using t test**

Just like the case we saw with z-test, t-test is actually more suitable for comparison of two populations samples because in practice population standard deviations for both populations are not always known.

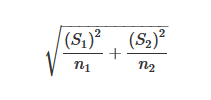
We assume a normal distribution of samples and though the population standard deviations are unknown, we assume them to be equal.

Also, samples are independent to each other.

Let’s assume two independent samples with size n1 and n2:

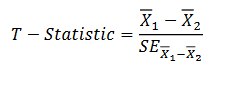
**Degree of freedom** = n1 + n2 -2

**Standard Error(SE):**



**Variance(Sample)** = (∑[X-X(mean)]^2 + ∑[Y-Y(mean)]^2))/(n1 + n2 -2)

Test statistic t in this case is given as:



Q) The means of two random samples of sizes 10 and 8 from two normal popultaion is 210.40 and 208.92. The sum of sqaures od deviation from their means is 26.94 and 24.50 respectively.Assuming population with equal variances, can we consider the normal populatiojns have equal mean?(Significance level =5%)

Ans.

n1 =10 , n2= 8 , X(mean) = 210.40 , Y(mean) = 208.92

std. Deviation(sample) =[ (26.94 + 24.50)/(10 + 8 - 2)]^0.5 = 1.79

H0: Population means are equal

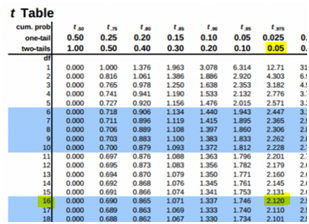
H1: Population means are not equal (two tailed test)

Standard errror = 1.79 \* (1/10 + 1/8)^0.5 = 0.84

t(test) = X(mean)-Y(mean)/0.84 = 1.48/.84 = 1.76

Degree of freedom = 10 +8 -2 = 16

Let’s look for critical value in the t-table for significance 5%(two tailed) and d.o.f 16:



t(0.005) = 2.120 and t(-0.005) = -2.120

We can see that,t (- 0.005) < t-score(test) = 1.76 <t (0.005)

So, the value lies in non-rejection region and we cannot reject our null hypothesis.

## **Paired Sample t-Tests**

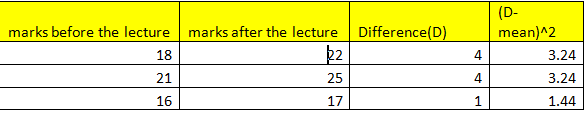
A paired t-test is used to compare two population means where you have two samples which are not independent e.g. Observations recorded on a patient before and after taking medicine, weight of a person before and after they started working out etc.

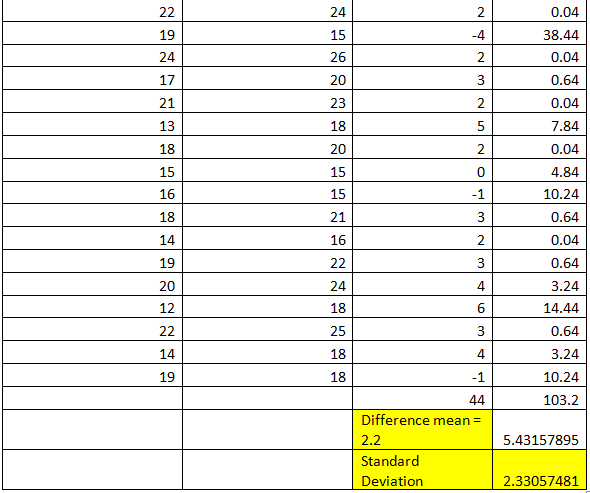
Now, instead of two separate populations, we create a new column with difference of the populations, and instead of testing equality of two population mean we test the hypothesis that mean of the population difference is zero. Also, we assume the samples are of same size. Population variances are not known and not necessarily equal.

Standard error = Deviation of differences/(n^0.5)

t= D(mean)/ standard error, where D(mean) is the men of the differences.

Q) A group 20 students were tested to see how many of them have improved marks after a special lecture on the subject.





H0: Difference mean >= 0

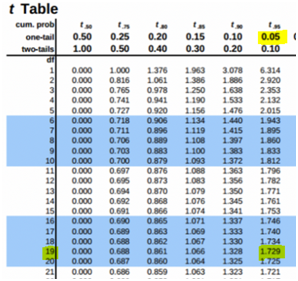
H1: Difference mean < 0

Standard error = 2.33 /(20)^0.5 = 0.52

t= 2.2 / 0.52 = 4.23

D.o.f = 19

On significance level 5%. 19 d.o.f and a one tail test, let’s calculate our critical level:



t(5%) = -1.729

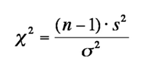
Since, t is greater than critical t, thus it lies in non-rejection region and hence we cannot reject the null hypothesis.

# **Testing of Hypothesis for population Variance Using Chi-Squared test**

Till now we were dealing with hypothesis testing for the means of various samples, but sometimes it is also necessary or desired to test the variances of the population under study i.e. let’s we obtained certain variance for a sample which is different than the population variance, now we need to find out if the variances are within acceptable limit or does it varies more than the desired variance of the population.

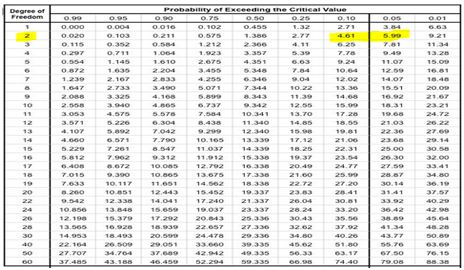
The chi-square test for variance is a non-parametric statistical procedure with a chi-square-distributed test statistic that is used for determining whether the variance of a variable obtained from a particular sample has the same size as the known population variance of the same variables.

The test statistic of the chi-square test for variance is calculated as follows:



where, n is sample size, s is sample deviation, σ is population std. Deviation

As similar with other tests, the critical value is obtained through a chi table on the basis of degree of freedom and significance level.



We will see about it with an example:

Q) The variance of a certain size of towel produced by a machine is 7.2 over a long period of time. A random sample of 20 towels gave a variance of 8. You nee to check if the variability for towel has increased at 5% level of significance, assuming a normally distributed sample.

Ans.

n = 20

sample variance = 8

population variance = 7.2

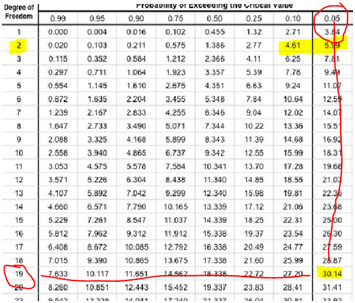
H0: variance <= 7.2

H1: variance > 7.2 (Right tailed test)

Using chi squared test,

ϗ-square = (20-1) \* 8/7.2 = 21.11

Critical value for D.o.f = 19 and 5% significance level,



Critical value = 30.14

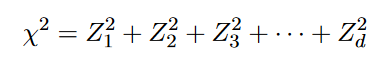
Here, chi value is less than the critical value, thus we donot reject the null hypothesis.

# **Chi-Squared Testfor Categorical Variables**

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, theχ2variable is the sum of the squares of a set of normally distributed variables.

Suppose that a particular valueZ1is randomly selected from a standardized normal distribution. Then suppose another valueZ2is 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. Theχ2variable 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 χ2 distribution with d degrees of freedom.

## **Chi Squared test For Goodness Of fit**

Chi Square test for testing goodness of fit is used to decide whether there is any difference between the observed (experimental) value and the expected (theoretical) value.

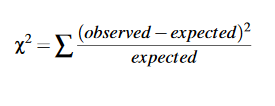
A goodness of fit test is a test that is concerned with the distribution of one categorical variable.

The null andalternative hypotheses reflect this focus:

H0: The population distribution of the variable is the same as the proposed distribution

HA: The distributions are different

The chi-square statistic is calculated as:



Where, Observed= actual count values in each category

Expected= the predicted (expected) counts in each category if the null hypothesis were true.

Let’s see an example for better understanding:

**Q) A survey conducted by a Pet Food Company determined that 60% of dog owners have only one dog, 28% have two dogs, and 12% have three or more. You were not convinced by the survey and decided to conduct your own survey and have collected the data below,**

**Data: Out of 129 dog owners, 73 had one dog and 38 had two dogs**

**Determine whether your data supports the results of the survey by the pet.**

**Use a significance level of 0.05**

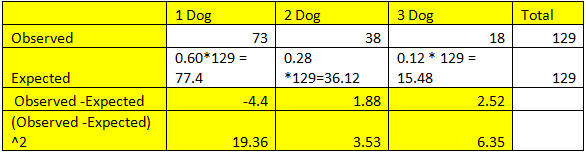
**Ans:**  E(1 dog) =0.60

E(2 dog) = 0.28

E(3 dogs) = .12

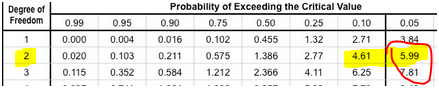
H0: proportions of dogs is equal to survey data

H1: proportions of dogs is not equal to survey data



Chi statistics = 19.36/77.4 + 3.53/36.12 + 2.52/15.48 = 0.7533

Let’s see the critical value using d.o.f 2 and significance 5%:



# 

# 

# **Analysis of Variance (ANOVA)**

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

When we have only two samples, t-test and ANOVA give the same results. 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 have a compounded effect on the type 1 error.

**Assumptions in ANOVA**

1) Assumption of Randomness: The samples should be selected in a random way such that there is no dependence among the samples.

2) The experimental errors of the data are normally distributed.

3) 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 said to be 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.

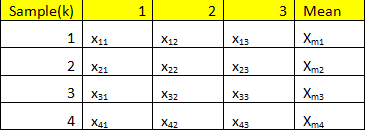
**The ANOVA model:**

Mathematically, ANOVA can be written as:

xij = μi + 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 mean plus error.

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



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 last column.

**Grand Mean**

Mean is a simple or arithmetic average of a range of values. 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

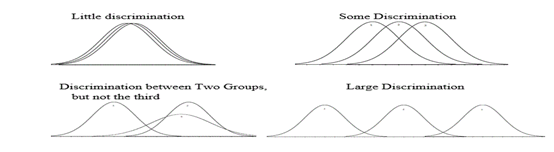
**For our dataset, k = 4**

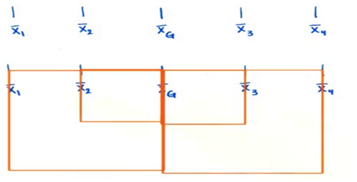
Xgm= (Xm1+Xm2+ Xm3+ Xm4)/4

**Between Group Variability (SST)**

It refers to variations between the distributions 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, 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.

Degree of Freedom for within variability:

Where, N is the total number of observations.

In our dataset example we have k =4 and 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 od Freedom = N-k = 12 -4 = 8

MeanSSwithin(MSSE) = SSwithin/ 8

Total Sum of Squares(TSS)

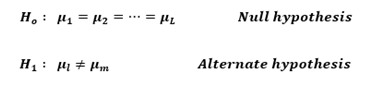
TSS = SSbetween + SSwithin = SST + SSE

**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 is different from the rest of the sample means. In mathematical form, they can be represented as:

where µ1 andµm belong to any two sample means 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 is given by the F-statistic.



F-Statistic

The statistic which measures if the means of different samples are significantly different or not is called the F-Ratio. Lower the F-Ratio, more similar are the sample means. In that case, we cannot reject the null hypothesis.

F = MeanSSbetween / MeanSSwitihn

F = MSST / MSSE with k-1 and N-k degrees of freedom.

This above formula is pretty intuitive. The numerator term in the F-statistic calculation defines the between-group variability. As we read earlier, as between group variability increases, sample means grow further apart from each other. In other words, the samples are more probable to be belonging to totally different populations.

This F-statistic calculated here is compared with the F-critical value for making a conclusion.

F-critical is calculated using the F-table, degree of freedoms and Significance level.

If observed value of F is greater than the F-critical value then we reject the null hypothesis.

