Unit-2

Outline:

1. Introduction to Statistics
    a. Terminologies in Statistics
2. Probability
    a. Introduction
    b. Random Variables
    c. Binomial Distribution
    d. Continuous Random Variables
    e. Central Limit Theorem
    f. Normal Distribution
    g. Area under normal distribution
    h. Confidence Interval
    i. Margin of Error
    j. Z Scores
3. Types of Statistics
    a. Descriptive Statistics
        i. Introduction
        ii. Mean
        iii. Median
        iv. Mode
        v. Standard Deviation
        vi. Variance
    b. Inferential Statistics
    c. Introduction
    d. Importance
    e. Hypothesis Testing, Types of Errors
    f. T-tests and Types of T-tests
    g. ANOVA
    h. Chi-square test

1. Introduction to Statistics

Math and Stats are the building blocks of Machine Learning algorithms. It is important to know the techniques behind various Machine Learning algorithms in order to know how and when to use them. Now the question arises, what exactly is Statistics?

"Statistics is a Mathematical Science pertaining to data collection, analysis, interpretation and presentation."

Statistics – Math and Statistics for Data Science.

Statistics Applications – Math and Statistics for Data Science.

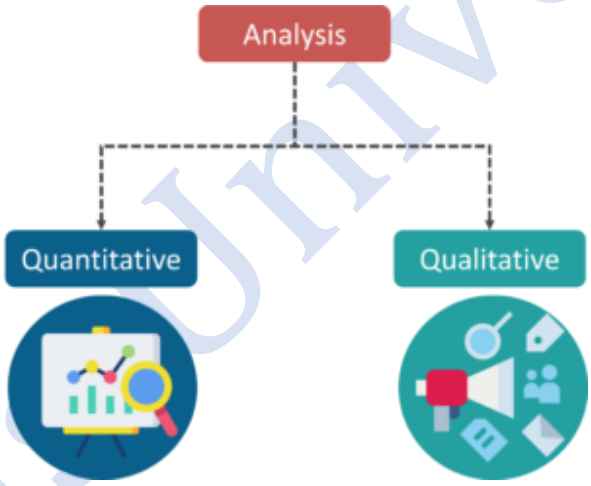
The field of Statistics has an influence over all domains of life, the Stock market, life sciences, weather, retail, insurance and education are but to name a few.
Moving ahead let's discuss the basic terminologies in Statistic.

2. Terminologies in Statistics

- ***Population*** *is the set of sources from which data has to be collected.*
- *A* ***Sample*** *is a subset of the Population*
- *A* ***Variable*** *is any characteristics, number, or quantity that can be measured or counted. A variable may also be called a data item.*
- *Also known as a statistical model, a statistical* ***Parameter*** *or population parameter is a quantity that indexes a family of probability distributions. For example, the mean, median, etc. of a population.*

Types of Analysis

An analysis of any event can be done in one of two ways:



Types of Analysis – Math and Statistics for Data Science.

1. **Quantitative Analysis:** Quantitative Analysis or the Statistical Analysis is the science of collecting and interpreting data with numbers and graphs to identify patterns and trends.
2. **Qualitative Analysis:** Qualitative or Non-Statistical Analysis gives generic information and uses text, sound and other forms of media to do so.

For example, if I want a purchase a coffee from Starbucks, it is available in Short, Tall and Grande. This is an example of Qualitative Analysis. But if a store sells 70 regular coffees a week, it is Quantitative Analysis because we have a number representing the coffees sold per week.

3. Probability

a) Introduction

What is Probability?

Simply put, probability is an intuitive concept. We use it on a daily basis without necessarily realizing that we are speaking and applying probability to work.

Life is full of uncertainties. We don't know the outcomes of a particular situation until it happens. Will it rain today? Will I pass the next math test? Will my favorite team win the toss? Will I get a promotion in next 6 months? All these questions are examples of uncertain situations we live in.

Experiment – are the uncertain situations, which could have multiple outcomes. Whether it rains on a daily basis is an experiment.

Outcome- is the result of a single trial. So, if it rains today, the outcome of today's trial from the experiment is "It rained"

Event- is one or more outcome from an experiment. "It rained" is one of the possible event for this experiment.

Probability is a measure of how likely an event is. So, if it is 60% chance that it will rain tomorrow, the probability of Outcome "it rained" for tomorrow is 0.6.

Why do we need probability?

In an uncertain world, it can be of immense help to know and understand chances of various events. You can plan things accordingly. If it's likely to rain, I would carry my umbrella. If I am likely to have diabetes on the basis of my food habits, I would get myself tested. If my customer is unlikely to pay me a renewal premium without a reminder, I would remind him about it.

So knowing the likelihood might be very beneficial

b) Random Variables

To calculate the likelihood of occurrence of an event, we need to put a framework to express the outcome in numbers. We can do this by mapping the outcome of an experiment to numbers.

Let's define X to be the outcome of a coin toss. X = outcome of a coin toss

Possible Outcomes: 1 if heads, 0 if tails

Let's take another one.

Suppose, I win the game if I get a sum of 8 while rolling two fair dice. I can define my random variable Y to be (the sum of the upward face of two fair dice)

Y can take values = (2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12)

A few things to note about random variables:

Each value of the random variable may or may not be equally likely. There is only 1 combination of dice, with sum 2{(1,1)}, while a sum of 5 can be achieved by {(1,4), (2,3), (3,2), (4,1)}. So, 5 is more likely to occur as compared to 2. On the contrary, the likelihood of a head or a tail in a coin toss is equal and 50- 50.

Sometimes, the random variables can only take fixed values, or values only in a certain interval. For example in a dice, the top face will only show values between 1 and 6. It cannot take a 2.25 or a 1.5. Similarly, when a coin is flipped, it can only show heads and tails and nothing else. On the other hand, if I define my random variable to be the amount of sugar in orange. It can take any value like 1.4g, 1.45g, 1.456g, 1.4568g as so on. All these values are possible and all infinite values between them are also possible. So, in this case, the random variable is continuous with a possibility of all real numbers.

Don't think random variable as a traditional variable (even though both are called variables) like y=x+2, where the value of y is dependent on x. Random variable is defined in terms of the outcome of a process. We quantify the process using the random variable.

c) Binomial Distribution

Most of the times, the situations we encounter are pass-fail type. The democrats either win or lose the election. I either get a heads or tails on the coin toss. You either win or lose your football game (assuming that there is always a forced outcome). So there are only two outcomes – win and lose or success and failure. The likelihood of the two may or may not be the same.

Let us understand this through an interesting example.

Let's say your football team is playing a series of 5 games against your opponent. Whoever wins more games (out of 5) wins the title.

Let us say, your team might is more skilled and has 75% chances of winning. So, there is a 25% chance of losing it.

What is the probability of you winning the series? Is it 75% or is it something else? Let us find out. What are the possible scenarios in playing 5 games?

WWWWW, WWWWL, WWWLL, WWLLL, WLLLL, LLLLL, LWWWW and so on….

So for the first game, there are two possibilities, you either win or lose, again for the second game we have two possibilities. Assuming that the first game has no effect on the outcome of the second – No one gets tired, no one gets under pressure after losing etc.

So let's define our random variable X to be a number of wins in 5 games. Remember probability of winning is 0.75 and losing is 0.25. Assume that a tie doesn't happen.

X=Number of wins in 5 games

So the first game has 2 outcomes – win and lose, second again has 2 and so on.

So total possibilities is 2*2*2*2*2 = 32

P (X=0) denotes the probability that you lose all the games and there is only one way that can happen i.e. {LLLLL} = 0.25*0.25*0.25*0.25*0.25 (multiplying the probabilities of losing the each time, lost first time **and** second time **and** third time and so on..)

P(X=1) denotes the probability that you win only 1 game i.e.(WLLLL or LWLLL or LLWLL or LLLWL or LLLLW). So there are 5 cases where you win 1 game = 5*0.75*0.25*0.25*0.25*0.25=0.0146

While we can count each of these possible outcomes, it becomes very exhaustive and intensive exercise. Let us take help of combinatorics here. Choose 2 wins out of 5 games = ${}^5C_2$ ()

so, the Probability for getting **k** successes in **n** Bernoulli trails is given by:

$P(X=k) = {}^nC_k\, p^k\, q^{n-k}$, [here p is the probability of success and q is the probability of failure] Let's see how this comes.

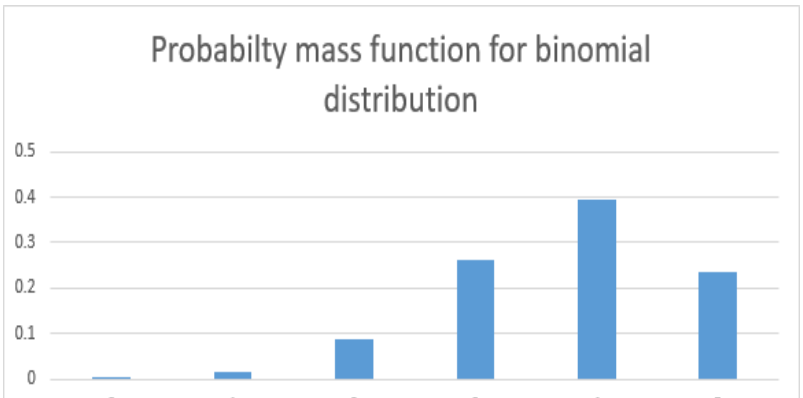
P(X=2) denotes the probability that you win 2 games. So there are ${}^5C_2$() = 10 cases where you win 2 games. Hence probability = 10*0.75*0.75*0.25*0.25*0.25=0.088

P(X=3) denotes the probability that you win 3 games. So, there are ${}^5C_3$() =10 cases where you win 3 games. Hence probability = 10*0.75*0.75*0.75*0.25*0.25=0.264

Similarly, P(X=4) = 0.395

P(X=5) = 0.237

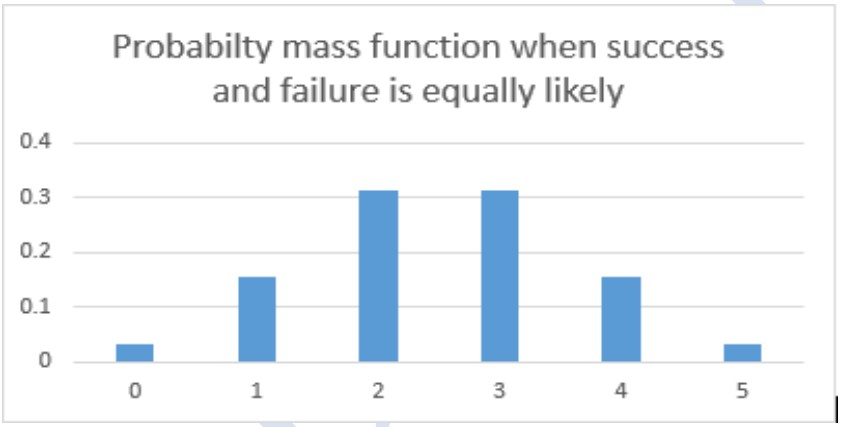
What we just calculated were discrete probabilities for a Binomial distribution. If we look at these probabilities we get something like:



As you can see the probability of winning the series is much higher than 0.75.

The general definition of a binomial distribution is the discrete probability distribution of the number of success in a sequence of n independent Bernoulli trials (having only yes/no or true/false outcomes).

If the events are equally likely to occur i.e. p = q = 0.5, the probability distribution looks something like the graph below. Here the probability of success and failure is the same.



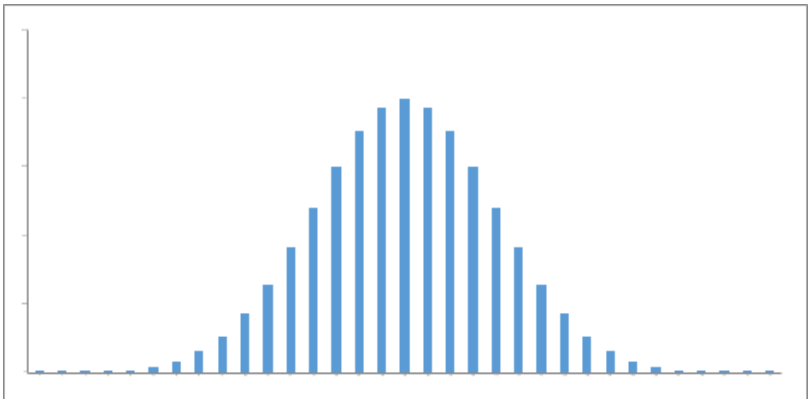
What difference do we see in the two probability distributions? The first one is skewed towards right. Reason being the likelihood to win is more, hence more wins are more likely than more losses.

In the second case when wins and losses are equally likely, so the distribution is symmetrical.

Let's assume that probability of winning and losing is equal. p=q=0.5

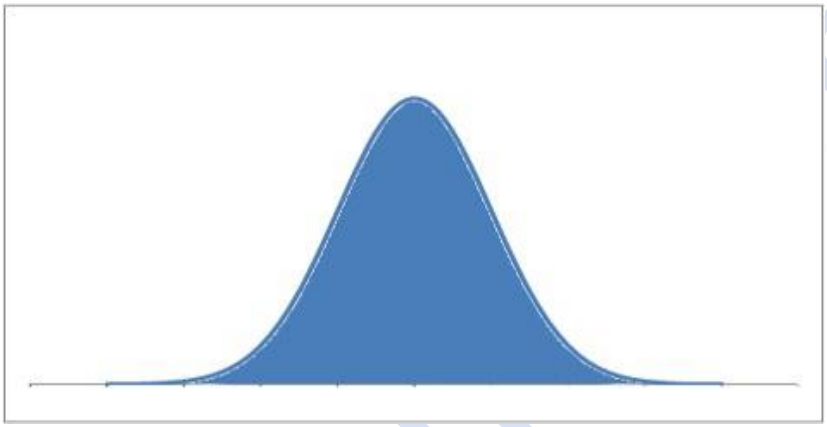
Now, What if I increase my number of trials? What if I play 20 games of football with a probability of winning and losing to be 50-50? There are a lot more possibilities and combinations. The bars get thinner and thinner.

The bars get thinner and thinner.



What if I play an infinite number of times with equal probability for winning and losing?

The bars get infinitely small and the probability distribution looks something like a continuous set of bars which are very close, almost continuous. This now becomes a probability density function. Notice that this now becomes a continuous function.



Let's point out some interesting things that happened.

The probabilities are high for the mean values of the random variables. When we were playing 5 football games, the random variable (X = the wins) could take values between 0 – 5. The mean value being 2.5. The probability is highest for 2 and 3. When we move towards the continuous curve, the probability is highest for the exact mean

The probabilities are low as we move away from the mean.

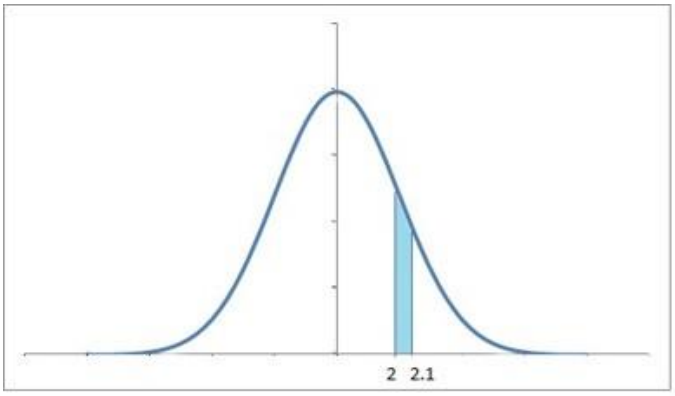
d) Continuous Random Variables

Let's see some cases where the random variables are continuous. Let's say the weatherman is trying to measure the amount of rainfall that will happen tomorrow.

Let's say the rainfall likely to happen is around 2 cm. But will it be exactly 2 cm? It can be 2.001 or

2.000001 Or 2.000000001 and an infinite number of values in between. It's even impossible for us to measure if it's exactly 2 cm.

So, we calculate the probability of it, being in a range. We calculate the probability of rainfall being in the range of 2 cm to 2.01 cm. It will be the sum of probabilities for all values between 2 and 2.01. The area under the probability density function with limits 2 and 2.01 will give us that.

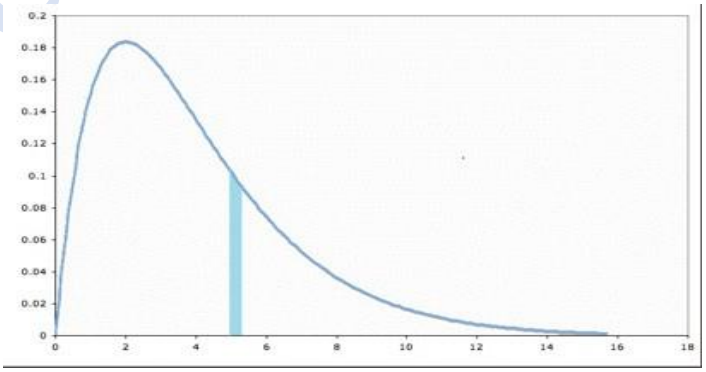


The probability density function may or may not be symmetrical.

Life of an insect

Suppose there is an insect whose lifespan ranges from 0 to 16 days. We're looking for the probability that it will die in around 5 to 6 days. Again we would need the sum of probabilities for all values between 5 days and 6 days.

We look at the probability density function and find the area of the graph under the limits of 5 and 6. We can use definite integration under the desired limits for the probability density to find the area. We're often interested in the probability of a range of values rather than the probability of an exact value.



We can now imagine that the probability at a particular point would be the area of the thinnest possible bar we can imagine. To calculate the probability at x, we would need the area from x to x+Δ, where Δ is very small.

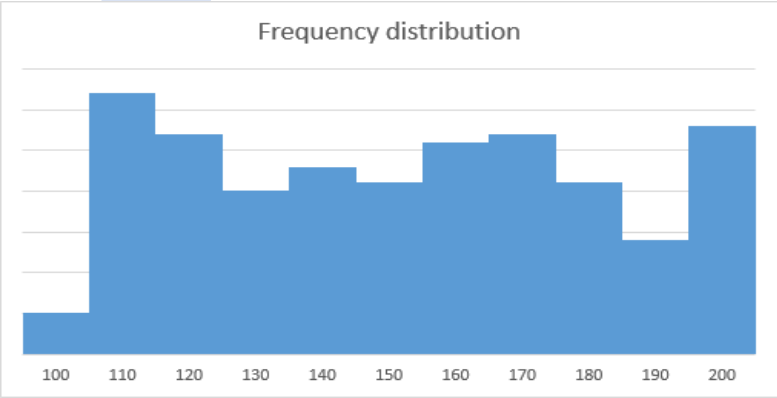
The total probability density function would then be the collection of all such areas / probabilities. The formula of the probability density function can be written as:

e) Central Limit Theorem

$$f_X(x) = \lim_{\Delta \to 0^+} \frac{P(x < X \leq x + \Delta)}{\Delta}$$

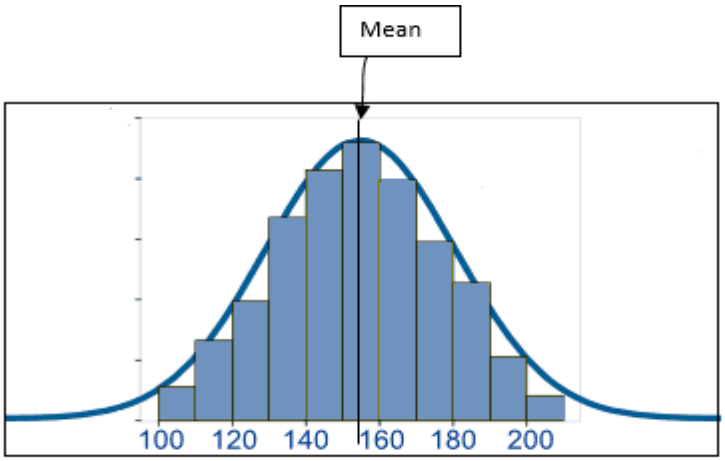
So when you have huge amount of data, you can be confused how to make sense of it. It is difficult to know what's happening underneath it. To tackle this problem, what we do is take a small chunk of data & look at it. But we won't be satisfied with just a single chunk. We'd try to look at multiple chunks to be sure of results.

Let's say we have the cholesterol levels of all the people in India, we can look at the mean, median and mode of the data. Maybe plot a histogram with sensible ranges and look at the data. Let's assume this is how the data looks like. The mean of this data is 153.2



But this huge amount of data is really tough to process. To process it, we take the data of some 50 people and calculate their mean.

We again take a sample of some 50 people and calculate the mean and we keep doing that for quite a number of times. We now plot the means of these samples.



We see that these sample means form a frequency distribution which looks very symmetrical. The frequency around the mean of the actual data is the highest and gradually reduces as we move away from the mean on the either side.

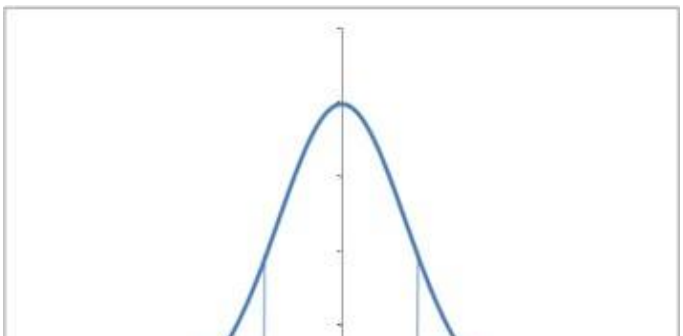
So when we take means of cholesterol levels of 50 people, again and again, we observe the mean values are around 150-160. Only a few mean values is more than 170 and less than 140. There are very, very few over 190 or less than 110.

We can easily convert the frequencies to see probabilities. If we divide the frequency of a bin (range like 110 to 120) by the total number of data points, we get the probabilities of each bin. So, now the frequency distribution becomes a probability distribution of the same shape. The probability distribution approaches more and more towards

The probability distribution approaches more and more towards symmetry, when the sample size that we use to create those means, is very large. As the sample size approaches infinity, the probability distribution becomes a perfectly symmetrical where the center of the curve is the mean of the population. The curve is known as normal distribution.

1. Normal Distribution

The normal distribution informally called as a bell curve looks like this



The equation of the normal distribution happens to be:

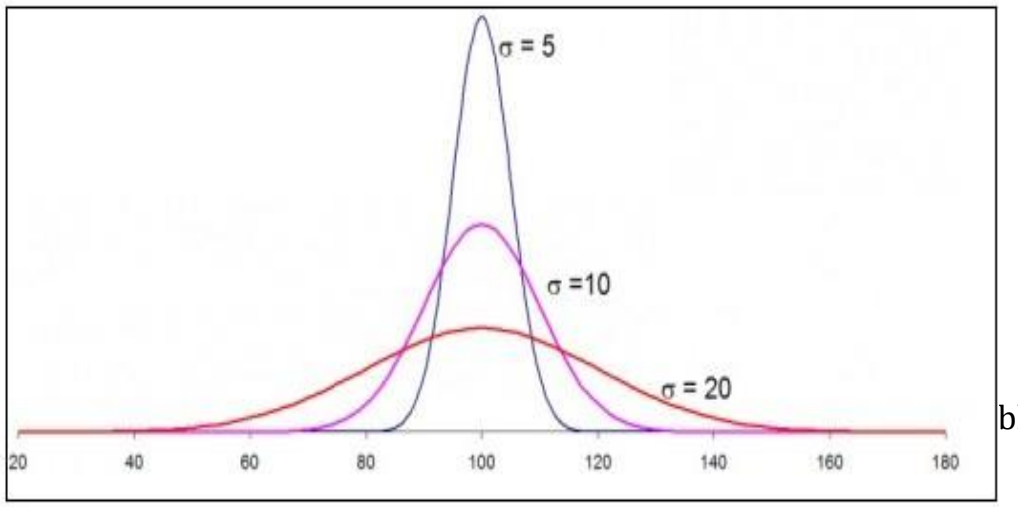
$$f(x|\mu, \sigma^2) = \frac{1}{\sqrt{(2\sigma^2\pi)}}\left(e^{-\frac{(x-\mu)^2}{2\sigma^2}}\right)$$

Here μ is the mean of the data while σ is the standard deviation of the data.

The normal distribution is perfectly symmetrical about the mean. The probabilities move similarly in both directions around the mean. The total area under the curve is 1, since summing up all the possible probabilities would give 1.

The distribution might vary a bit depending upon how spread the data is. If the data has a very high range and standard deviation, the normally distributed curve would be spread out and flatter, since a large number of values would be sufficiently away from the mean.

Also, if a lot of values are away from the mean, the probability for data being around the mean also drops. Similarly, if the standard deviation is low, which means most of the values are near around the mean, there is high probability of the sample mean being around the mean and the distribution is a lot skinnier. The higher the standard deviation, the thicker and flatter the curve.



Let's summaris

Area under a pr ble to be in that range.

If I have a population data and I take random samples of equal size from the data, the sample means are approximately normally distributed

There is large probability for the means to be around the actual mean of the data, than to be farther away

Normal distributions for higher standard deviations are flatter as compared to those for lower standard deviations

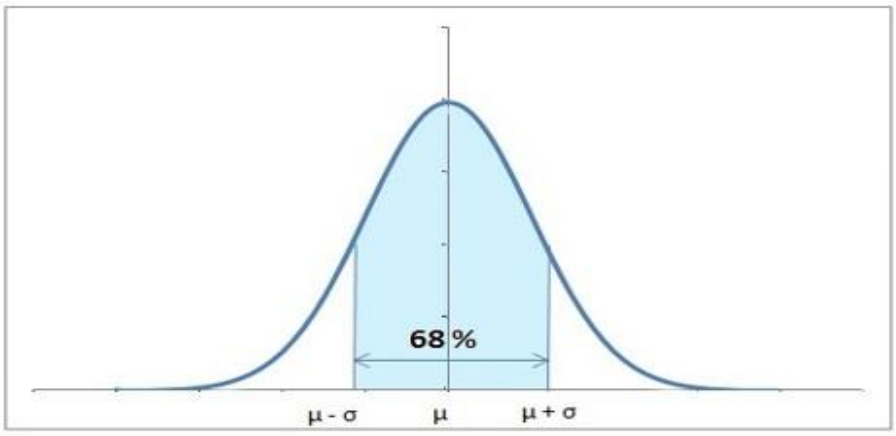
2. Area under normal distribution

Now, let's say I have a dataset of cholesterol levels of a number of patients and we need to calculate the probability of how many patients are healthy. The mean value (μ) for cholesterol of all the patients is equal to 150 and standard deviation (σ) is equal to 15. The probability density function is a normal distribution given by the above equation.

We need to calculate the probability of cholesterol levels to be between 135 (150-15) and 165 (150+15) – the healthy cholesterol range.

Can you see that the healthy patients that we are talking about are one standard deviation on either side of the mean? This means we need to calculate the area under the curve with 135 and 165 as limits. Don't worry, this area for normal distribution is already calculated for us and is ~68%.

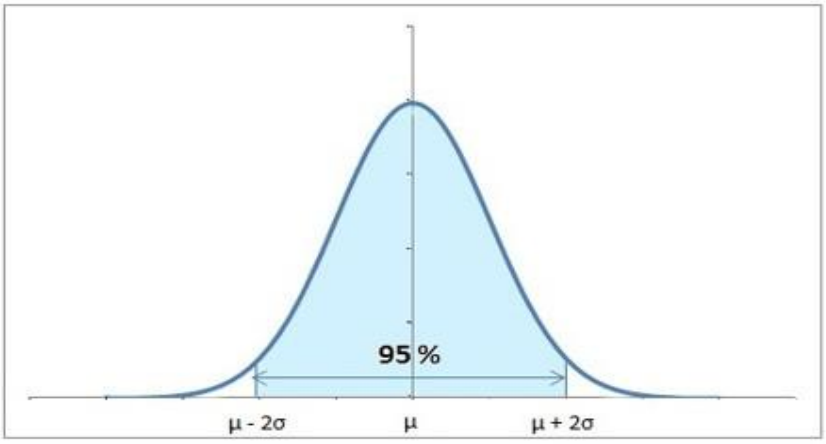
So always for a normally distributed data, around 68% of the data falls within 1 standard deviation of the mean. So probability of the data being within 1 standard deviation if the mean = 0.68



Let's also calculate the probability of being 2 standard deviations away from the mean. Let's say we need to warn the patients who are two standard deviations away.

This means 150+30 and 150-30.

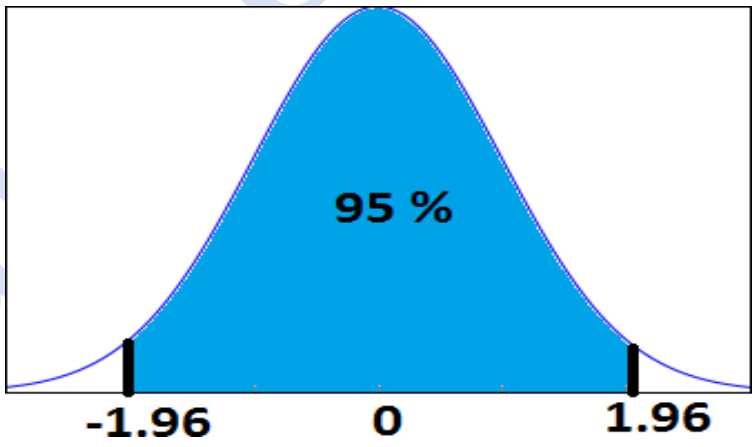
The range of area to be calculated now is 120 to 180. To your surprise, 95% of the values fall in this range.



So, 95% of the patients have their cholesterol levels between 120 and 180. And the remaining 5% are really critical and different from the average values.

3. Confidence Interval

The confidence interval is a type of interval estimate from the sampling distribution which gives a range of values in which the population statistic may lie. Let us understand this with the help of an example.



We know that 95% of the values lie within 2 (1.96 to be more accurate) standard deviation of a normal distribution curve. So, for the above curve, the blue shaded portion represents the confidence interval for a sample mean of 0.

Formally, Confidence Interval is defined as,

$$C.I = \bar{X} \pm Z_{a/2}\, \sigma/\sqrt{n}$$

Whereas,

$\bar{X}$ = the sample mean

$Z_{a/2}$ = Z value for desired confidence level α

σ = the population standard deviation

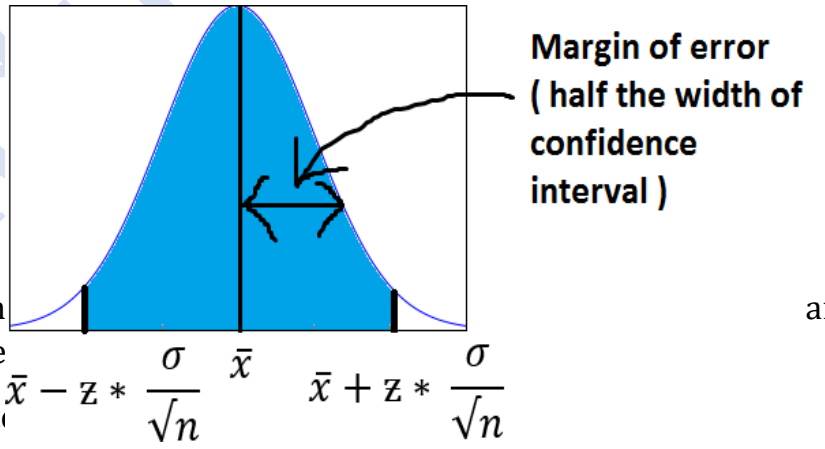
For an alpha value of 0.95 i.e 95% confidence interval, z=1.96.

4. Margin of Error

Now there is one more term which you should be familiar with, *Margin of Error*. It is given as {(z.σ)/√n} and defined as the sampling error by the surveyor or the person who collected the samples. That means, if a sample mean lies in the margin of error range then, it might be possible that its actual value is equal to the population mean and the difference is occurring by chance. Anything outside margin of error is considered *statistically significant*.

And it is easy to infer that the error can be both positive and negative side. The whole margin of error on both sides of the sample statistic constitutes the Confidence Interval. Numerically, C.I is twice of Margin of Error.

The below image will help you better visualize Margin of Error and Confidence Interval.



The shaded portion and half of it is Margin of Error which can be

Interesting points t

1. Confidence Intervals can be built with difference degrees of confidence suitable to a user's needs like 70 %, 90% etc.

2. Greater the sample size, smaller the Confidence Interval, i.e more accurate determination of population mean from the sample means.
3. There are different confidence intervals for different sample means. For example, a sample mean of 40 will have a difference confidence interval from a sample mean of 45.
4. By 95% Confidence Interval, we do not mean that – The probability of a population mean to lie in an interval is 95%. Instead, 95% C.I means that 95% of the Interval estimates will contain the population statistic.

Example

Calculate the 95% confidence interval for a sample mean of 40 and sample standard deviation of 40 with sample size equal to 100.

Solution:

We know, z-value for 95% C.I is 1.96. Hence, Confidence Interval (C.I) is calculated as: C.I= [{x(bar) – (z*s/√n)},{x(bar) – (z*s/√n)}]

C.I = [{40-(1.96*40/10},{ 40+(1.96*40/10)}]

C.I = [32.16, 47.84]

f) Z Scores

We will encounter a lot of cases, where we would need to know the probability for the data to be less than or more than a particular value. This value will not be equal to 1σ or 2σ away from the mean.

The distance in terms of number of standard deviations, the observed value is away from the mean, is the Standard score or the Z score.

A positive Z score indicates that the observed value is Z standard deviations above the mean. Negative Z score indicates that the value is below the mean.

Observed value = μ+zσ [μ is the mean and σ is the standard deviation]

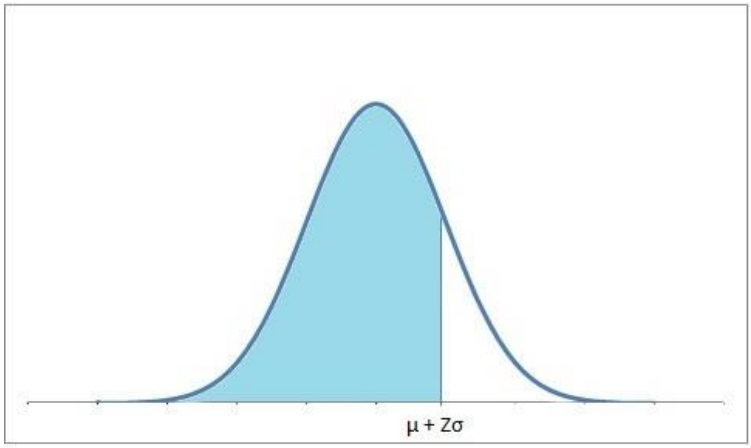
In our cholesterol example, let's see where 172 falls on the distribution. We will calculate the Z score to find the percentage of people having cholesterol less than 172.

172 = 150+Z*15

Here, we see that 172 is 1.47 {(142-150)/15} standard deviations more than the mean. This 1.47 is known as the z value.

Now, we would need to use these limits to calculate the area under the curve. Remember that the area under the curve is 1. Let's calculate the probability of people having a cholesterol level of less than 172.

To your happiness, you will never have to actually calculate the area under the normal curve, we have the z table that can be used to calculate the probabilities for particular z values. The rows of the Z table have the Z score in tens, while the hundredths decimal is given by the columns. The value is the area under the curve less than that Z score.

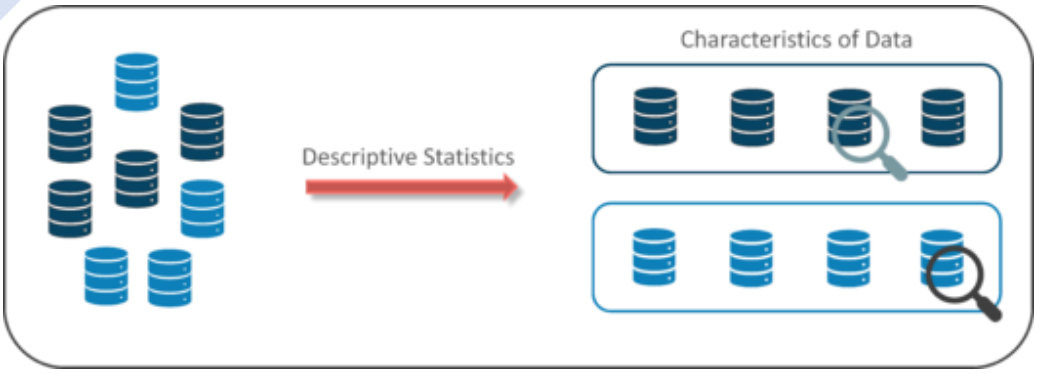


For a particular z score, we can look into the table to find the probability for values to fall less than that particular z value. It can be –ve or +ve. If we look out for 1.47, we find that ~93% data falls less than that. Therefore, 93% patients have cholesterol less than 172. Also, we can safely say that 7% have cholesterol more than 172.

4. Types of Statistics

a) Descriptive Statistics

Descriptive Statistics helps organize data and focuses on the characteristics of data providing parameters.



Descriptive Statistics – Math and Statistics for Data Science

Suppose you want to study the average height of students in a classroom, in descriptive statistics you would record the heights of all students in the class and then you would find out the maximum, minimum and average height of the class.

Descriptive Statistics Example – Math and Statistics for Data Science

Understanding Descriptive Analysis

When we try to represent data in the form of graphs, like histograms, line plots, etc. the data is represented based on some kind of central tendency. Central tendency measures like, mean, median, or measures of the spread, etc. are used for statistical analysis. To better understand Statistics lets discuss the different measures in Statistics with the help of an example.

Using descriptive Analysis, you can analyse each of the variables in the sample data set for mean, standard deviation, minimum and maximum.

If we want to find out the mean or average horsepower of the cars among the population of cars, we will check and calculate the average of all values. In this case, we'll take the sum of the Horse Power of each car, divided by the total number of cars:

***Mean** = (110+110+93+96+90+110+110+110)/8 = 103.625*

If we want to find out the center value of mpg among the population of cars, we will arrange the mpg values in ascending or descending order and choose the middle value. In this case, we have 8 values which is an even entry. Hence we must take the average of the two middle values.

The mpg for 8 cars: 21,21,21.3,22.8,23,23,23,23

***Median** = (22.8+23)/2 = 22.9*

If we want to find out the most common type of cylinder among the population of cars, we will check the value which is repeated most number of times. Here we can see that the cylinders come in two values, 4 and 6. Take a look at the data set, you can see that the most recurring value is 6. Hence 6 is our Mode.

Measures of the Spread

Just like the measure of center, we also have measures of the spread, which comprises of the following measures:

1. **Range:** It is the given measure of how spread apart the values in a data set are.
2. **Inter Quartile Range (IQR):** It is the measure of variability, based on dividing a data set into quartiles.
3. **Variance:** It describes how much a random variable differs from its expected value. It entails computing squares of deviations.
    1. ***Deviation** is the difference between each element from the mean.*
    2. ***Population Variance** is the average of squared deviations*
    3. ***Sample Variance** is the average of squared differences from the mean*
    4. **Standard Deviation:** It is the measure of the dispersion of a set of data from its mean.

Basic Formulas:-

Statistics Formula

$$\text{Mean } \bar{x} = \frac{\sum xi}{N} \qquad \text{Variance} = \frac{\sum (xi - \bar{x})^2}{N}$$

$$\text{Median} = \begin{cases} \frac{(N+1)^{th}}{2} \text{ term; when N is even} \\ \frac{\frac{N^{th}}{2} \text{ term} + \left(\frac{N}{2} + 1\right) \text{term}}{2} \text{ ; when N is even} \end{cases}$$

$$\text{Mode} = \text{The value in the data set that occurs most frequently}$$

Standard Deviation is Under Root of Variance.

b) Inferential Statistics

1) Introduction

Statistics is one of the key fundamental skills required for data science. Any expert in data science would surely recommend learning / upskilling yourself in statistics.

However, if you go out and look for resources on statistics, you will see that a lot of them tend to focus on the mathematics. They will focus on derivation of formulas rather than simplifying the concept. I believe, statistics can be understood in very simple and practical manner. That is why I have created this guide.

In this guide, I will take you through Inferential Statistics, which is one of the most important concepts in statistics for data science. I will take you through all the related concepts of Inferential Statistics and their practical applications.

This guide would act as a comprehensive resource to learn Inferential Statistics. So, go through the guide, section by section. Work through the examples and develop your statistics skills for data science.

2) Importance

Suppose, you want to know the average salary of Data Science professionals in India. Which of the following methods can be used to calculate it?

1. Meet every Data Science professional in India. Note down their salaries and then calculate the total average?
2. Or hand pick a number of professionals in a city like Gurgaon. Note down their salaries and use it to calculate the Indian average.

Well, the first method is not impossible but it would require an enormous amount of resources and time. But today, companies want to make decisions swiftly and in a cost-effective way, so the first method doesn't stand a chance.

On the other hand, second method seems feasible. But, there is a caveat. What if the population of Gurgaon is not reflective of the entire population of India? There are then good chances of you making a very wrong estimate of the salary of Indian Data Science professionals.

In simple language, Inferential Statistics is used to draw inferences beyond the immediate data available. With the help of inferential statistics, we can answer the following questions:

Making inferences about the population from the sample.

Concluding whether a sample is significantly different from the population. For example, let's say you collected the salary details of Data Science professionals in Bangalore. And you observed that the average salary of Bangalore's data scientists is more than the average salary across India. Now, we can conclude if the difference is statistically significant.

If adding or removing a feature from a model will really help to improve the model. If one model is significantly better than the other?

3) Hypothesis Testing, Types of Errors

Before we go into the theoretical explanation, let us understand Hypothesis Testing by using a simple example.

Example: Class 8th has a mean score of 40 marks out of 100. The principal of the school decided that extra classes are necessary in order to improve the performance of the class. The class scored an average of 45 marks out of 100 after taking extra classes. Can we be sure whether the increase in marks is a result of extra classes or is it just random?

Hypothesis testing lets us identify that. **It lets a sample statistic to be checked against a population statistic or statistic of another sample to study any intervention etc.** Extra classes being the intervention in the above example.

Hypothesis testing is defined in two terms – Null Hypothesis and Alternate Hypothesis.

Null Hypothesis being the sample statistic to be equal to the population statistic. For eg: The Null Hypothesis for the above example would be that the average marks after extra class are same as that before the classes.

Alternate Hypothesis for this example would be that the marks after extra class are significantly different from that before the class.

Hypothesis Testing is done on different levels of confidence and makes use of z-score to calculate the probability. So for a 95% Confidence Interval, anything above the z-threshold for 95% would reject the null hypothesis.

Points to be noted:

1. We cannot accept the Null hypothesis, only reject it or fail to reject it.
2. As a practical tip, Null hypothesis is generally kept which we want to disprove. For eg: You want to prove that students performed better after taking extra classes on their exam. The Null Hypothesis, in this case, would be that the marks obtained after the classes are same as before the classes.

Types of Errors in Hypothesis Testing

Now we have defined a basic Hypothesis Testing framework. It is important to look into some of the mistakes that are committed while performing Hypothesis Testing and try to classify those mistakes if possible.

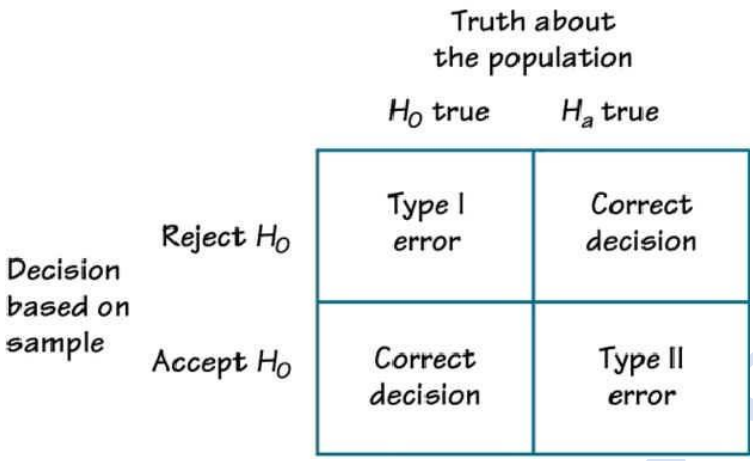
Now, look at the Null Hypothesis definition above. What we notice at the first look is that it is a statement subjective to the tester like you and me and not a fact. That means there is a possibility that the Null Hypothesis can be true or false and we may end up committing some mistakes on the same lines.

There are two types of errors that are generally encountered while conducting Hypothesis Testing.

Type I error: Look at the following scenario – A male human tested positive for being pregnant. Is it even possible? This surely looks like a case of False Positive. More formally, it is defined as the incorrect rejection of a True Null Hypothesis. The Null Hypothesis, in this case, would be – Male Human is not pregnant.

Type II error: Look at another scenario where our Null Hypothesis is – A male human is pregnant and the test supports the Null Hypothesis. This looks like a case of False Negative. More formally it is defined as the acceptance of a false Null Hypothesis.

The below image will summarize the types of error:



4) T-tests and Types of T-tests

T-tests are very much similar to the z-scores, the only difference being that instead of the Population Standard Deviation, we now use the Sample Standard Deviation. The rest is same as before, calculating probabilities on basis of t-values.

The Sample Standard Deviation is given as:

$$s = \frac{\sqrt{\sum(x-\bar{x})^2}}{(n-1)}$$

Where n-1 is the Bessel's correction for estimating the population parameter.

Another difference between z-scores and t-values are that t-values are dependent on Degree of Freedom of a sample. Let us define what degree of freedom is for a sample.

The Degree of Freedom – It is the number of variables that have the choice of having more than one arbitrary value. For example, in a sample of size 10 with mean 10, 9 values can be arbitrary but the 1oth value is forced by the sample mean.

Points to note about the t-tests:

1. Greater the difference between the sample mean and the population mean, greater the chance of rejecting the Null Hypothesis. Why? (We discussed this above.)
2. Greater the sample size, greater the chance of rejection of Null Hypothesis.

Different types of t-tests

1. Sample t-test

This is the same test as we described above. This test is used to:

Determine whether the mean of a group differs from the specified value. Calculate a range of values that are likely to include the population mean.

For eg: A pizza delivery manager may perform a 1-sample t-test whether their delivery time is significantly different from that of the advertised time of 30 minutes by their competitors.

$$t = \frac{\bar{X} - \mu}{\frac{s}{\sqrt{N}}}$$

Where, **X(bar)** = sample mean

μ = population mean

s = sample standard deviation

N = sample size

2. Paired t-test

Paired t-test is performed to check whether there is a difference in mean after a treatment on a sample in comparison to before. It checks whether the Null hypothesis: The difference between the means is Zero, can be rejected or not.

	A	B	C	D	E	F	G
1	Before	After		t-Test: Paired Two Sample for Means			
2	1.2689	-1.3681					
3	-2.3645	0.2332			*Before*	*After*	
4	0.2698	0.5236		Mean	0.7479	0.598906667	
5	0.3456	0.1452		Variance	6.303513117	2.787580174	
6	-3.4156	-3.4256		Observations	15	15	
7	6.1458	2.1253		Pearson Correlation	0.644292336		
8	3.1569	3.1526		Hypothesized Mean Difference	0		
9	0.1235	-1.196		df	14		
10	2.1023	1.5631		t Stat	0.30041793		
11	-1.3698	1.4785		P(T<=t) one-tail	0.384136606		
12	1.8896	0.5645		t Critical one-tail	1.761310136		
13	0.1463	0.2589		P(T<=t) two-tail	0.768273211		
14	-2.3512	0.6587		t Critical two-tail	2.144786688		
15	2.1253	2.1452					
16	3.1456	2.1245					
17							

The above examp l and that there is no significant differe e is not less than the alpha value (0.05

$$t = \frac{\bar{d}}{S_d / \sqrt{n}}$$

Where, **d (bar)** = mean of the case wise difference between before and after,

S_d = standard deviation of the difference

n = sample size.

3. 2-sample t-test

This test is used to determine:

- Determine whether the means of two independent groups differ.
- Calculate a range of values that is likely to include the difference between the population means.

$$t = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{s_1^2}{N_1} + \frac{s_2^2}{N_2}}}$$

The above formula represents the 2 sample t-test and can be used in situations like to check whether two machines are producing the same output. The points to be noted for this test are:

1. The groups to be tested should be independent.
2. The groups' distribution should not be highly skewed.

where, **X1 (bar)** = mean of the first group.

S_1 = represents 1st group sample standard deviation.

N_1 = represents the 1st group sample size.

Practical Example:-

We will understand how to identify which t-test to be used and then proceed on to solve it. The other t- tests will follow the same argument.

Example: A population has mean weight of 68 kg. A random sample of size 25 has a mean weight of 70 with standard deviation =4. Identify whether this sample is representative of the population?

Step 0: Identifying the type of t-test

Number of samples in question = 1

Number of times the sample is in study = 1 any intervention on sample = No

Recommended t-test = 1- sample t-test.

Had there been 2 samples, we would have opted for 2-sample t-test and if there would have been 2 observations on the same sample, we would have opted for paired t-test.

Step 1: State the Null and Alternate Hypothesis

Null Hypothesis: The sample mean and population mean are same.

Alternate Hypothesis: The sample mean and population mean are different.

Step 2: Calculate the appropriate test statistic

df = 25-1 =24

$t = (70-68)/(4/\sqrt{25}) = 2.5$

Now, for a 95% confidence level, t-critical (two-tail) for rejecting Null Hypothesis for 24 d.f is 2.06. Hence, we can reject the Null Hypothesis and conclude that the two means are different.

5) ANOVA

ANOVA (Analysis of Variance) is used to check if at least one of two or more groups have statistically different means. Now, the question arises – Why do we need another test for checking the difference of means between independent groups? Why can we not use multiple t-tests to check for the difference in means?

The answer is simple. Multiple t-tests will have a compound effect on the error rate of the result. Performing t-test thrice will give an error rate of ~15% which is too high, whereas ANOVA keeps it at 5% for a 95% confidence interval.

To perform an ANOVA, you must have a continuous response variable and at least one categorical factor with two or more levels. ANOVA requires data from approximately normally distributed populations with equal variances between factor levels. However, ANOVA procedures work quite

well even if the normality assumption has been violated unless one or more of the distributions are highly skewed or if the variances are quite different.

ANOVA is measured using a statistic known as F-Ratio. It is defined as the ratio of Mean Square (between groups) to the Mean Square (within group).

Mean Square (between groups) = Sum of Squares (between groups) / degree of freedom (between groups) Mean Square (within group) = Sum of Squares (within group) / degree of freedom (within group)

Source	*SS*	*df*	*MS*	*F*
Between	SS_b	k-1	MS_b	MS_b/MS_w
Within	SS_w	N-k	MS_w	
Total	$SS_b + SS_w$	N-1		

$$\sum_{j=1}^{p}\sum_{i=1}^{n_j}(X_{ij} - \bar{X}_j)^2 = SS_{w/in}$$

$$\sum_{j=1}^{p} n_j(\bar{X}_j - \bar{X})^2 = SS_{Betw}$$

$$\sum_{j=1}^{p}\sum_{i=1}^{n_j}(X_{ij} - \bar{X})^2 = SS_{Tot}$$

Here, **p** = represents the number of groups

n = represents the number of observations in a group

X_j (bar) = represents the mean of a particular group

X (bar) = represents the mean of all the observations

Now, let us understand the degree of freedom for within group and between groups respectively.

Between groups: If there are k groups in ANOVA model, then k-1 will be independent. Hence, k-1 degree of freedom.

Within groups: If N represents the total observations in ANOVA (∑n over all groups) and k are the number of groups then, there will be k fixed points. Hence, N-k degree of freedom.

Steps to perform ANOVA

1. Hypothesis Generation
    1. Null Hypothesis : Means of all the groups are same
    2. Alternate Hypothesis : Mean of at least one group is different
2. Calculate within group and between groups variability
3. Calculate F-Ratio
4. Calculate probability using F-table
5. Reject/fail to Reject Null Hypothesis

There are various other forms of ANOVA too like Two-way ANOVA, MANOVA, ANCOVA etc. but One-Way ANOVA suffices the requirements of this course.

Practical applications of ANOVA in modeling are:

1. Identifying whether a categorical variable is relevant to a continuous variable.
2. Identifying whether a treatment was effective to the model or not.

Practical Example

Suppose there are 3 chocolates in town and their sweetness is quantified by some metric (S). Data is collected on the three chocolates. You are given the task to identify whether the mean sweetness of the 3 chocolates are different. The data is given as below:

	Type A	Type B	Type C
	643	469	484
	655	427	456
	702	525	402
$\overline{X}$	666.67	473.67	447.33
S	31.18	49.17	41.68

Here, first we have calculated the sample mean and sample standard deviation for you.

Now we will proceed step-wise to calculate the F-Ratio (ANOVA statistic).

Step 1: Stating the Null and Alternate Hypothesis

Null Hypothesis: Mean sweetness of the three chocolates are same.

Alternate Hypothesis: Mean sweetness of at least one of the chocolates is different.

Step 2: Calculating the appropriate ANOVA statistic

In this part, we will be calculating SS(B), SS(W), SS(T) and then move on to calculate MS(B) and MS(W). The thing to note is that,

Total Sum of Squares [SS(t)] = Between Sum of Squares [SS(B)] + Within Sum of Squares [SS(W)].

So, we need to calculate any two of the three parameters using the data table and formulas given above.

As, per the formula above, we need one more statistic i.e Grand Mean denoted by X(bar) in the formula above.

X bar = (643+655+702+469+427+525+484+456+402)/9 = 529.22

SS(B)=[3*(666.67-529.22)^2]+ [3*(473.67-529.22)^2]+[3*(447.33-529.22)^2] = 86049.55

SS (W) =[(643-666.67)^2+(655-666.67)^2+(702-666.67)^2]+[(469-473.67)^2+(427-473.67)^2+(525- 473.67)^2] + [(484-447.33)^2+(456-447.33)^2+(402-447.33)^2]= 10254

MS(B) = SS(B) / df(B) = 86049.55 / (3-1) = 43024.78

MS(W) = SS(W) / df(W) = 10254/(9-3) = 1709

F-Ratio = MS(B) / MS(W) = 25.17 .

Now, for a 95 % confidence level, F-critical to reject Null Hypothesis for degrees of freedom (2,6) is 5.14 but we have 25.17 as our F-Ratio.

So, we can confidently reject the Null Hypothesis and come to a conclusion that at least one of the chocolate has a mean sweetness different from the others.

Note: ANOVA only lets us know the means for different groups are same or not. It doesn't help us identify which mean is different. To know which group mean is different, we can use another test know as Least Significant Difference Test.

6) Chi-square test

Sometimes, the variable under study is not a continuous variable but a categorical variable. Chi-square test is used when we have one single categorical variable from the population.

Let us understand this with help of an example. Suppose a company that manufactures chocolates, states that they manufacture 30% dairy milk, 60% temptation and 10% kit-kat. Now suppose a random sample of 100 chocolates has 50 dairy milk, 45 temptation and 5 kitkats. Does this support the claim made by the company?

Let us state our Hypothesis first.

Null Hypothesis: The claims are True

Alternate Hypothesis: The claims are False.

Chi-Square Test is given by:

$$\chi_c^2 = \sum \frac{(O_i - E_i)^2}{E_i}$$

Where,

O_i = sample or observed values

E_i = population values

The summation is taken over all the levels of a categorical variable.

E_i = = **[n *** p_i **]** Expected value of a level (i) is equal to the product of sample size and percentage of it in the population.

Let us now calculate the Expected values of all the levels. E (dairy milk)= 100 * 30% = 30

E (temptation) = 100 * 60% =60

E (kitkat) = 100 * 10% = 10

Calculating chi-square = [(50-30)^2/30+(45-60)^2/60+(5-10)^2/10] =19.58

Now, checking for p (chi-square >19.58) using chi-square calculator, we get p=0.0001. This is significantly lower than the alpha (0.05).

So we reject the Null Hypothesis.