Engineering Metrology

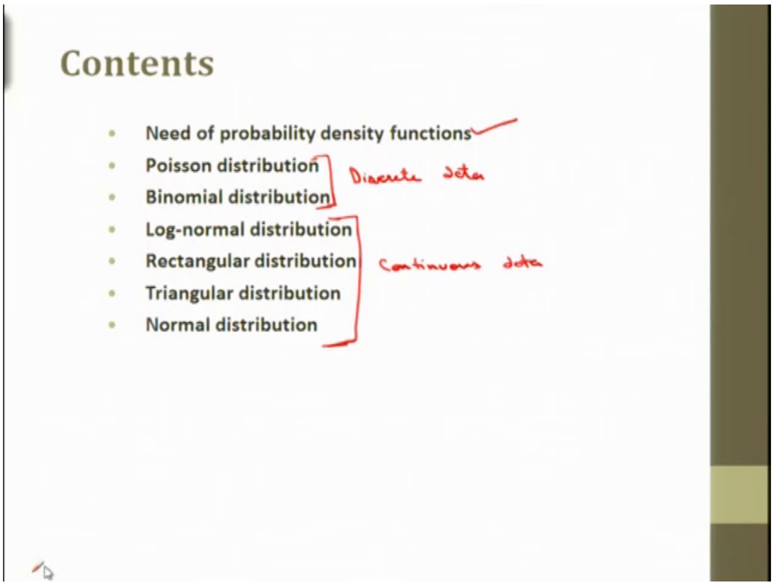
Prof. J. Ramkumar
Dr. Amandeep Singh Oberoi
Department of Mechanical Engineering & Design Programme
Department of Industrial & Production Engineering
Indian Institute of Technology, Kanpur
National Institute of Technology, Jalandhar

Lecture - 42
Probability distributions

Good morning. Welcome back to the course on Engineering Metrology and we have discussed the statistical part that is used in engineering metrology. In this lecture I will discuss about probability distributions.

I have given a brief information on what are frequency distributions, how do we construct a frequency distributions. In metrology what happens there are multiple the distributions which are used or not in any even in metrology like in statistics the probability distributions are there.

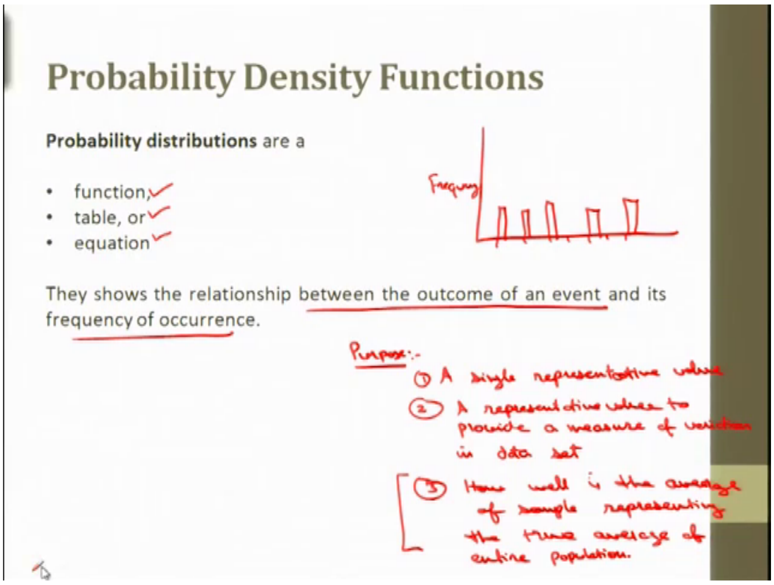
(Refer Slide Time: 00:48)



But the contents I like to cover here would be the need of probability distribution or we also called them probability density functions. Then I will discuss a few distributions Poisson and binomial distributions, the log normal, rectangular, triangular and normal distribution.

In this lecture I will just introduce these distributions and I will discuss more about the normal, binomial and chi square distribution that I will discuss in the next lecture and also try to solve some problems to have the deep understanding of how the distributions are used.

(Refer Slide Time: 01:25)



Now, what are distributions? Probability distributions as I said in the previous lectures is a function or a table or it can be an equation that shows the relationship between the outcome of an event and its frequency of occurrence. So, what is the outcome of an event? It is frequency and it is an outcome here how many times the specific frequency, specific event has occur.

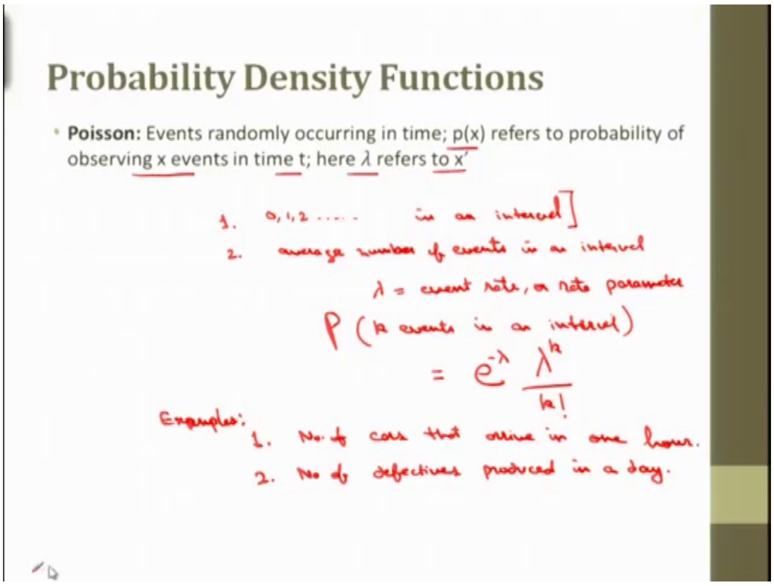
Now, why do we need all the distributions? Why do we need for Poisson, binomial, log normal, rectangular, triangular all the distributions? These are actually, the Poisson and binomial I will discuss from discrete view point for discrete data and this will for continuous data. However, they can also be continuous binomial distribution, and on the other hand we can have a discrete normal distribution as well but more commonly used distribution for the discrete purpose for the discrete data are the Poisson and binomial and normal is generally used for continuous data.

Now, what happens when we have made the observations, we have different kinds of calculations like we have go no go gauge, we learn how to take measurements using various instruments, how to measure lengths, how to measure temperature, pressure, strain force. Most of these values would fall in a continuous state. Sometimes what happens in a industrial measurements the measurements are not continuous and not continuous sometime it can be continuous but it is not just the way we look at the height length. For instance we can say like in go, no go gauge whether it is go or it is no go, ok.

So, I can say whether the product that have produced is acceptable or non acceptable, whether it is defective, or non defective there is no in between, ok. So, it is kind of 0 1 0 1, we have binary kind of data I have. Then what distribution would fit here. Then sometimes there is a the interval is fixed, the time space or the observation space is fixed that out of within a week how many times the defect sector or I can say within an hour how many times the defect sector, how many pieces per hour or how many pieces per 20 pieces. So, in that case the time space is fixed we have distribution for that.

Then sometimes we do not have a plenty of data, we do not we are not very clear about distribution we called it is a lack of knowledge or lack of information, then in that case we can use uniform or triangular distributions. So, we need to select distribution carefully based upon the type of observations that we have taking.

(Refer Slide Time: 04:44)

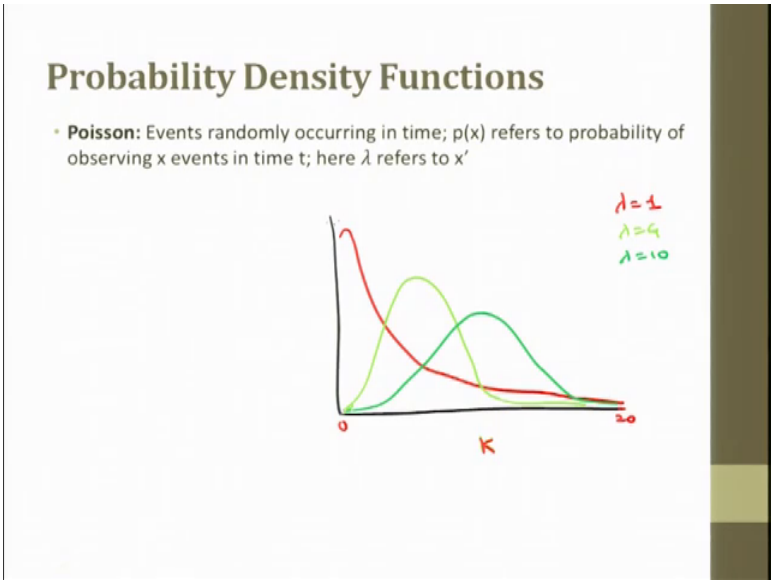


Now, the first issue distributional let to discuss here is Poisson distribution. And now here the events randomly occurring in time p x refers to probability of observing x events in the time t and lambda refers to x dash here. So, what is Poisson distribution? Poisson distribution in event can occur 0 1 2 for so many number times in an interval some event occurs in an interval. Let me say there is a some noise (Refer Time: 05:21) in the machine that we are using about in manufacturing. How many times does the specific peak noise value or sound peak sound value of occurs in an hour or how many times is the setup disturb in between.

So, we can have multiple examples in this. So but the thing is that the an interval the space is fixed here, ok. I can just put a certain key points here. Then average number of events is lambda here that is lambda is average number of events in an interval. So, I can call lambda is event rate, ok. The lambda is event rate also it is called the rate parameter. So, the probability of observing k events in an interval is given by the situation if we have to observe k events in an interval, ok. The probability of observing k in events in an interval is equal to e power minus lambda lambda power k by k factorial. So, this is Poisson distribution.

So, the very first application of Poisson distribution for the soldiers who were accidentally killed by horse kick ok, so this was an example. Now, there can be certain examples in Poisson distribution that number of specific crimes that happens in a month, ok. Then I can put certain examples here, then we can have number of arrivals in a service centre, number of cars that arrive in one hour for service, ok. Another example could be the number of defectives produced in a day by a machine, ok. Then I can have like cars we can (Refer Time: 08:14) say the number of patients those walk in clinic to get treated in an hour, then similar like in that specific times space in an hour or in a day when the specific number is there this kind of distribution Poisson distribution can be used.

(Refer Slide Time: 08:39)



Now, the Poisson distribution can take various forms here, depending upon the value of lambda the Poisson distribution maybe like this. So, Poisson distribution is highly skewed if the value of lambda is small. So, lambda equal to 1, so lambda equal to 1 the distribution is somewhat like this. So, I have discussed about Skewness, I will take up this Skewness in detail when we will discuss in normal distribution.

So, it is highly positively skewed distribution when the value of lambda is 1, ok. If value of lambda is greater than 1 or it is the let me say the level of lambda is somewhere equal to lambda is equal to 4, if lambda is equal to 4 the distribution can take some value like this somewhere like this. So, this is actually I am presenting k, this is k here and I am taking this out of 20 like the map with the value here is 0 times in 20 times, ok. So, this is for lambda equal to 4, so lambda even greater than 4 let me say lambda is equal to 10. It would be somewhere like this, it will be close to the normal distribution, ok.

So, this is the characteristic or the behaviour of the Poisson distribution. So, what are we bothered about here in the probability density function or in the probability distribution? So, what is the mean value of the variable based upon a finite number of measurements and how well this value represents the entire population? So, do the variations found and share that, so we can meet tolerances a population as a whole or how good are this results what essentially is happening this distributions for the whole population. Like somewhere did in a multiple experiments and whatever to develop this distributions like have by conducting maybe 1000s of experiments.

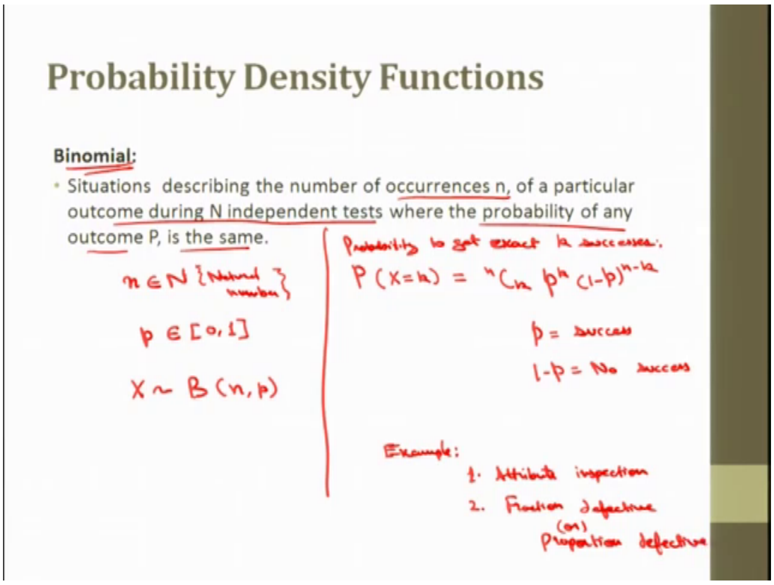
But if we have a limited sample we pick the sample like number of defective in an hour, ok, did you distribution for that. We pick for (Refer Time: 11:12) for 8 hours, 10 hours, [FL] have some sample, ok. Number like we 5 let me say 5 defects per hour, maybe number 1 is 5 defectives per hour, number 2 is 2 defectives per hour, 3 defectives per hour something like that and which we put pick a sample of maybe 10 items like this. But the population is the probability distribution Poisson distribution.

Now, how well our sample fixes with the population, how well the curve fitting happens here curve fitting if you know the term how well the curve fitting happens here based upon that Poisson distribution is best suitable for the events that occurs in a specific time, when interval is fixed. So, this is happening in probability density functions we choose the distribution based upon this criteria. So, what is the target? I will put the target for probability density functions here.

The purpose of probability distributions, purpose is a single representative value would that that tells us the average of the measure data. We need to have number 1 a single representative value, representative value means that is should be close to your target value. Then this representative value that provides a measure of (Refer Time: 12:45) variation in the measured data set then representative value to provide a measure of variation in data set. Then third purpose is we need to know how well the average of the data set represents the true average value of the entire population that is just I said, how well is the average of sample representing the true average of entire population?

So, point number 3 we need to establish in interval within which the true average value of the population is expected to lie and this interval quantifies the probable range of this random error and this is known as random uncertainty, ok. So, we need to work on that for that I have just mention the application and the use and the applicability of the Poisson distribution.

(Refer Slide Time: 14:34)



Now, next is binomial distribution. Now, this describe the situations when the number of occurrence is and of a particular outcome during N independent tests with the probability of any outcome P is same. So, how do we represent the binomial distribution if probability of the random variables x is equal to k as to be determined this is equal to n c k probability of k and 1 minus p, this is actually representative p this is small p the probability 1 minus p power n minus k.

So, p is actually success and 1 minus p is no success you can even call it failure. Say as we say n small n belongs to capital N and (Refer Time: 15:46) it is some natural number. So, small n cannot be sum number with decimal places small n belongs to capital N. So, capital N is actually number, ok. I will put it here it is a natural number and small p belongs to 0, n to 1. So, probability can vary from 0 to 1.

Then the variable X or the random variable X is representing a binomial distribution with n number of occurrence is n probability of success as p. So, the probability of get a exact k success is this is actually probability to get exact K success is if it is small k here, this is equal to this.

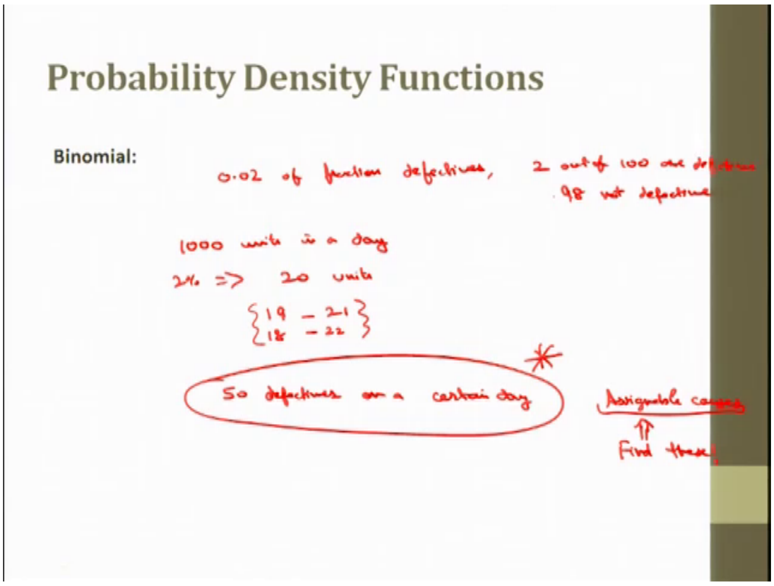
Now, it is (Refer Time: 17:00) means tests that in an industry it is not always possible to measure the quality, in such situations quality could be said to be is a satisfactory or non satisfactory. We use go, no gauge, no go, defective, not defective. We cannot just quantify the data all the time the measurements which we need to take, if I taking measurements that can be continue date that can be specific value. If we just telling that whether it is success or failure these are the most common terms I can use. Success can be accepted, not defective, then go, no go is, ok, both are going in. So, this can be like if this kind of data is there binomial distribution is used, ok.

So, in this case is although it is possible to measure the quality but there is not much time equipment personnel to make the measurements I like to. In other way I can say that the resource is the limited sometimes and we just need to say whether yes or no. So, like go, no go gauges are set. So, these are known as attributes. So, it is actually most applicable in attributes kind of inspection, ok, number 1 is attribute inspection, ok.

Now, inspections inspection by measurements is known as variable inspection which is attribute inspection we will do this in control charts. I am just giving brief idea what is attribute inspection when we talk about the qualitative data that is attribute inspection when we call called about talk about it quantitative data that is variable inspection. So, binomial data it is actually just said before the binomial and Poisson. So, these are more used for the discrete distribution.

There is another term that is a corrected to the binomial distribution that can be the fraction it defective or proportion defective I can say fraction defective, or we can call it proportion defective.

(Refer Slide Time: 19:32)



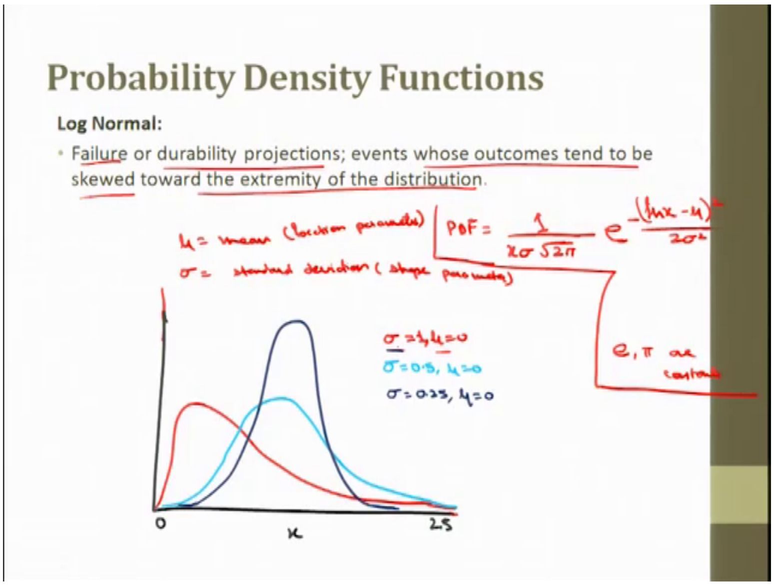
So, fraction defective I like to puts some light on this as well like if you say if the whole aggregate of a production has certain fraction defective. Let me say 0.02 of fraction defectives, ok. So, then that means, 2 out of 100, 2 out of 100 pieces are defective, 2 out of 100 it is a defective. Now, it is assumed that the production runs over a large number of days and that the aggregate production is large, the days production is a part of the aggregate production that is now overall production that happens in a month. So, this is 2 out of 100 is just a small number. So, this is a sample size out of 100 to a defective I can say 2 defective and 98 not defective 98 not defective.

But if I take the whole universe the whole population in this case now what happens. Let me say if in a day 1000 units are produced if you are produced 1000 per day, 2 percent means 2 percent implies 20 this is 20 units, ok. Now, it can vary from like 20 if the 20 units are produced defective in a day it can be 19 units, someday it can be 21 unit, someday it can even vary from 18 to 22 or something 23 something. So, if sometimes it happens that if in on one days if we have about if you say 50 defectives on a certain day, now this is alarming. 50 defective in a certain they that means, on a 20 has to be average 20 is one mean that the set of purpose of the probability distribution is to find the average and see that does our do our average fit in the population or not, ok.

Now, 20 is the average and the variability is your maybe from 18 to 22. If some if an someday I am having 50 defective that means, this is alarming some condition it case some deterioration has occurred of some assignable causes are there assignable causes which are to be traced, ok. I will put it like this find this, find this causes what are the assignable causes otherwise the random variation as we discussed before random variation is due to the non assignable causes. The assignable causes they can be systematic error random error this random error we have mostly dealing with when you are talking about distributions or we are talking again in the statistics in itself we are talking of the random uncertainty.

So, if causes assignable so 20 to 50 defective is have happened that might mean that there is some big cause or there is some reason there is some reason because of this is happened. Maybe some deterioration is happened to the machine, maybe some setup is a little changed or maybe the worker who is working is not the skills to the level. So, we need to find that. So, I just need touch reflection of defective with binomial distribution because these are always related. So, it is actually 2 out of 100 that is it can we can say success and non-success converted into in this way as well.

(Refer Slide Time: 23:32)



So, next is log normal distribution. So, it is failure or durability projections; events whose outcomes tends to be skewed towards the extremity of the distribution. First I just put the PDF of this, PDF of log normal distribution is equal to 1 by x sigma under 2 pi this is actually log normal which should be derived from the normal distribution only, I will take normal distribution in the end because I like to explain that in detail. So, this is minus of natural log of x minus mu whole square by 2 sigma square.

Now, what are the parameters here? Here x is a random variable, mu is mean, sigma is standard deviation, this is for the population, ok. Now, we need to note here that we have constants here e comma pi are constants, e and pi constants here. So, if you are performing some measurements later finite and for instance length, height, weight you are most likely going to end up with the data that resembles in log normal distribution. For the finite number let me say I can have if let me say this is my variable x and it to be varies from 0 to, it will be actually it varies from 0 to infinity.

But I am taking 0 to 2.5 here. So, I can have different types of colors here different types of the level of Skewness depending upon the variation and u, I will take mu as 0, mu 0 I will take value of sigma is equal to 1 first, it will be like this. Again positively skewed and highly skewed. So, what happens actually? To give a better understanding on this think of it calibrating gauge block, if you begin the calibration you know the target length.

To perform the measurements at a repeated point on a gauge block in such a way that majority of your observation maybe majority of your measurements or measurement results will be centred around the actual length of the gauge block and some measurements will be larger than and a fewer would be lesser than that. Because you are limited with the let me say this is a gauge block I am just using this two measure lengths of some measurement term instrument it is manufactured, let me say something is manufactured I am just measuring because this is my length I am limited with this length only. So, most of the observation would be close to this only gauge block is this thing. So, we are limited by the length of the gauge block here.

So, if we see the realistic situation we cannot measure less than the length of the block. So, measurement is also finite or limited. So, we need to make sure that two consider lognormal distribution the data has to be or the performing measurements that are that are to be finite. So, it may prevent us from encountering measurement errors and which calculate uncertainties. So, if sigma is equal to 1 and mu is equal to 0, it can take this shape. If the sigma, sigma that means, the shape parameter I can that put it here as well sigma is standard deviation this is actually the shape parameter. By shape parameter I mean because it is determine the shape. So, we can see that. So, sigma is equal to 1 and mu is equal to 0 this is the shape that is coming.

So, also the standard deviation of the log normal distribution it is determining forms the historical data. So, sometime we may be able to estimate it using current data. So, the shape parameter does not change the location or height of the graph, the location or height is not change. What changes the location? So, mean is the location parameter here, mean is the location parameter so that atom is the location of that where is graph located. Also we can use median to that may determine the how the graph shrinks or stretches that can determine.

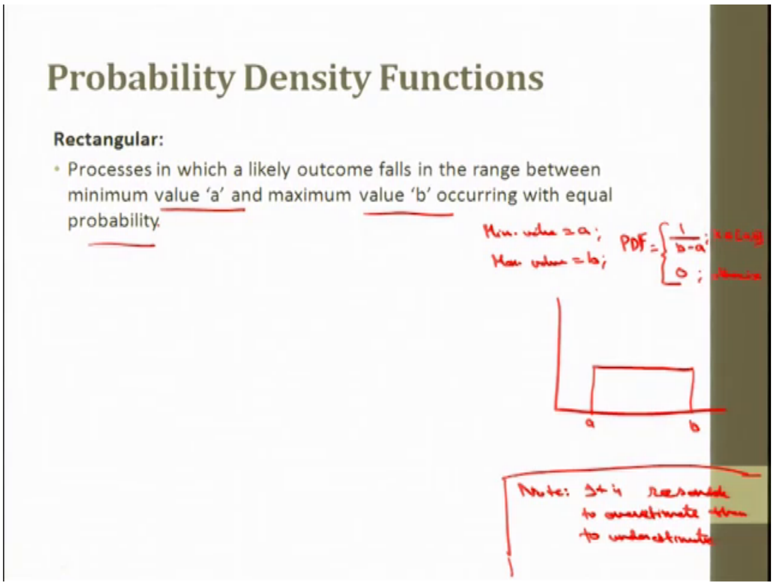
So, for value of sigma is equal to 0.5 and mu is equal to 0 the Skewness as we say that because the what is sigma; sigma. Or what is standard deviation? Standard deviation actually spread the spread more the is the value of standard deviation most spread would be there. So, if this dispersion or spread the values reduce the spread would be (Refer

Time: 29:28) contracted. So, what we will see? We have an a way distribution like this here of from 0 to 2.5.

So, can I can even make one more curve. So, standard deviation is equal to 0.25 and mu is equal to 0 this curve would be further shrinked, it will further more closer but will have some like this because the area has always to be shrinked but it will be lengthier now, it could be sorry taller now. So, this is sometimes known as courtesies, if you courtesies (Refer Time: 30:19) something different than the central line is central tendency is exactly same and we have different heights of the curve then courtesies happen that will also discuss. So, this is the log normal.

You can just to see the approximate area actually the actual area or the exact area below the red, blue and dark blue curve would be exactly same. Only thing is that because of the value of the sigma have because the value of standard deviation the curve is getting contracted here.

(Refer Slide Time: 30:48)



Next comes is rectangular distributions: Processes in which a likely outcome falls in the range between a minimum value 'a' and the maximum value 'b' occurring with equal probability in that case rectangular distribution is used.

So, some time what happens that we have some data or we have some observation but we are not sure that what kind of distribution would it follow we just know the minimum value is a, minimum value is equal to a and maximum value is equal to b in between what is happening we do not know. We do not have this get information about that. We can say that, there is nothing happening this is uniform distribution that is what we call rectangular it can be plotted like this, this is rectangular distribution. The minimum value where is a and the maximum value is b and PDF of this can be given by this is a and PDF is equal to it is 1 by b minus a, ok.

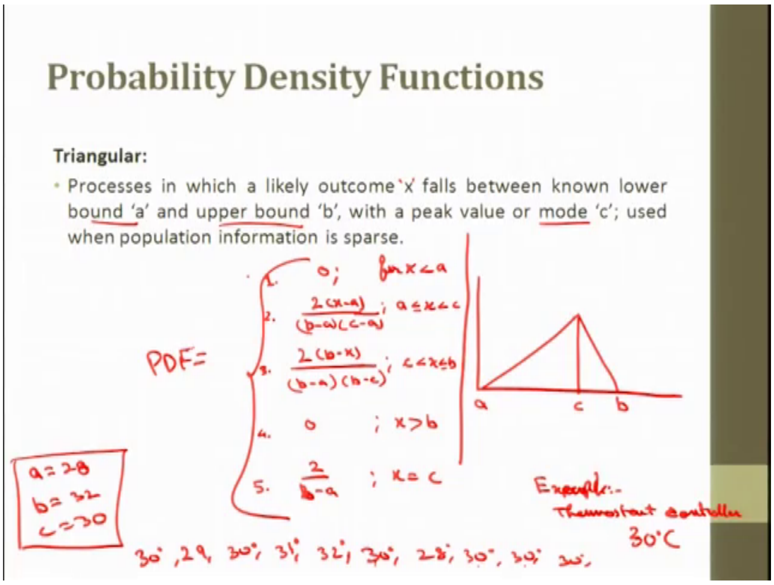
If x is between; if x is between a and b, otherwise this PDF is 0, ok. I will put 0 otherwise. So, we just know the minimum value and maximum value if outlines are the we will just take the PDF to be 0, if outlines means value is greater than the maximum value of lesser than the minimum value. But if it is between a and b rectangular distribution can be used.

Now, rolling is single die can be example of a discrete uniform distribution but we are more concerned with the continuous distributions here. The processes in which a likely outcome falls in the range between minimum maximum, so the most common example of this one is when you do not have any information as a just said and the day of the like if you can see, ok, day of the week of the hottest day of an year is about equally likely to be any of the 7 days. Ever I can say the last digit of the security number of the you can say the cricket captain or of the divided teams that would be always between 0 and 9. So, these are silly examples but the thing is that the minimum and the maximum numbers are known there.

So, in measurement uncertainty when its estimation is to be done it often occurs when it is necessary to make choice between two alternatives of which may be possible lead to somewhat over estimated uncertainty and the other one may maybe lead to underestimate in uncertainty, in such situation it is usually reasonable to rather somewhat overestimate than underestimate uncertainty. I will put a note here. It is reasonable to overestimate then to underestimate, ok. If this is the case we can use the rectangular distribution, ok.

For instance, if I during go and no go gauge is all by just measuring something and the upper tolerances is having high limit or is higher we can over estimated it is upper permissible to have to overestimate under estimating is not allowed, not very much allowed, so then rectangular distribution is (Refer Time: 34:35).

(Refer Slide Time: 34:37)



This is another distribution like rectangular distribution that is triangular distribution. In triangular distribution also we do not have a much of data but the thing is that this one value that is the mode value. So, processes in which are likely outcome x falls between known lower bound a and upper bound b with the peak value mode this is used when population information is sparse we do not have much information but the population. But we have the lower value and higher value but there is middle value let me repeating multiple times here repeating maximum more of times occur is the mode in that case standard distribution is there.

So, it can be something like this. I call it is value a, value d and a mode value c. It takes it shape of a triangle that is why it is known as triangular distribution, ok. I will make some other angle it is triangular distribution. So, PDF for this says PDF triangular distribution is it different conditions for that it is equal to 0 4 x less than a, as we had in a rectangular distribution. It is 2 into x minus a by b minus a into c minus a I am sorry this is b here, ok. The mode is c minimum value a maximum value b upper bound is b lower bound is a, ok.

So, this is the PDF for probability (Refer Time: 36:31), this is for a is less than equal to x is less then c, ok. Then this is equal to 2 into b minus x by b minus a into b minus c you can see b is the upper bound, so b minus x. So, x a was lower bound though we had x minus a in the second part and third part b minus x this is for c is less than x is less than equal to b and again it is equal to 0 for x greater than b, ok. I have put one more here it is equal to actually 2 over b minus a for x exactly equal to c.

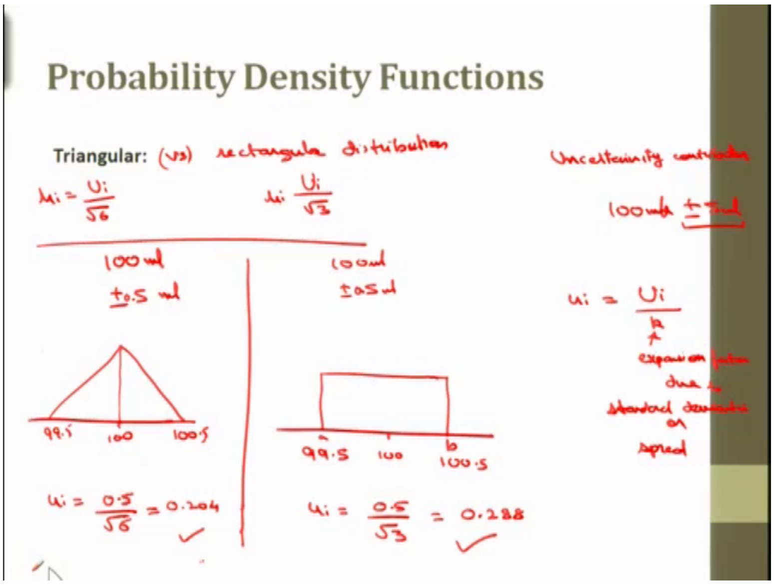
So, this is PDF 1, 2, 3, 4 and 5 cases this is PDF 4 triangular distribution. So, triangular distribution represents in non maximum minimum and estimated central value it is commonly referred to as lack of knowledge distribution when we do not have large information of very good information for the population data. So, it is typically used by relationship between variable is known but the data is spares data is small.

For example, I can say if I like me say (Refer Time: 38:23) metrology lab I am using at thermostat controller that is constantly trying to achieve target temperature set point, ok. For example, I can put here there is thermostat controller thermostat controller that is the set to achieve a target set point, most of the values with the just floating around let me say this is value is let me say this it is 30 degrees, 30 degree centigrade. That is most of the value would be close to 30 degree centigrade.

You know most of time the thermostat will just give 30 degree centigrade like you said you can set thermostat control you can say thermostat of air conditioner. Air conditioner temperature is set to 18 degrees now here, temperature is set into 22 degrees here. The most of the temperature would be 22 degrees, 22 degrees central value it can be 21 degrees, 23 degrees so values can vary for but the most of times a value is 22. What is happening? The value is, I can set thermostat controller if I take 30 degrees. So, 30 degrees it would value would be like; let me check if we take 10 values with 30 degrees, then 29, 30, 31, 32, then 28, then 30, then 30, then 30, 2 3 4 5 6 7 8 9. Then let me say first value was also 30.

You can see the more the more value have 1 2 3 4 5, 30 is repeated 7 6, 30 repeats 6 time and the value of a here is the minimum value 28, the value of b here is the maximum of the upper bound that is 32 and the mode value here is 30, ok. So, in this case triangular distribution can be used so when mode value is here. Sometimes mode value is a not there sometimes we can also use rectangular distribution in case of triangular distribution.

(Refer Slide Time: 40:41)



Now, what determines this (Refer Time: 40:41) between triangular and rectangular you can put triangular versus rectangular distribution.

So, before comparison I like to introduce the term uncertainty contributor, uncertainty contributor. What is this? For instance I have a jar, I get a jar from the market which has to measure 100 ml, 100 ml jar, so this is a kind of a gauge, 100 ml. If I fill the jar completely, then 100 ml liquid should be there, but it might be marked there this 100 ml plus minus 100 ml or it is plus minus 5 ml. So, this plus minus 5 ml it is my uncertainty contributor, uncertainty contributor that is the this specifications are these are non specification limits specific limits are there on my measuring instrument itself, that this can measure from 95 ml to 105 ml, ok.

This is uncertainty contributor and this uncertainty contributor when we talk about the actual uncertainty I will call it u i, this is equal to uncertainty contributor which is actually capital U i this divided by k it k is my expansion factor, ok. Then for normal this value of k maybe the deviation standard deviation the mode plus (Refer Time: 42:29) is the let us say this is the value of u i here small u i. So, this k can be; so in case of triangular distribution u i the value of u i that is generally used is equal to uncertainty contributor by under root of 6 in case of rectangular it is U i by under root 3.

So, based upon based upon this information if I use this data if i, ok. Let me take this jar only, this jar 100 ml 100 ml this is plus minus 5 ml. So, I can call this 100 ml as my mean value, on the centre value and plus minus 5 can be the variation. So, if I try to draw a rectangular distribution, ok. I will I will try to draw the rectangular distribution and a triangular distribution 100 ml and plus minus 5 ml. For triangular distribution I can have something like this. So, as I height from a to b I can even put from minus a to a and centre at 0, ok.

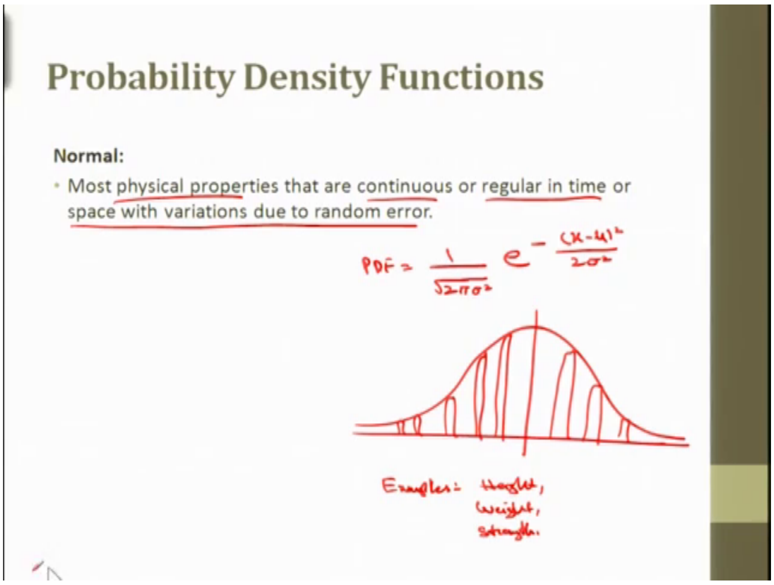
So, in this case in my specific case from a to b this value of a this is equal to 90, 5 ml is not a very good figure if I will put 0.5, ok. I will put 0.5 I will then I will put 99.5 I will put 100.5 as b and the centre value is 100 here but I am following rectangular distribution like this. So, what is my u i here? U i is equal to the uncertainty contributor that is 0.5 divided by under root 3 which is equal to 0.288, ok.

And in this case in triangular distribution I can have something like this 99.5 and 100.5 in most of the times it would be 100. So, would be like this, but in this case a value of u i is equal to 0.5 by under root 6 which is equal to 0.204. So, if I see my uncertainty contributor here it is then reduced by smallest standard deviation or the expansion factor, ok, expansion factor this is expansion factor due to standard deviation or spread, ok.

So, I will pick the triangular distribution based upon the value of u i because, this value is lower if I need to follow the criteria or the tip that I just mentioned before that it is reasonable to overestimate if we need to just pick 100 ml but we cannot afford less than 100 ml we can better choose rectangular distribution if we cannot afford underestimation but overestimation can be allowed, ok. So, instance if we are a like milk vendor is trying to measure using this 100 ml jar or if you are doing some experiments when it we will be need 100 ml of water and we even consider evaporation of might happen some evaporation might happen all the losses are there.

We cannot afford the underestimation here, underestimation we cannot afford 99.5 or maybe if it is 5 ml is the uncertainty contributor, we cannot afford 95 ml. We can afford 105 it is a acceptable 105 acceptable. So, in that case rectangular distribution is there, in that case it will be selected.

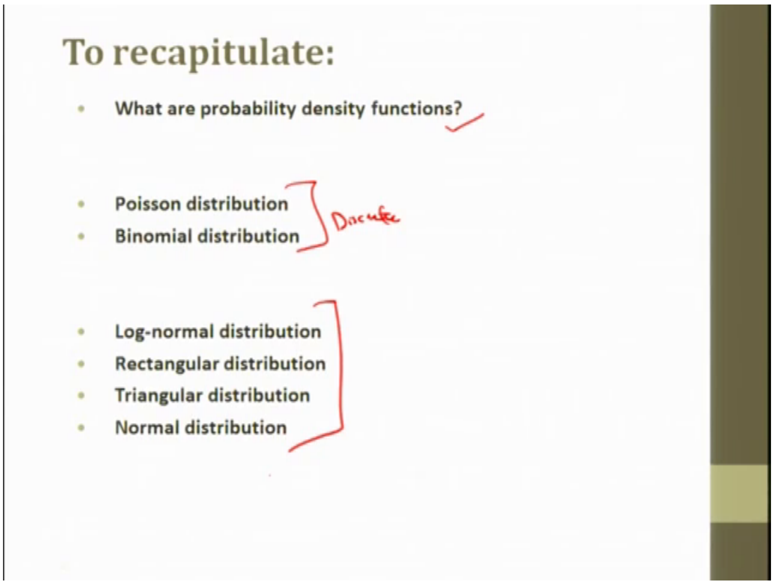
(Refer Slide Time: 47:18)



So, next is normal distribution. Most physical properties that are continuous or regular in time or space with variation due to random error follow the normal distribution the PDF for normal distribution is 1 by under root 2 pi sigma square. So, this sigma would can come out as well here e power minus x minus mu square over 2 sigma square and it is asymptotic distribution, I will explain it asymptotic that its infinity at the extreme ends and it is symmetrical (Refer Time: 48:10) distribution like this, ok. So, this is my normal distribution.

So, to get better understanding we can imagine to collect may be 100 measurements of samples and create a histogram graph with you results like if I (Refer Time: 48:26) 100 histograms like this and so on, ok. So that we will see that the frequency if we are plot the frequency we will see that it will follow normal distribution it can be examples, can be example it can be height, weight, then strength something like that, ok. This is our normal distribution.

(Refer Slide Time: 49:01)



So, to recapitulate within what are the probabilities density functions, we discussed about a few discrete probability distributions, then we discussed about a few continuous probability distributions. We will discuss in log, but the normal distribution in detail and try to solve some problems with them. And we will meet in next lecture where I will take this forward.

Thank you.