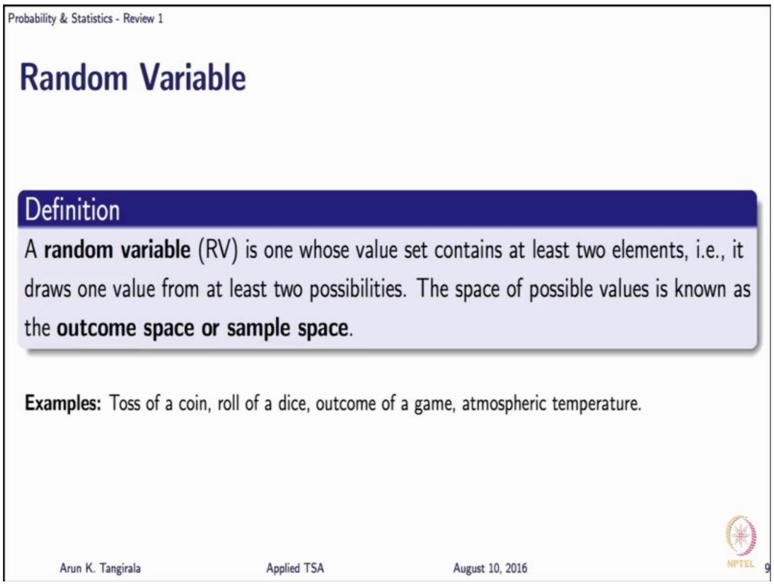
Applied Time-Series Analysis
Prof. Arun K. Tangirala
Department of Chemical Engineering
Indian Institute of Technology, Madras

Lecture - 08
Lecture 04B - Probability and Statistics Review (Part 1)-2

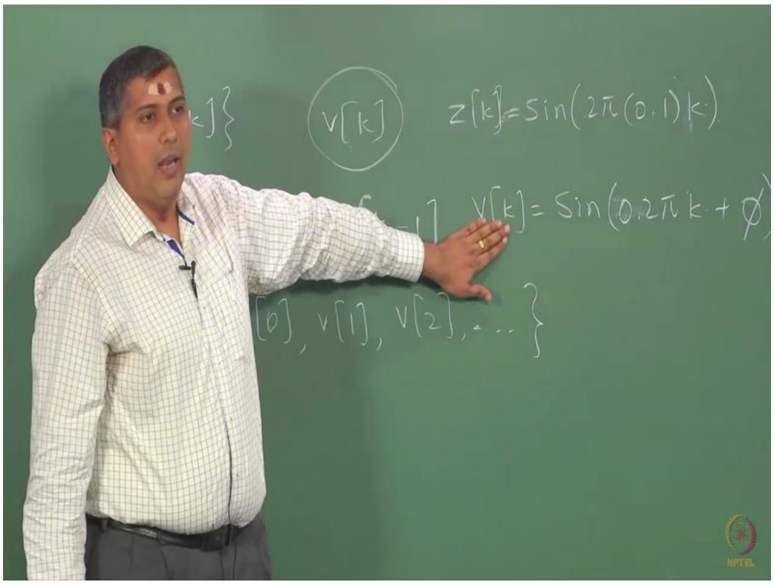
So, let us now begin with the review of the notion of a random variable. Now hopefully all of you have sat through the NPTEL course or even otherwise you are familiar with the notion of a random variable.

(Refer Slide Time: 00:24)



A random variable can be defined in two different ways, you can say that a random variable is that variable whose value set; that means, the set of possible values that it can take as more than one element whereas, a deterministic variable has only one possible value and importantly when we think of random variable. So, what is happening here?

(Refer Slide Time: 00:51)

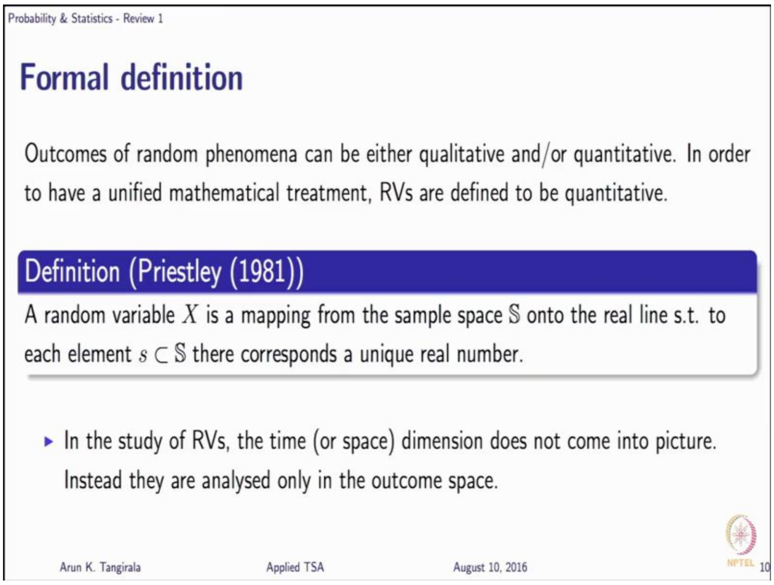


We said we are interested in a random signal, let us say our index is here 0, 1, 2 and so on and here you have v minus 1 and so on. So, this is your infinitely long random signal that will consider for all theoretical purposes. We are interested in analyzing this random signal, but at this moment now when we are reviewing the theory of random variables we are freezing time. We are saying let us now stand at a single instant in time and understand how the random signal is characterized alright. So, you are standing at some instant k right and asking how you characterize this and at any instant the random signal is a random variable.

That is one way of looking at; it is a very convenient way of looking at it. We will give a formal definition of random process and a signal later on, but right now freeze time, there is no notion of time at all we are only talking of a random variable and it is very important to get this clear in our minds.

At any instant, the random signal is a random variable which means at any instant this random signal they has many possible values that it could take. I keep saying this; this is not necessarily the truth, the truth is possibly that it can take on only one value; it is only our imagination because I do not know what value it will take, I am listing all possible values that is all.

(Refer Slide Time: 02:35)



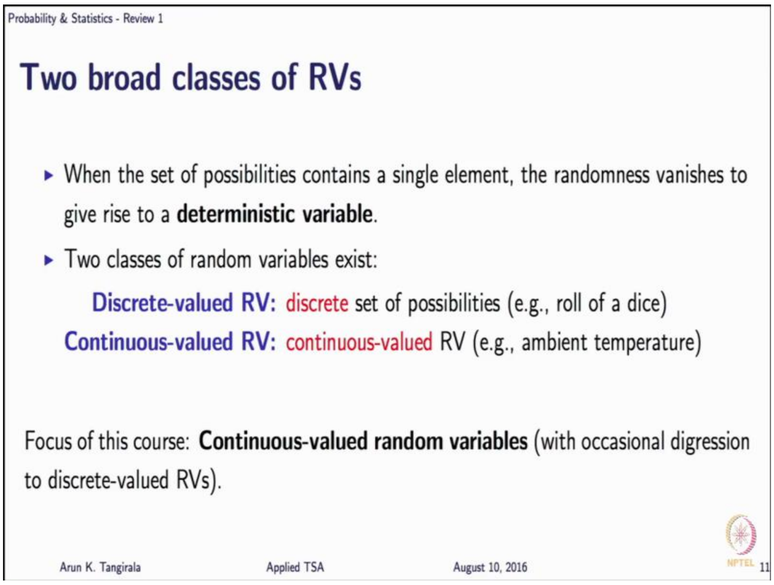
So, that is one way of looking at a random variable. The formal way of looking at a random variable is that it is a mapping from the outcome space, outcomes are all the possible values to the real number space which means the moment I say a variable is random; that means, it can take on only numerical values, we will not talk of qualitative outcomes at all. If you have qualitative outcomes it is your responsibility to map the qualitative outcomes to some numerical values and of course, that mapping can play a role, we will not talk about that at all.

We will assume that mapping has been done and proceed with random variables; obviously, the difference between a random variable and a deterministic variable is; at any instant in time I mean even if you just forget time, the deterministic variable will take on only one value and that is a only possible value for it. For example, if I say sin 2 pi times 0.1 right and let us say times k. So, this is I am sorry; let me use a different notation here let us say there is some z which is a deterministic variable. You cannot call this as a random variable because there is nothing random on the right hand side, it is deterministic, at any instant k, it can only take on one value alright.

On the other hand if I have a variable as follows let us say omega is known, let us call it as some 0.2 pi k plus phi. So, let us call this as v k, at any instant; v is determined by this function alright.

If the phase phi is deterministic then v is deterministic, there is nothing random about it at all. On the other hand, if the phase has randomness in it, if a phase is a measure is telling you when the signal, how the signal is position with respect to the 0 reference axis or you can say it is a lag, if that phase is random then that imparts randomness to be and then it becomes a random variable or you can even say random signal. Of course, I am still maintaining the time dependence here because ultimately we will handle random signals, but what I am trying to tell you here is you should be a priori convinced that a random variable is indeed random and you should be clear in your mind what is the source of randomness, we will come to that with an example later on. So, formally a random variable is a mapping from your qualitative space to the numerical space.

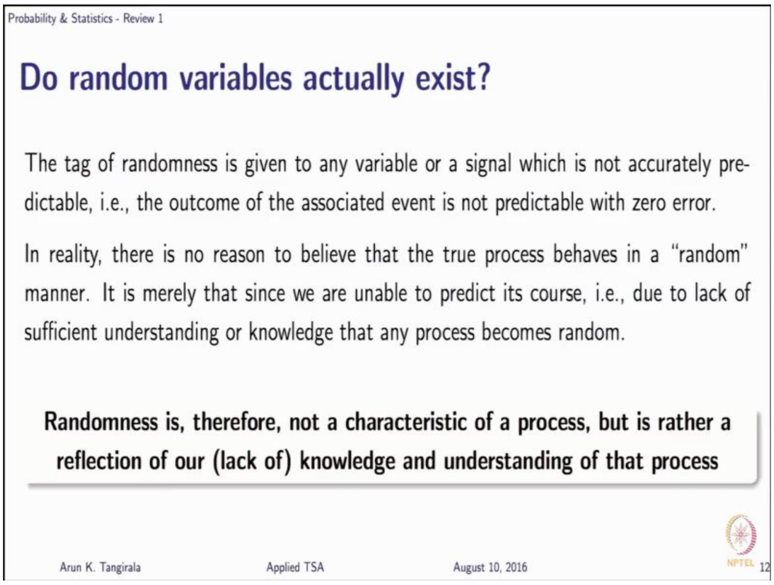
(Refer Slide Time: 05:26)



And there are generally two broad classes of random variables as you have all must have learned the discrete value random variable and a continuous valued random variable. The focus in this course is on continuous valued random variables, the reason for focusing on continuous valued random variables is most of the times we would run into situations where the signals that I am looking at are continuous value. For example, temperature, pressure or even stock market index or rainfall anything that you take are mostly continuous value.

There are of course, some situations like salaries for example, you know they are discrete value, populations are discrete valued and so on in which case we will give special attention, but by and large we will be dealing with continuous valued signals and therefore, the focus is on continuous valued random variables.

(Refer Slide Time: 06:33)



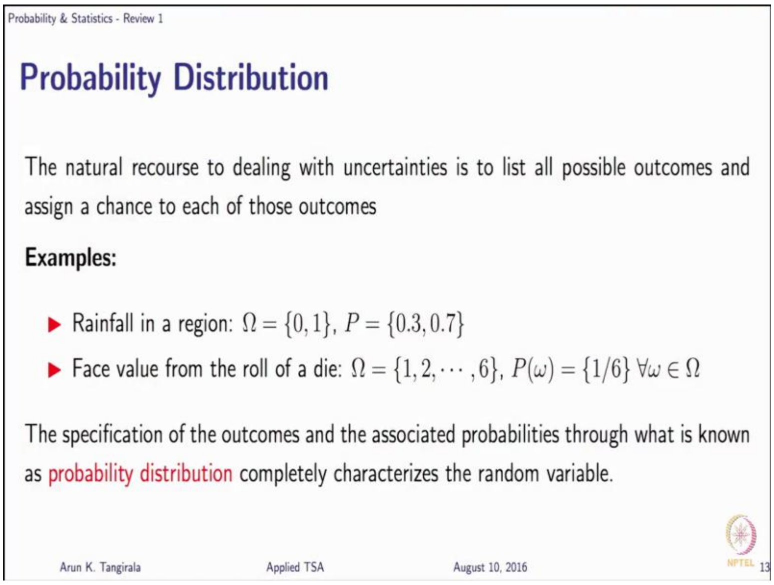
So, in passing I also want you to think and keep asking again and again until you are convinced of the philosophy, of this entire theory of random variables. This question that I want you to think is do random variables actually exist, we never know. You may say yes and I may say no and we can keep arguing forever, but the fact is that even if in reality the process is deterministic because of uncertainty; because of our ignorance, we are assuming the process to be random.

In the eyes of the creator maybe the world is deterministic, but from the eyes, from the human viewpoint no there are many possibilities and we think we have actually touched upon one possibility and we keep living in this dream world sometimes that could have been possible, no I could have obtained this grade or maybe you know I could have probably had gone a bit earlier, I would have eaten that dish that got over and so on but we never know maybe it was all predestined, you were supposed to go there at that time and so on. But if you resign to such a conclusion then unfortunately there is a danger of becoming very lazy and not doing anything.

Therefore, uncertainty is important for us to keep moving forward to keep aspiring for better and higher things. So, let it be that way right, but do not keep on the other hand thinking that there are infinite possibilities when that can also lead to depression. So, be careful, there is a balance that is required. So, randomness is not necessarily a characteristic of the process, it is our ignorance that we are blaming the process to be random saying how this process is so complicated is random. It is not complicated; our ignorance is so bad that we think it is complicated. So, it is very convenient for humans to blame on other things and that is what we keep doing. So, it is a euphemistic term for ignorance, so let us get back to the math from philosophy.

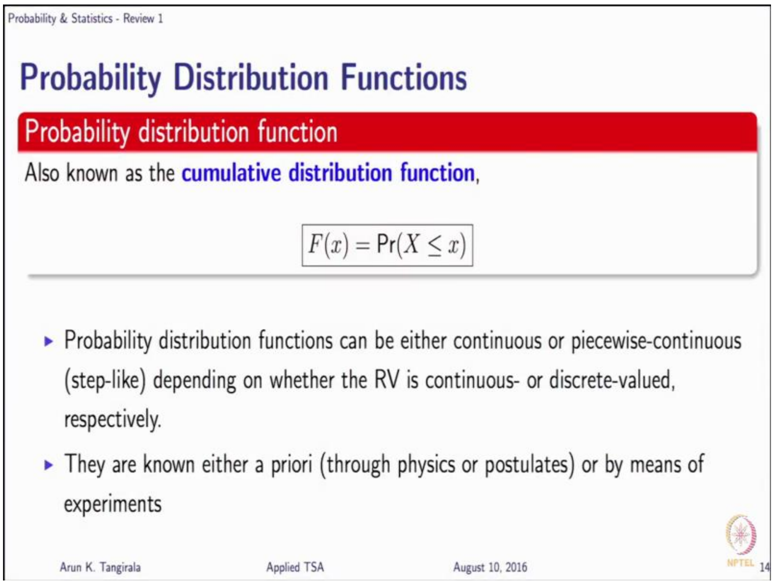
Now, that we have decided that many processes are complicated and we are going to treat them as random. Natural recourse is to list all the outcomes and assign chances to it which we call as probabilities and the distribution of probabilities across outcomes is what we call at least qualitatively as probability distribution.

(Refer Slide Time: 08:49)



And formerly the probability distribution is denoted as I have said earlier with this upper case F.

(Refer Slide Time: 08:52)

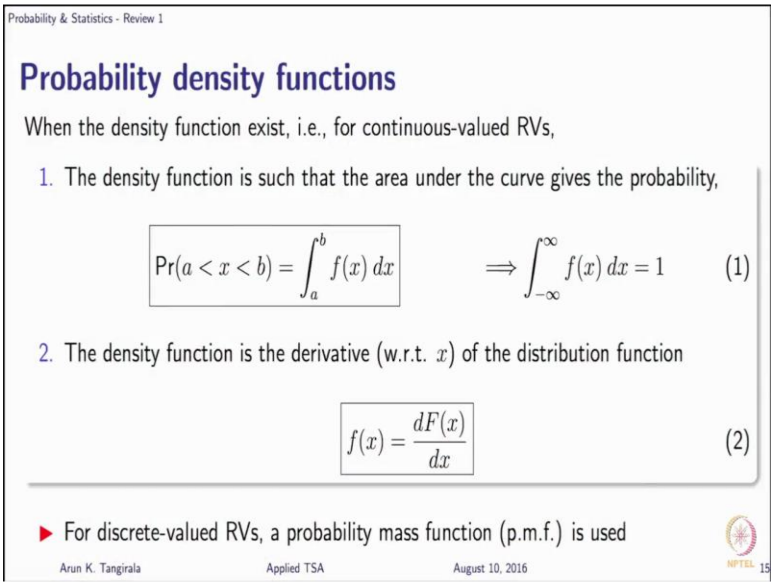


So, for a random variable x, the probability distribution is denoted by big F of small x notice that we are using small x which means it is actually a function; there is nothing random about this function; we remember that. F of x is for all practical and theoretical purposes a mathematical function, there is nothing random about that function; the randomness is with regards to the variable alright and it is also known as a cumulative distribution function because of the way it is defined. It is the probability distribution function is a probability of x taking on values less than some pre specified value and of course, you can have probability distributions that are either continuous or discrete in the sense let me say piece wise continuous or step like functions; depending on whether you are dealing with a continuous valued random variable or a discrete valued random variable.

I am kind of keeping things to a less rigorous way that is necessary for us to keep moving forward. If some of you are thinking that no I should be talking of boreal sets and so on and maybe probability measures and so on, unfortunately we will not use those terms. Remember boreal consists the most part of boreal set is b o r e, so I can end up boring you, this is with due respect to boreal, but we do not want to get into that, just keep moving forward. And if you look at a rigorous probability theory if you take a book solely devoted to probability theory, tell you that probabilities are actually measures. So, you have to understand the entire theory of at least the basics of measures to be able to understand probabilities and their measures on boreal sets and so on, but will not get into that.

So let us learn whatever is required and keep moving forward. Now do I know the probability distribution for a given random phenomenon; a priori, mostly no. There are some class of processes for which I can theoretically derive the probability distribution perhaps, but most of the times it is derived by means of experiments. So, one has to perform experiments and then look at all possible outcomes and try to find out how many times a particular set of outcomes have occurred, draw the histogram; you remember your high school tally tables kind of things and then you derive the probability distribution empirically, but we will not learn how to do that in this course. I am just telling you that most of the times, the probability distribution functions are derived in an empirical way. Now since we are going to deal with continuous valued random variables, it is convenient to work with probability density functions.

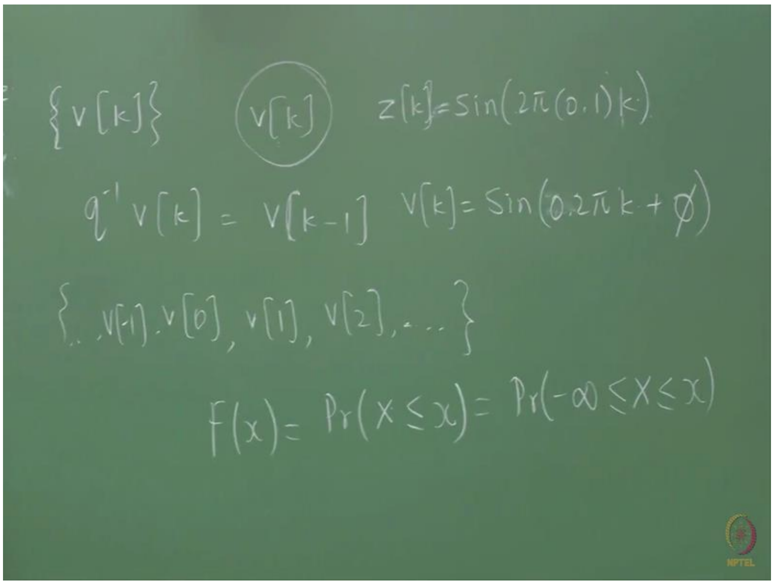
(Refer Slide Time: 11:48)



And I keep saying this for those of you who are still uncomfortable with probability distribution and the notion of probability density. Hopefully you are comfortable with these notions in mechanics where we talk of mass distribution density and so on and it offers a very good analogy, if you have understood that well then this should not be so much of a problem. In mechanics we talk of densities for objects that have masses for distributed over a continuum, when the mass is contiguously distributed then only we speak of densities. Likewise here, in the case of random variables when the random variable is defined over a continuum and the distribution exists and so on then you can define a probability density function. The probability density function is also abbreviated as pdf distribution functions are also abbreviated as pdf, but to avoid the confusion we will use cdf for distributions because there is another term to it cumulative distribution function and pdf for density functions.

There are two ways one can define probability density function and those are shown on the slide for you. One way is to define in such a way that the area under the density gives you the probability and the other way is to define it as the derivative of the cdf. Both give you identical results; do not worry about pathological situations where things break down and so on. Now one point that I want to make is the misconception that generally prevails among many beginners in probability theory.

(Refer Slide Time: 13:38)

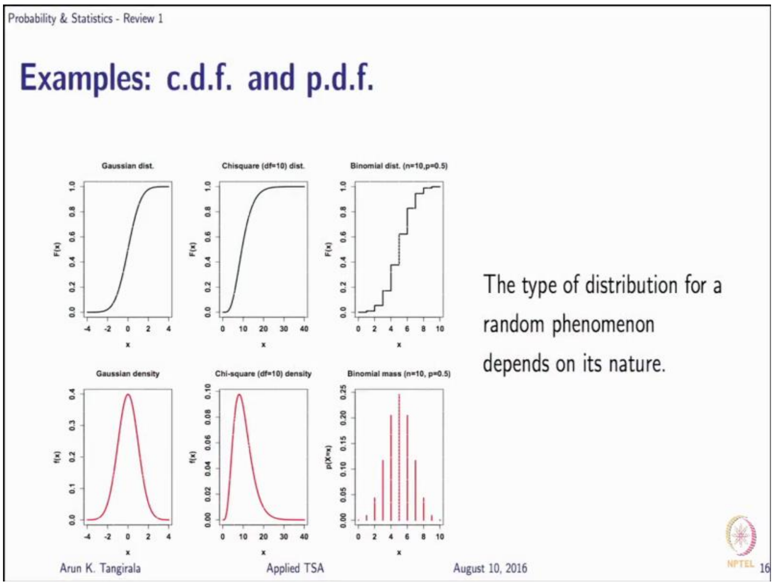


Which is that this small f of x that is the density function is the probability of x taking on a value pre-specified value small x; this is wrong. For continuous valued random variables this interpretation is absolutely wrong, it is as bad as saying the density, if I give you mass density let us say per unit length, it is the mass at that point, we know from mechanics that mass how is it defined it at any for any point because mass is also measure actually.

So, it is incorrect and inappropriate to speak of probability of x taking on a specific value x, when it is continuous value. You can talk about it when it is discrete value in which case we talk of probability mass functions, but for continuous valued, random variables please (Refer Time: 14:38) this interpretation, you can only look at the area under the density; however, small that interval may be infinitesimally small, but it has to be an interval and the length of that interval cannot be 0.

So, you can say that the probability for example, that x takes on in values in an interval x plus d x sorry x 2; x plus d x is approximately f of x d x. This approximation is Ok, there is nothing seriously wrong with it and it gets better and better as the interval becomes smaller and smaller, but it cannot be 0; the length of the interval cannot be 0. So, always use density functions to calculate probabilities and of course, will use density functions to derive moments and so on, but this is how you should use density functions whenever you are interested in computing probabilities; do not evaluate density function at a specific value and say that is the probability that x will take on that value; that is absolutely wrong.

(Refer Slide Time: 16:01)



So, just to give you a few examples before we move along. So, shown on the top of this figure that you see are the some three very popular distribution functions that you see and at the bottom are the corresponding density and mass function. So, the first two correspond to continuous valued random variable case and the third one corresponds to the discrete valued case, I keep using continuous time; continuous valued case and the discrete valued case. As you can see, the distributions for the continuous valued ones are continuous where as the distribution for the discrete valued one is a step like function and the names of the distributions are given for example, the first one is Gaussian and the last one is for the what is it are you able to read the name.

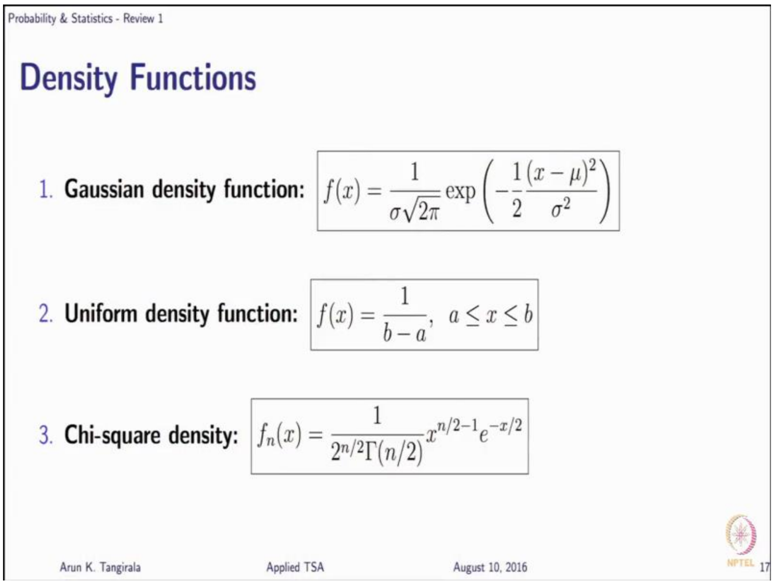
Student: Binomial.

Binomial very good, so you can see that although we do not have the notion of a density function for the discrete valued case, we have the notion of a probability mass function. In the case of discrete valued random variable, it is okay to ask this, so you can ask what is the probability this is for discrete case.

In which case you do not call it as a density function at all, you call it as a mass function. So, the question is now do I know the distribution or the density function for any random phenomenon a priori because that is a big piece of information. Remember for any random variable, the moment the probability distribution or let us say the density function for the continuous case is defined, you have kind of you have characterized the uncertainty completely; it does not mean that becomes predictable, but you have given all the information that is required to describe that random variable alright.

So, the golden piece of information that you require for a random variable is either the pdf or the cdf. do I have that priori unfortunately the answer is no, do I know at least the shape of it or the type of the distribution, maybe yes that at least has been very nicely studied for example, Gaussian distributions is associated with a certain class of phenomena; Poisson distribution is associated with some other phenomena for example, the number of accidents that occur in industry or you know the number of vehicles going past you when you are standing at some point in the road. So, that the Poisson distribution describes nicely so phenomenal it does not mean that it is going to give you an accurate description remember, the moment you have the pdf or cdf; you can draw many other inferences, you can make predictions those predictions are not going to be accurate, but they are going to help you that is the point.

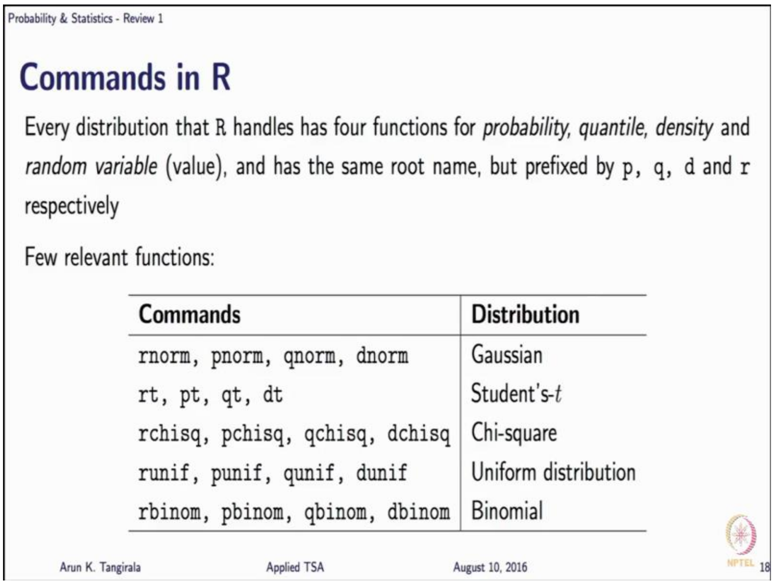
(Refer Slide Time: 19:04)



So, a lot of times this distribute type of distributions may be known, but you do not know the so called parameters of the distribution, what we mean by this is for example, if you these are; I am showing you just three of the many popular density functions. If you take the Gaussian density function there are two parameters mu and sigma, you may know for a given process that a random process or random phenomenon that the associated distribution is Gaussian, but you may not know mu and sigma which have to be estimated experimentally or you may know that some process has uniform density function in which case it is kind of easier you just need to know the intervals of the outcomes and so on.

In many other situations, we may not even know the type of distribution there are enough processes out there for which we do not know the distributions; in which case we kind of fit a distribution to the data and we will not get into that in this course in a typical statistics course probability and statistics course you will be perhaps taught how to fit distributions. In this course predominantly we will be dealing with these three density functions and by and large will assume that the observations that I am dealing with are coming out of a random process that follows a Gaussian distribution; strictly speaking joint Gaussian distribution, but I am not described what a joint Gaussian distribution yet is we are still in the univariate case even when it comes to random variables. So, to summarize I may know the type of distribution, but I may not know the parameters or I may not know the distribution at all in which case I have to discover it from data.

(Refer Slide Time: 20:32)

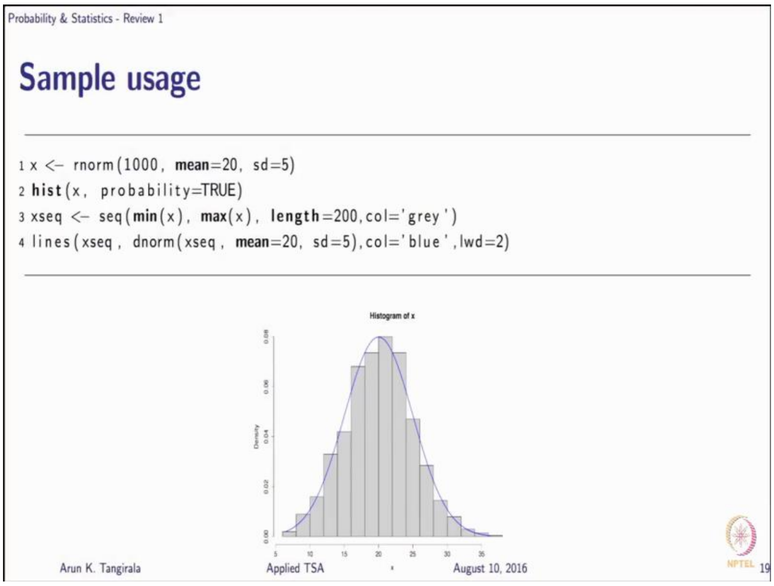


Now, there are these commands in r that will help you to visualize a lot of these distributions, this is not exhaustive list of commands, but quite useful corresponding to some of the most popular ones that you will come across and you can see for every distribution there are four different types of commands you have r, p, q and d, the r norm for example, gets you a random observation; observation at random drawn from a random variable that from that follows a Gaussian distribution; that means, there is a random variable for which there are many possibilities. But those possibilities are distributed in a Gaussian way and r norm will get you either one number or as many numbers as you want from the distribution it will essentially sample that for you; p norm is the other way round it gives you the probability when you specify, when you want to know; what is the probability that this random variable which follows a Gaussian distribution will take on values within a specified interval.

And q norm will give you the quintile value itself that is you specify the probability that is remember that your f of x is defined as a probability that x takes on any value less than or equal to x. It is understood that you are looking at the left extreme which is typically denoted by minus infinity up to the value that you specify.

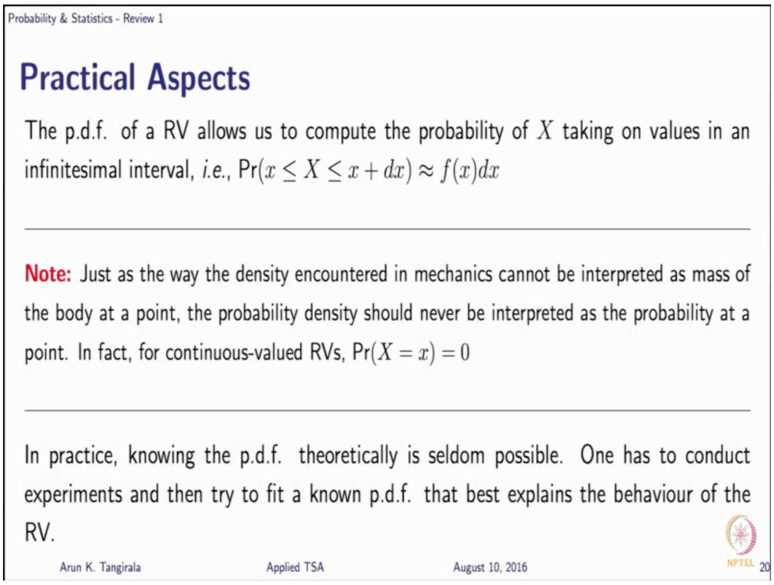
That minus infinity should not be thought of as minus infinity, it essentially it is saying the left extreme value for x. So, your q norm will help you figure out what is the sets for a specified probability, but you should look up the help on q norm whether it is actually looking at this cdf in some software packages, it may look at the complement of f of x. So, you have to see whether it is evaluating the probability from the left extreme to the point of interest or the right extreme to the point of interest and finally, the d norm itself will give you the density f small f of x. So, if I want to know the value of density function, the Gaussian density function at a specific value of x then d norm will help you and likewise for all other distributions. So, you should be quite comfortable with all these four because we will use all these four in some fashion or the other.

(Refer Slide Time: 23:04)



So, this is a sample usage of how to use this commands and what I have done is I have generated thousand numbers, I have drawn thousand numbers at random from a Gaussian distribution with pre-specified mean and standard deviation and then I have drawn the histogram of this thousand numbers that I have generated and it kind of indicates a Gaussian shape, I have also fit a Gaussian shape there and then of course, made the plot a bit more colorful for you.

(Refer Slide Time: 23:39)



So, just a sample usage, what will do tomorrow is move from the world of pdf s to the world of moments which is what we will keep working with in practice it is; remember we said we do not know the pdf it has to be derived empirically from experiments. But estimating pdf’s from data accurately and reliably is not an easy task, one needs large amounts of data and good estimation algorithms which are not yet available today. Yes we have improved, but still estimating pdf’s is a lot less accurate and reliable as compared to estimating so called moments what remember moments is mean variance and so on.

And the good news is that as far as linear random processes are concerned I do not have to worry about the pdf, I can be content with the knowledge of these, so for the first two moments mean and variance and of course, when you take this into the random signal world, we will talk about covariance auto covariance and so on. It turns out that for linear random processes that is more than enough only when you talk of non-linear random processes, we will have to look at the pdf’s or higher order moments and so on. So, tomorrow will review the moments and will also look at the bi-variate distributions and so.

Thank you.