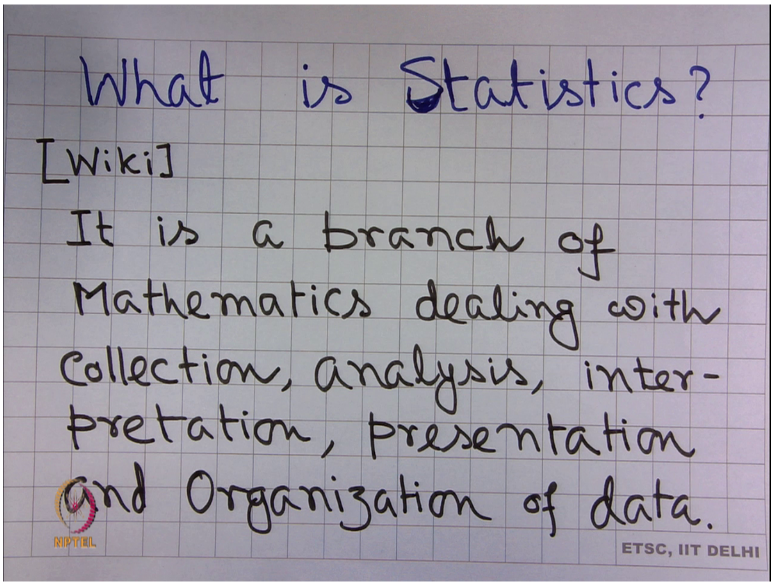
Statistical Inference
Prof. Niladri Chatterjee
Department of Mathematics
Indian Institute of Technology, Delhi

Lecture – 01
Statistical Inference

Welcome students to my MOOCs online lecture on Statistical Inference. I am planning to have about 20 lectures on this topic and this is the very first lecture of that series. It is assumed that the listeners of this course have some background of basic statistics and basic probability distributions.

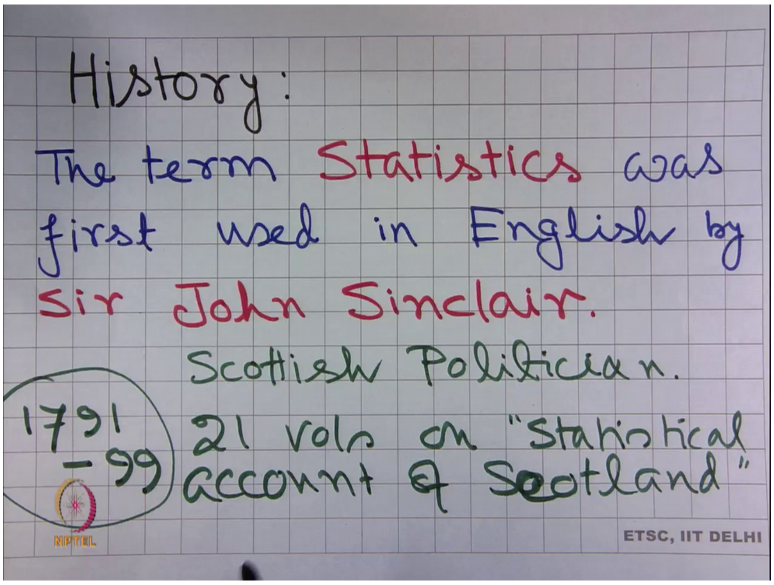
In this course of course, I will revise the probability part very quickly; and I will touch upon only those aspects which are which will be used in course of my lecture on this series. This course I expect will help under graduate students of statistics, maths and computing, computer science etcetera and also basic science students at honors level to understand the basics of statistical inference.

(Refer Slide Time: 01:25)



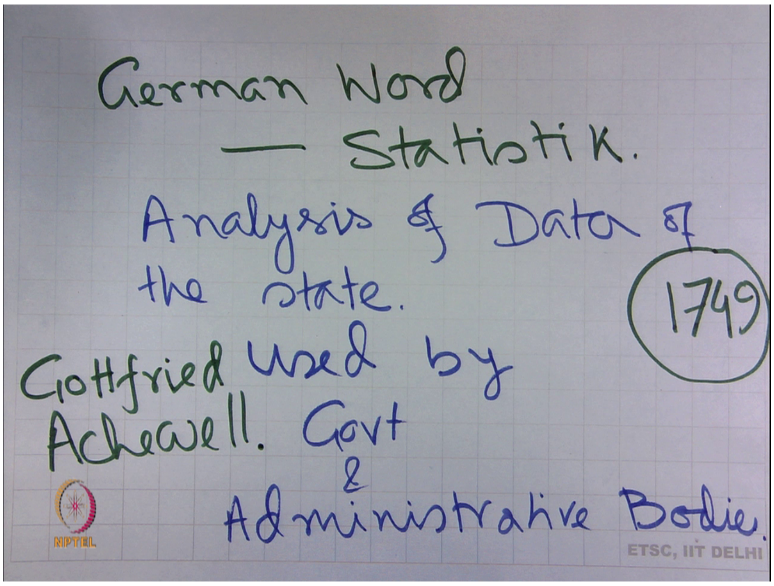
Before I go into the topic let me first explain what is statistics? As per Wikipedia, it is a branch of mathematics dealing with collection, analysis, interpretation, presentation and organization of data. So, one thing is very clear to us that statistics is something where we deal with data. This has become very important in this era of big data, when data is in abundance and we need to learn from this data.

(Refer Slide Time: 02:16)



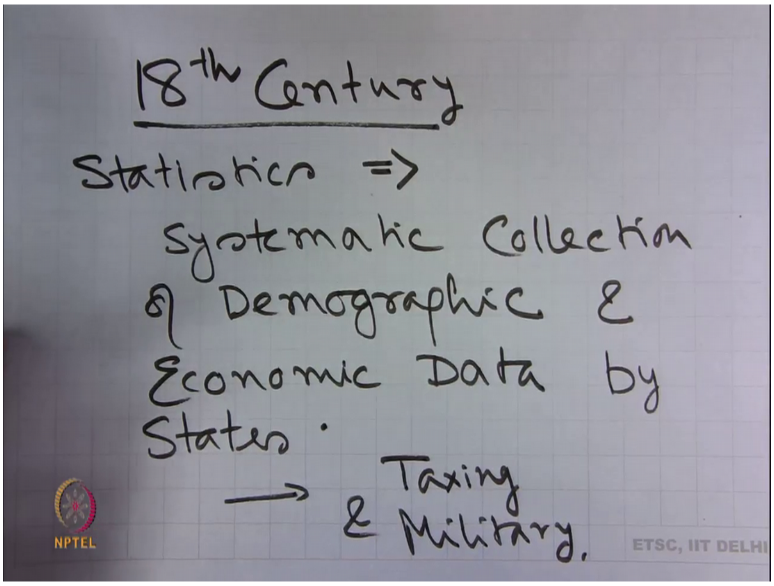
So, historically the term statistics was used first in English by Sir John Sinclair who is the Scottish politician; and he was a prolific writer he has written 21 volumes on Statistical Account of Scotland that was around the time 1791 to 99. So, it is more than 200 years that the word statistics is being used in English.

(Refer Slide Time: 03:34)



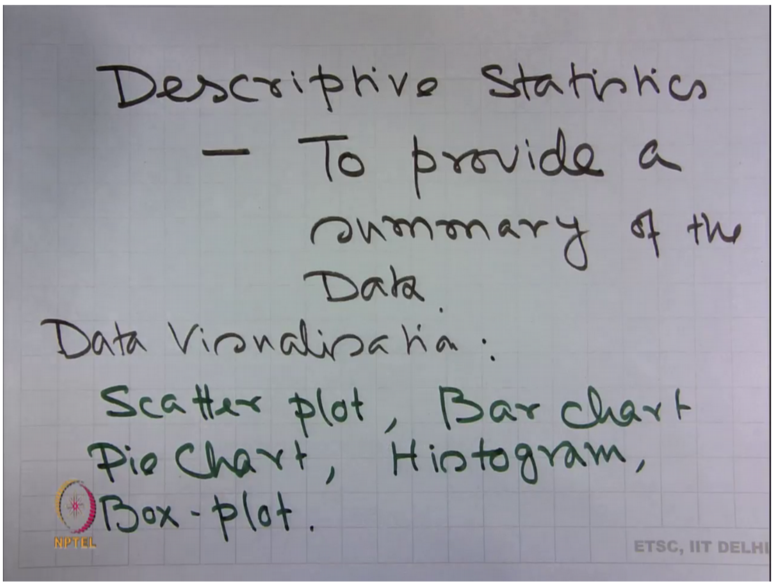
Originally people think that the word has come from a German word Statistik which means. So, at state level government was collecting data which was used by government and administrative bodies so that was around middle of the 18th century in particular it is around 1749 and the term was coined by Gottfried. So, it is more than 250 years old that the term statistics statistic or something related to that one that is invoked.

(Refer Slide Time: 05:21)



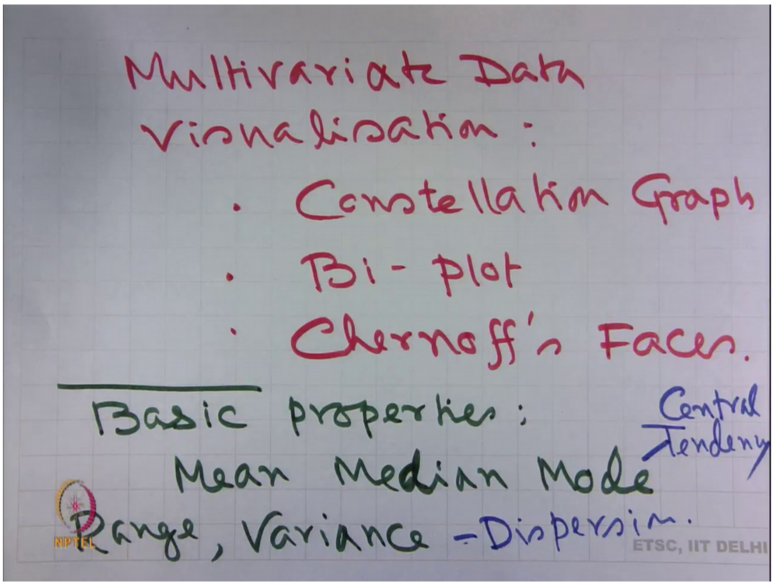
By the time 18th century statistics more or less stand for systematic collection of demographic and economic data by states. The basic purpose was taxing and military so that is the basic historical background of statistics.

(Refer Slide Time: 06:35)



So, statistics essentially has two parts; one is descriptive statistics which is basically to provide a summary of the data. I am not going to discuss descriptive statistics as I said that it is not the very first course of statistics. I assume people know some basics of descriptive statistics which may mean data visualization which is very very important for practical purpose. Because when you see it on a graph in a 2D or 3D, one can get much better intuitive idea of the data. And we have scatter plot you must be knowing all these things bar chart, pie chart, histogram, box-plot these are basic data visualization techniques.

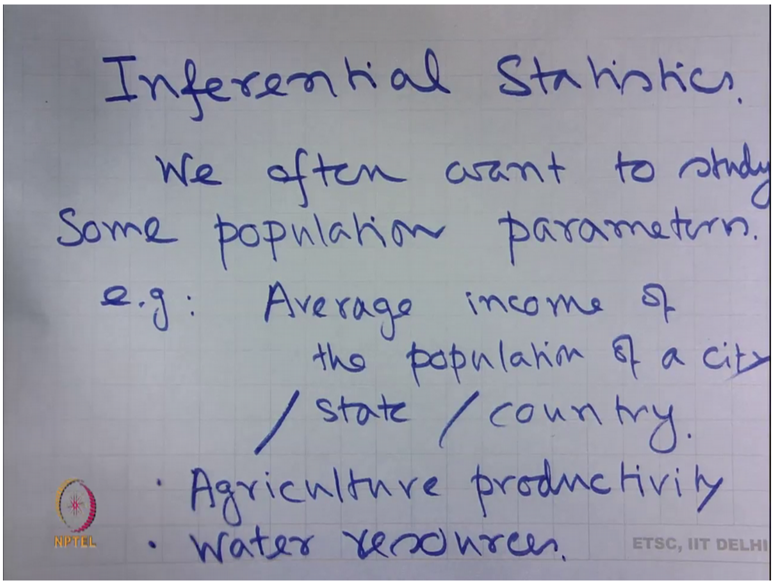
(Refer Slide Time: 08:24)



Also if you go for multivariate data, then one can visualize using many techniques, some of them are constellation graph, one can think of Bi-plot, one can think of Chernoff Chernoff's faces. Also apart from visualization, one can think of some basic properties such as mean, median mode.

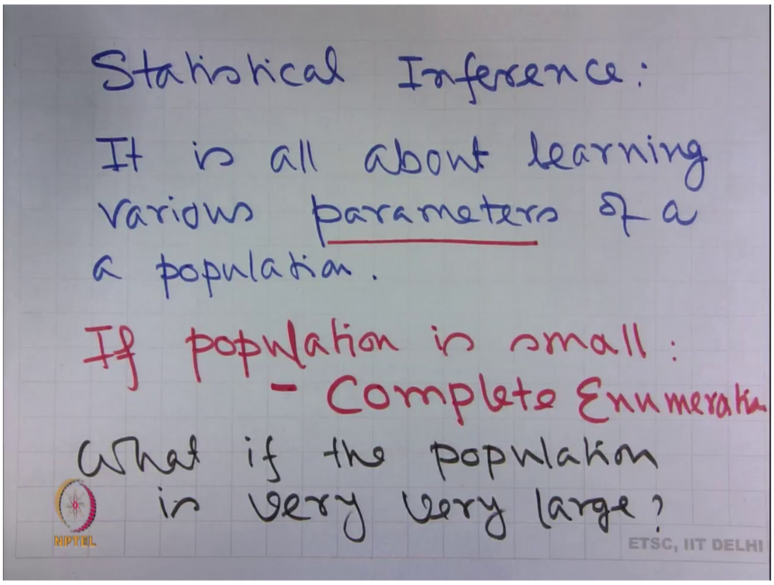
These are the central tendency. Similarly, one can think of range, variance, quartile deviation etcetera as a study study of dispersion. And similarly one can think of higher order movements like skewness, kurtosis etcetera. These are the techniques any statistician should learn for dealing with data because these are the basic processing of the data to understand what is going on there.

(Refer Slide Time: 10:35)



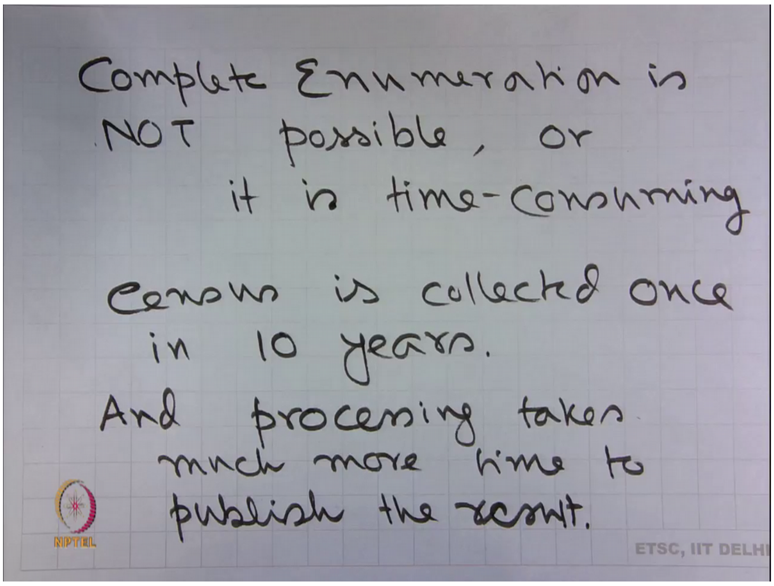
But statistics has an another purpose that is inferential statistics. We often want to study some population parameters. For example, you may likes to know the average income of the population of a city or state or country. We can think of the agriculture productivity, total water resources etcetera. How do you study this when the population is huge, what when the universe is huge, it is not possible to check each and every unit of it, and measure the relevant properties to come to overall figure with respect to the universe.

(Refer Slide Time: 12:41)



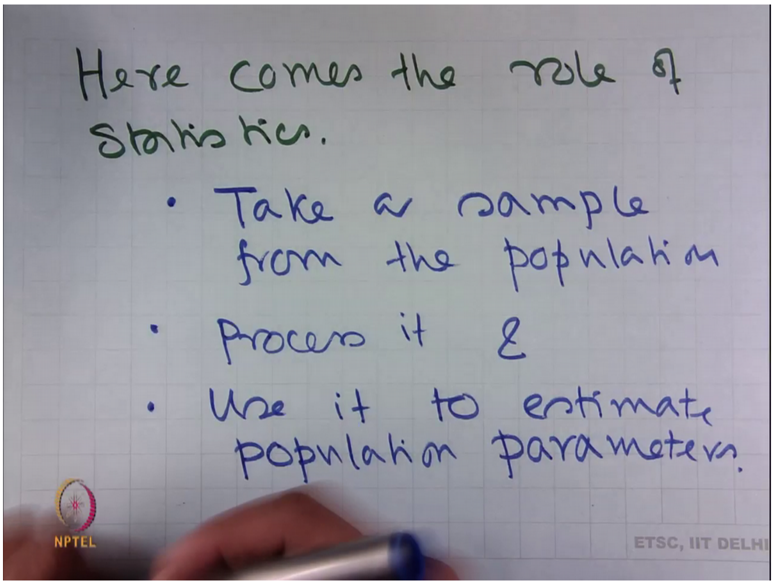
Here comes the utility of statistical inference it is all about learning various parameters of a population. If the population is small, one can actually study each and every individual unit and come to a conclusion about the population typically that is called complete enumeration complete enumeration. So, you are looking at all the members of the population, you are measuring the parameter that you are looking for which may be a weight, which may be height, which may be income which may be age. Similarly, you can think of the total volume of forestry in a country etcetera. If a population is small then it can be done very easily.

(Refer Slide Time: 15:02)



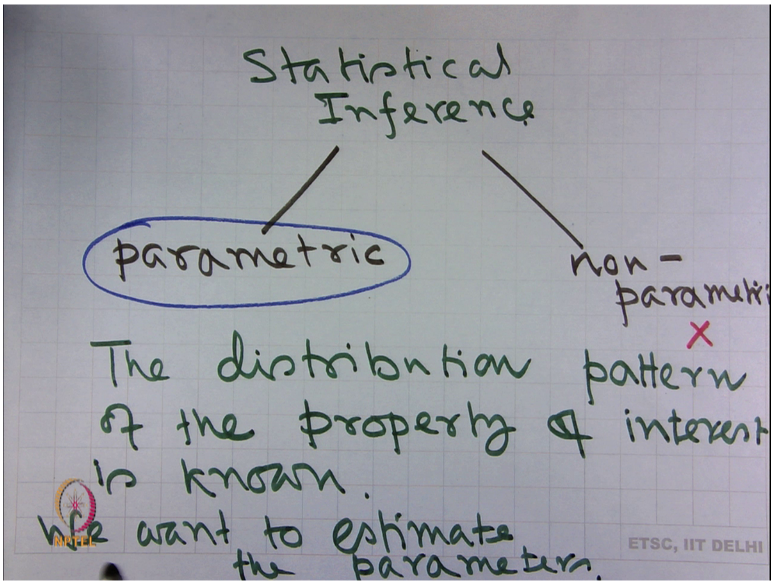
What if the population is very large, then complete enumeration is not possible or it is time consuming. For example, if you look at census data where surveyors actually go from household to household, and collect information about individuals and the household; together it is expensive and time consuming and that is the reason census is collected once in 10 years. And the processing takes much more time to publish the result. In practice that is not always affordable you cannot afford 10 years, 15 years to complete your study because lot of planning economic or otherwise have to be done in a much shorter span of time.

(Refer Slide Time: 17:00)



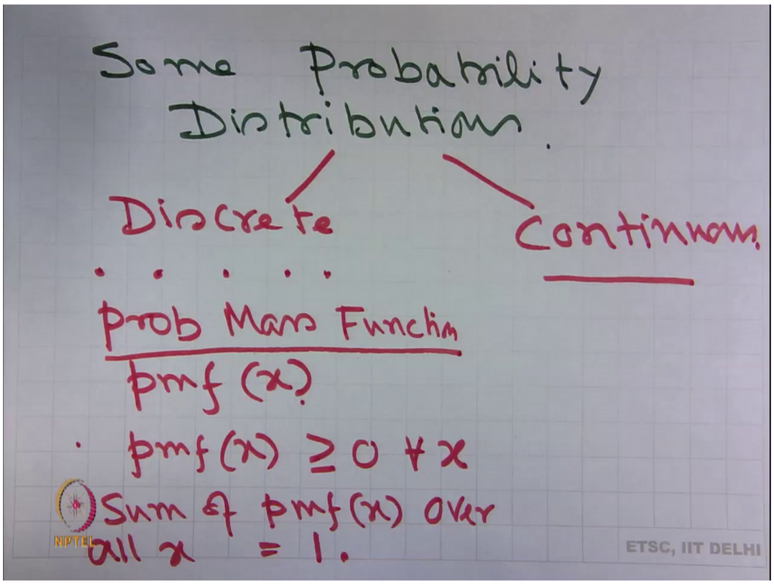
Here comes the role of statistics. So, what statistics will do take a sample from the population, process it, and use it to estimate population parameters. So, instead of a huge population considering completely, we will take a representative sample out of it, will process it. And from the results obtained after the processing we will try to infer about the whole population this is the science of statistics and in this series of lectures I will look into this aspect of statistics which is called statistical inference.

(Refer Slide Time: 18:41)



There are two basic approaches of statistical inference; parametric and non-parametric. In this series, I will be focusing on parametric inference; non-parametric I am not going to cover in these series of lectures parametric means the distribution pattern of the property of interest is known; and our job is to estimate the parameters. So, here comes the concept of probability, you must have studied different probability distribution, and you must have had the background of that one I will assume that much knowledge from your side.

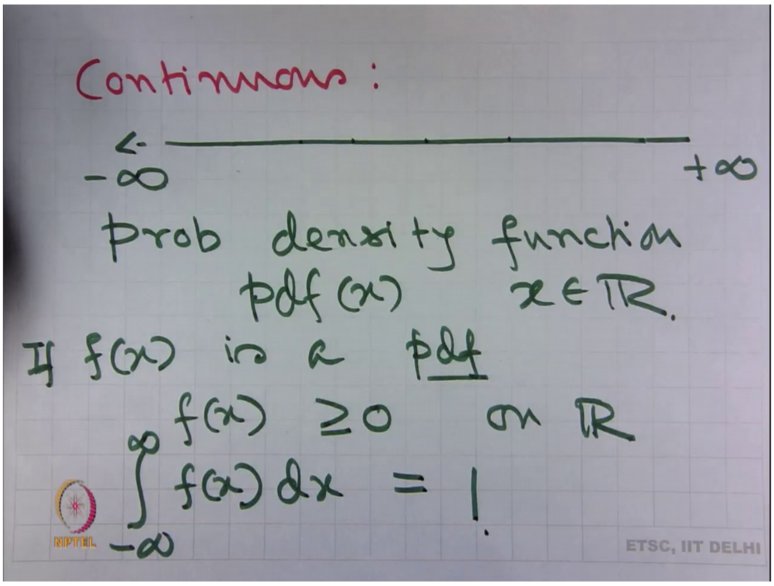
(Refer Slide Time: 20:51)



But to make it complete, I will first talk about some popular probability distributions. As you all know they can be of two types primarily there can be mixed also, but I am not considering that discrete and continuous. When the random variable takes discrete values, then we call it a discrete random variable and corresponding distribution is a discrete probability distribution.

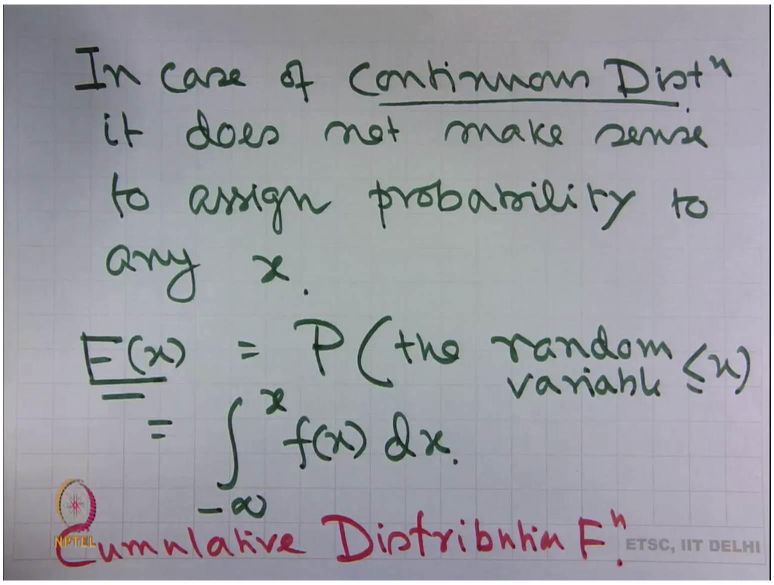
Otherwise if it is continuous range along the real line then we call it a continuous random variable with respect to discrete random variable we associate probability mass function in short we call it pmf of x where x is one possible value that the random variable can take. So, a pmf the basic properties is that pmf of x is greater than equal to 0 for all x and some of the values over all x is equal to 1. So, any discrete values, which are greater than equal to 0 and this sum up to 1, we can in principle consider that to be a probability mass function. It does not mean that any arbitrary selection of values which satisfy these properties can be modeled with some natural phenomenon, but for mathematical treatment we can consider that to be a valid pmf.

(Refer Slide Time: 23:32)



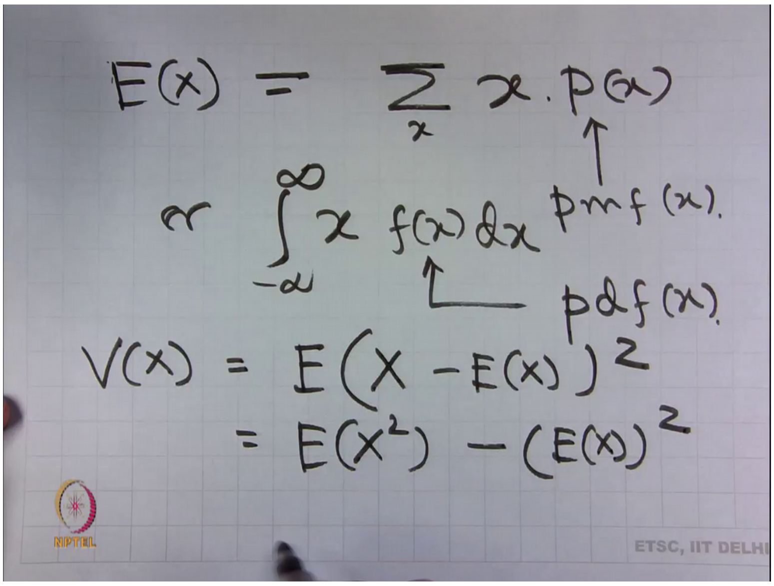
Continuous random variable we assume that it is spread over the entire real line minus infinity to plus infinity. And therefore, it does not make sense to assign a value to each one of them because thus total probability has to be 1. So, in this case, we talk about probability density function; in short pdf of x x belonging to R. So, if f x is a pdf, what if fx is a function which is the probability density function then f x has to be greater than equal to 0 on R everywhere, on R it is greater than equal to 0. And if you integrate it from minus infinity to infinity, that has to be 1.

(Refer Slide Time: 25:03)



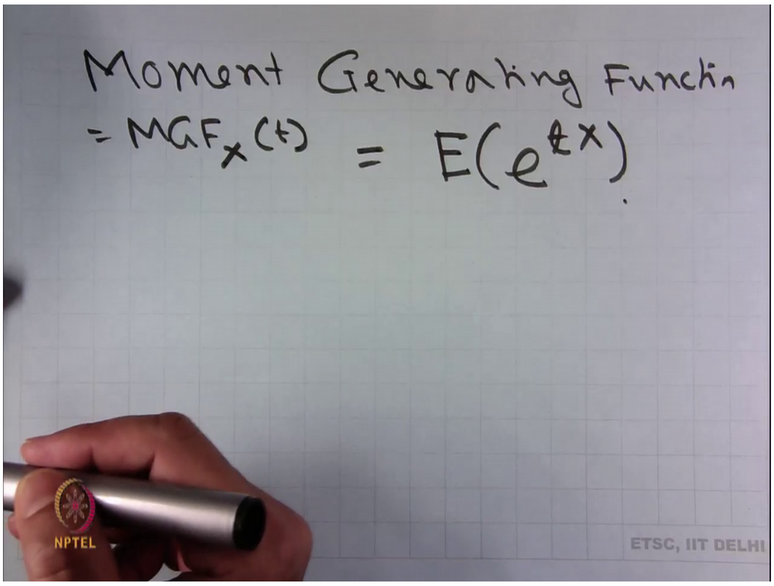
As you know in case of continuous distribution, it does not make sense to assign probability to any x. In fact, when we talk about continuous random variable or a continuous distribution, we look at f of x is equal to probability that the random variable less than equal to x which is obtained by integrating the probability density function from minus infinity to x. And this is called the cumulative distribution function. So, this is the cumulative distribution function. And this makes sense, this gives you the probability that random variable is taking a value less than equal to x.

(Refer Slide Time: 26:58)



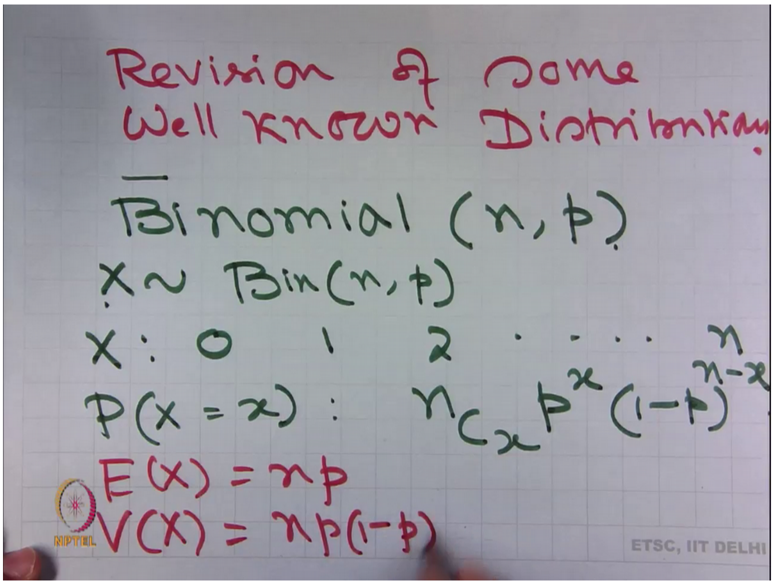
Corresponding to each random variable, we can assign expected value of x which is sigma over x x p x, p x is the corresponding probability mass function of x or it can be written as minus infinity to infinity x times f x dx, where f x is the corresponding pdf of x. As you all know variance of x is defined as expected value of X minus expected value of X whole square which can be written as the expected value of X square minus expected value of X whole square.

(Refer Slide Time: 27:57)



Also you know there is something called moment generating function which is called MGF of x for real value t is equal to expected value of E to the power e x this is called moment generating function because from here we can generate all the moments of the random variable. For example, first moment is expectation of x second moment is expectation of x square like that.

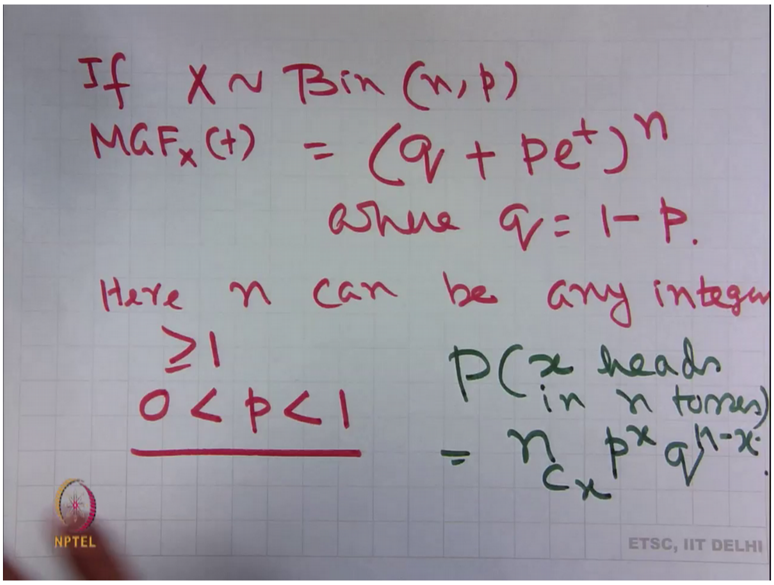
(Refer Slide Time: 28:47)



Now, let me revise some well known distribution all of you know. But if you are forgetting something, then you try to recap. Also I am not going to deal with all the discrete distributions that you might have studied, but it will be good if you have a revision of those random variables as well.

The first one that I look at is binomial. It has two parameters n comma p, it takes values if x is a random variable which is binomial with parameters n comma p then the possible values for x are 0, 1, 2 up to n. And probability x is equal to x that is the probability mass function at x is equal to n c x p to the power x 1 minus p whole to the power n minus x. These all of you should know. And you also know that expected value of x is equal to n p variance of x is equal to n p into 1 minus p.

(Refer Slide Time: 30:44)

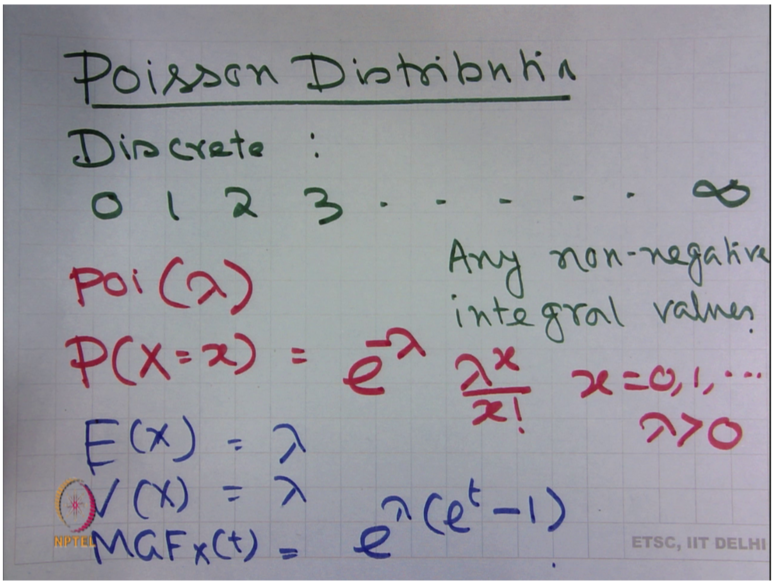


And if x is binomial n comma p then its MGF at t is equal to q plus p e to the power t whole to the power n, where q is equal to 1 minus p. Here n can be any integer greater than equal to 1, 0 less than p less than 1. So, for so the binomial distribution is defined for all n greater than equal to 1 integers; and for any value of p between 0 to 1.

As you know that binomial distribution is used to obtain, the distribution of the number of heads of a coin where probability of getting a head is p. And if the coin is tossed n times what is the probability that one will obtain x many heads. So, probability of x heads in n tosses right that probability is going to be ncx p to the power x q to the power n minus x so that is only a model. In reality when some experiment is going on n number of times where probability of success is p binomial distribution gives you the probability of obtaining certain particular value.

For example, if a machine is producing some items say nut bolts, they can be defective or they can be ok. Suppose, the probability of getting non-defective nut bolt is p and that machine has produced 10,000 many nut bolts in a day. So, as a producer or manufacturer, one may like to know how many of them are defectives or how many of them are non-defectives of course, there is nothing guaranteed it is a probability it is a random event and the probability can be estimated using binomial model.

(Refer Slide Time: 34:00)

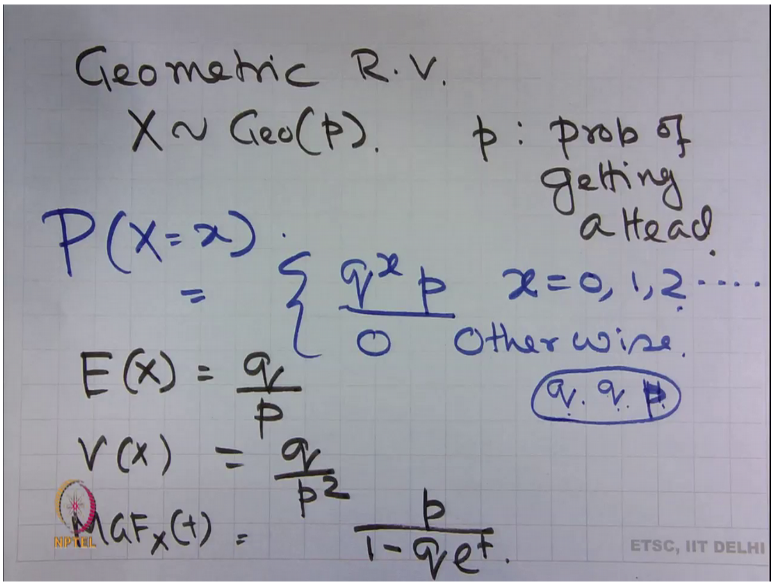


The next one is Poisson distribution. This is also a discrete distribution and it takes values 0, 1, 2, 3 up to infinity that means, it can take any non-negative integral values. Poisson distribution can be used to model the number of arrivals when there is a flow of incoming things. For example, the number of cars passing or say suppose there is a conveyor belt which is carrying the material the items produced by a machine, and then suppose the defective items are coming at a rate say 2 per minute, then what is going to be the expected number of defective items if the machine runs for half an hour.

Again this is a random variable, it is not fixed, but the number of defective items that are coming that will take different values different integral values and the probabilities can be modeled using Poisson random variable, it should have one parameter lambda. If we know the lambda then we should be able to know everything about the distribution. Probability x is equal to x is equal to e to the power minus lambda lambda power x upon factorial x x is equal to 0, 1 etcetera lambda greater than 0.

If you are recalling then you will be knowing that the expected value of x is equal to lambda; variance of x is equal to lambda; and the moment generating function of x the point t is equal to e to the power lambda into e to the power t minus 1. They are very easy to compute as it is not a first course on probability, I am not computing it, but it will be good if you revise these things. Another interesting point to note that for binomial random variable, variance was npq or np into 1 minus p. Since 1 minus p is less than 1, we knew that the variance of a random binomial random variable is less than the expected value. In this case, we can see that the expected value and the variance both are same.

(Refer Slide Time: 37:43)

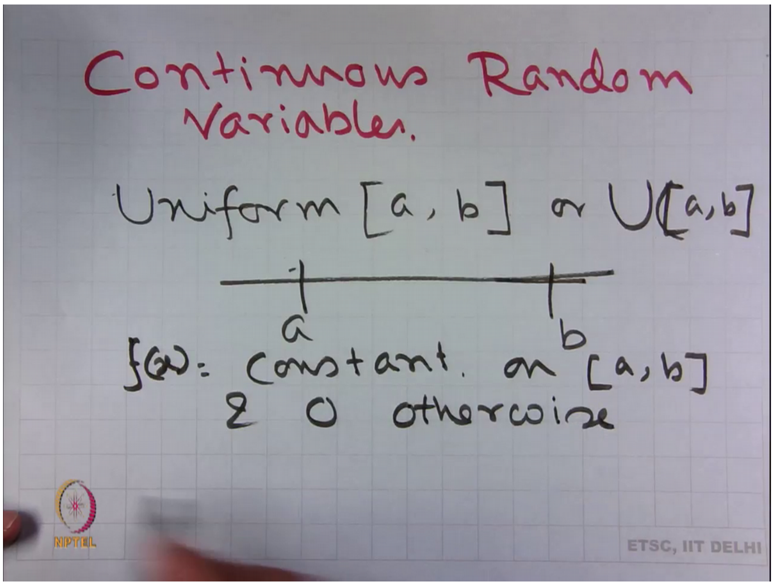


Now, I will consider another discrete random variable which is called geometric random variable. It also has one parameter p like binomial here also we look at tossing a coin. And p is the probability of getting a head. So, in geometric random variable, we want to study how much one has to wait to get one head.

So, probability x is equal to x is equal to q to the power x into p, where x is equal to 0, 1, 2 like that. And it is 0 otherwise; that means, suppose the geometric random variable takes a value two that means you have to make two tosses before getting a head. So, in the first toss, you got a tail whose probability is q; in the second toss, you get another tail whose probability is q. So, you have to wait for two tosses to get the head and its probability is p. So, overall probability is q square into p for x is equal to 2.

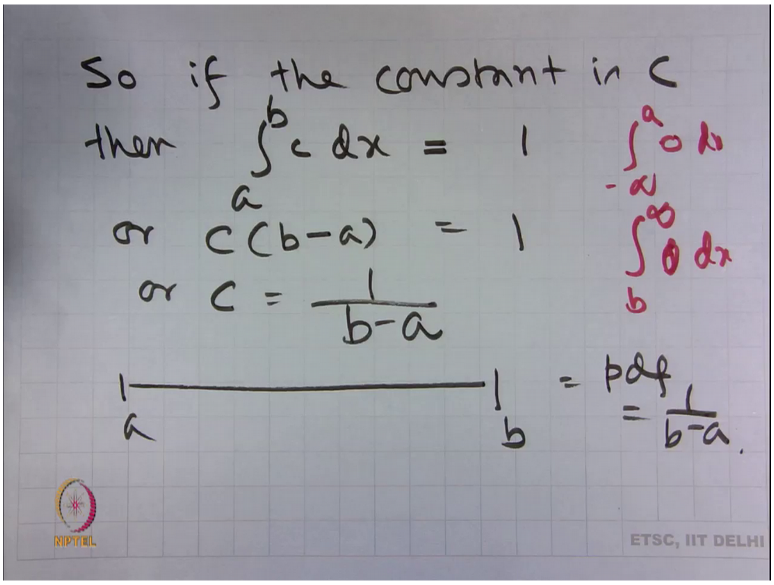
This is the basics of geometric random variable and it is used for modeling the waiting time. And as before I am giving you the values of expectation of X is equal to q by p; variance of x is equal to q by p square. And moment generating function of x, I suggest that you verify these results that will give you some practice of working on examples as in course of time I will give you some assignments, where you will be needing some practice of solving problems. And by solving this on your own you will get that required practice.

(Refer Slide Time: 40:36)



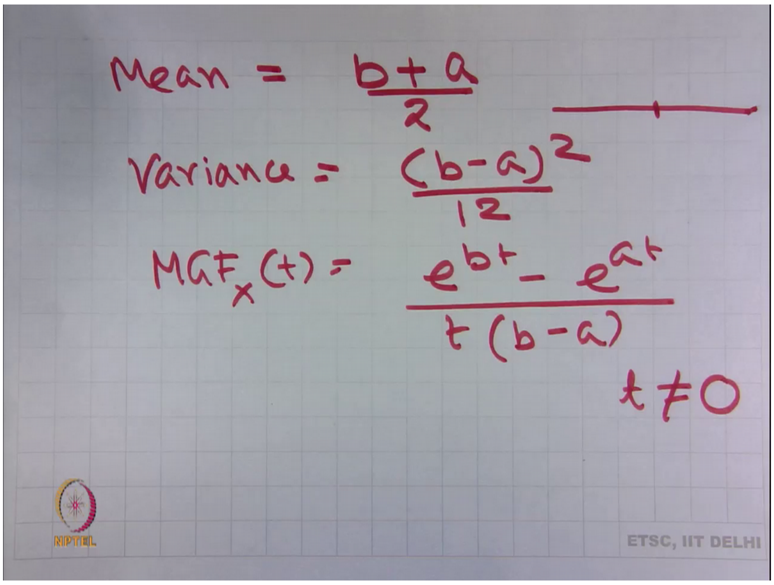
Now, let me look at some continuous random variables, the simplest one is perhaps uniform a, b or on the real line there are two points a and b. And the distribution of the random variable is uniform that means all values are equal likely therefore if we call it f x, then f x is basically constant on a, b and 0 otherwise.

(Refer Slide Time: 41:54)



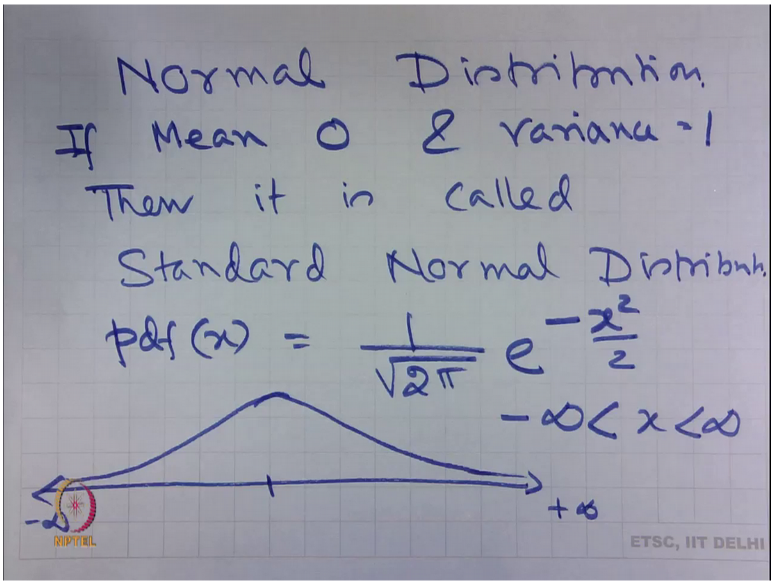
So, if the constant is c, then integration a to b c dx is equal to 1 or c into b minus a is equal to 1 or c is equal to 1 upon b minus a. Therefore, a uniform distribution on an interval a to b will have a constant density function is equal to 1 upon b minus a you may be wondering why I am integrating only from a to b why not from minus infinity to infinity that is because minus infinity to a in this region f x is 0 and also b to infinity in this region f x is 0. Therefore, when we add, when you integrate them they do not contribute to the overall integration.

(Refer Slide Time: 43:15)



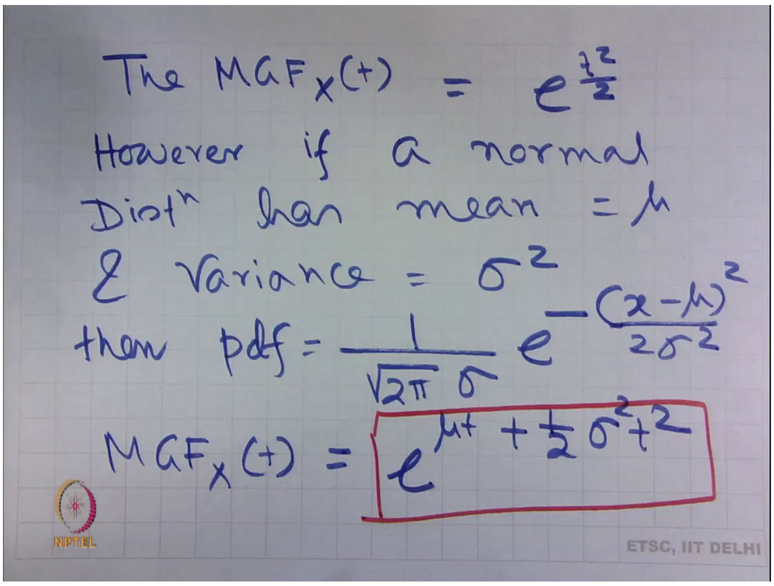
So, what is the mean of a uniform random variable mean is equal to b plus a by 2 or since all the points are uniform, it is the midpoint of this all right. The variance is equal to b minus a whole square upon 12. And the moment generating function of x at t is equal to e to the power bt minus e to the power at upon t into b minus a, when t is not equal to 0, so that is the basic properties of uniform distribution, the most well known continuous random variable is normal distribution.

(Refer Slide Time: 44:26)



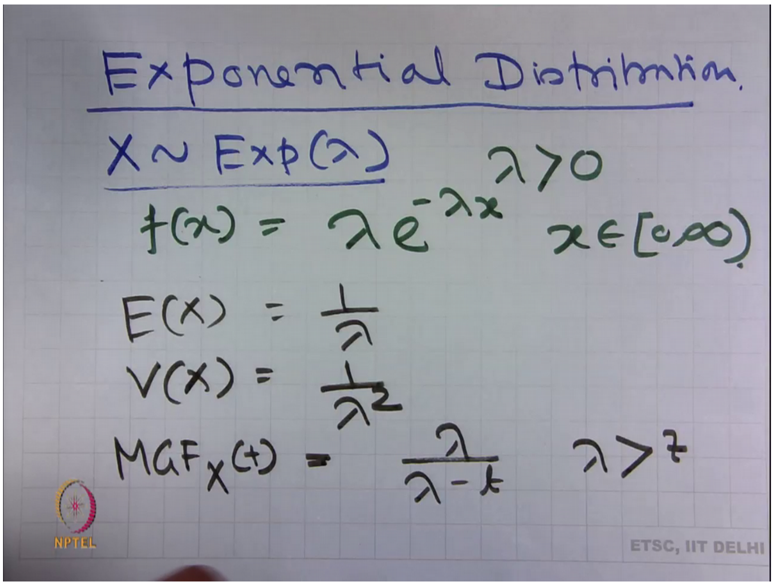
Typically, we have mean 0 and variance equal to 1, then it is called standard normal distribution pdf at x is equal to 1 over root over 2 pi e to the power minus x square by 2 minus infinity less than x less than infinity. So, you see that this is one continuous random variable that is defined over the entire real line. And if it is centered around 0 with variance is equal to 1, it is called a standard normal distribution. Many of you are familiar with a curve of this type which typically is used for standard normal.

(Refer Slide Time: 46:04)



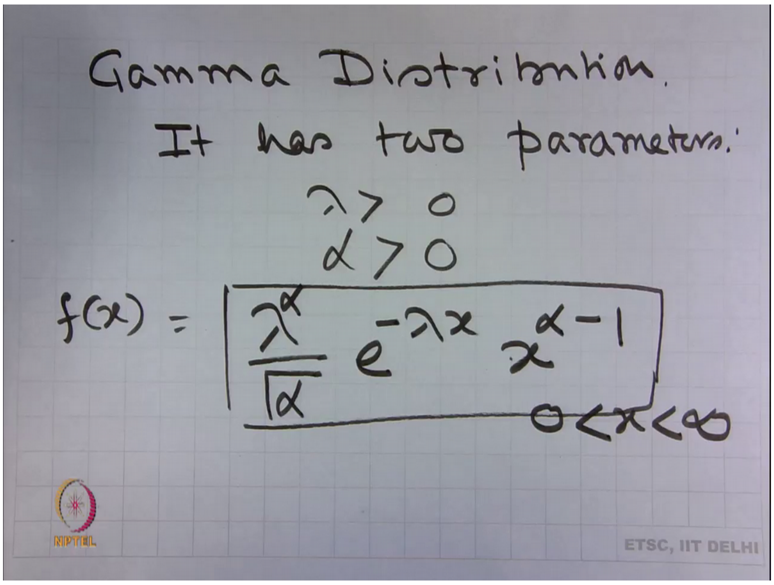
The MGF of x t is equal to e to the power t square by 2. However, if a normal distribution has mean equal to mu and variance equal to sigma square, then pdf is equal to 1 over root over 2 pi sigma into e to the power minus x minus mu whole square upon 2 sigma square ok. Mu can be any real number and sigma square being a variance of course has to be positive. And this is going to give you the pdf. In this case, the MGF is going to be e to the power mu t plus half sigma square t square. If you put mu is equal to 0, and sigma square is equal to 1, you get the moment generating function for the standard normal random variable.

(Refer Slide Time: 47:47)



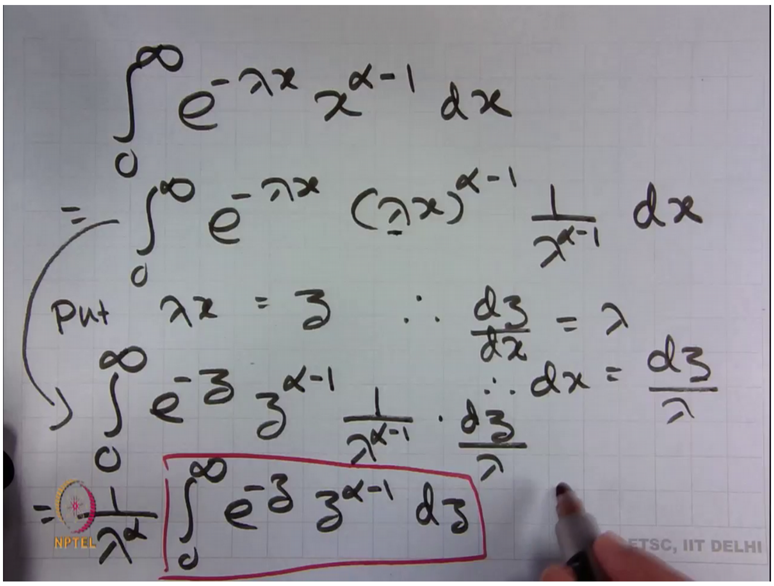
Our next focus is on exponential distribution. This is also used for modeling sum arrival, where lambda is typically the arrival rate. So, if x is a random variable which is following exponential distribution with parameter lambda where lambda is greater than 0 then f of x is equal to lambda e to the power minus lambda x for x belonging to 0 to infinity. It can be proved, or I will rather ask you to prove that expectation of X is equal to 1 upon lambda; variance of X is equal to 1 upon lambda square. And moment generating function of X at t is equal to lambda upon lambda minus t, of course it will be valid if lambda is greater than t.

(Refer Slide Time: 49:30)



Before I close I will give you one more distribution that will be used often in this course that is called gamma distribution. It has two parameters; lambda greater than 0 and alpha greater than 0. And f of x is defined as lambda power alpha upon gamma alpha e to the power minus lambda x x to the power alpha minus 1 0 less than. So, it is defined for non-negative x which is going from 0 to infinity; and this is going to be the corresponding PDF.

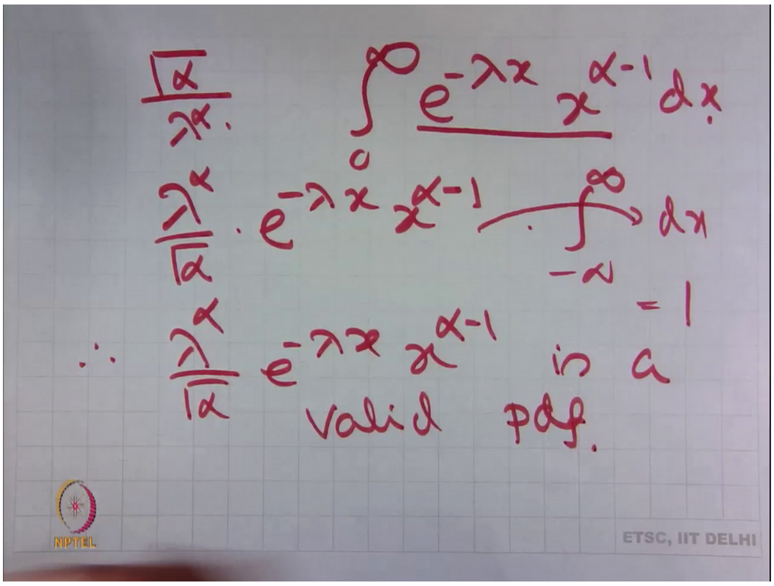
(Refer Slide Time: 50:45)



If we integrate this quantity e to the power minus lambda x into x to the power alpha minus 1 in the range 0 to infinity, then we can write it as 0 to infinity e to the power minus lambda x, lambda x to the power alpha minus 1 into 1 upon lambda to the power alpha minus 1 dx; I have used lambda power alpha minus 1. So, I have cancelled it put lambda x is equal to z, therefore, dz dx is equal to lambda therefore, dx is equal to dz upon lambda. So, this now I can write it as a is going from 0 to infinity, z is also going from 0 to infinity as lambda is positive.

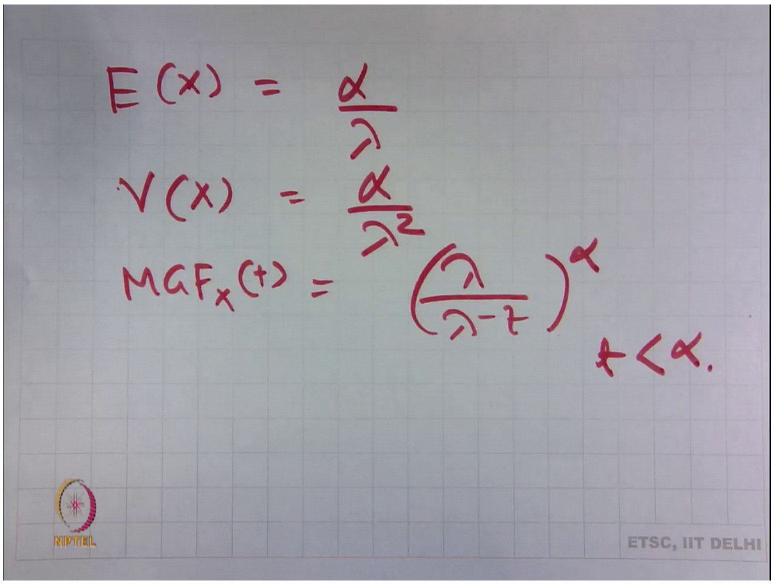
So, it is 0 to infinity e to the power minus z, z to the power alpha minus 1 1 upon lambda alpha minus 1 into dz upon lambda is equal to 1 upon lambda power alpha integration 0 to infinity e to the power minus z z to the power alpha minus 1 dz. If you remember your mathematics this is the famous gamma integral, and this is actually gamma alpha.

(Refer Slide Time: 52:36)



Therefore, the whole integration boils down to gamma alpha upon lambda power alpha. So, this is what we have obtained when I am integrating 0 to infinity e to the power minus lambda x x to the power alpha minus 1 dx. Therefore, if we multiply this by lambda power alpha upon gamma alpha, then we get this quantity which integrates to 1. Therefore, lambda power alpha upon gamma alpha e to the power minus lambda x x to the power alpha minus 1 is a valid pdf.

(Refer Slide Time: 53:32)



So, what are the mean and standard deviation expectation of X is equal to alpha upon lambda. Variance of X is equal to alpha upon lambda square. And moment generating function of X t is equal to lambda upon lambda minus t whole to the power alpha where t is less than alpha ok. So, these are some of the basic probability distributions or probability mass functions that I will be using during the course.

Also I will be using some distributions like chi square, t, f. In some of the subsequent lectures, I will derive those distribution what do they mean because I will be using them in later part of statistical inference when I will be doing testing of hypothesis. Thank you for your attention to this course.

Thank you.