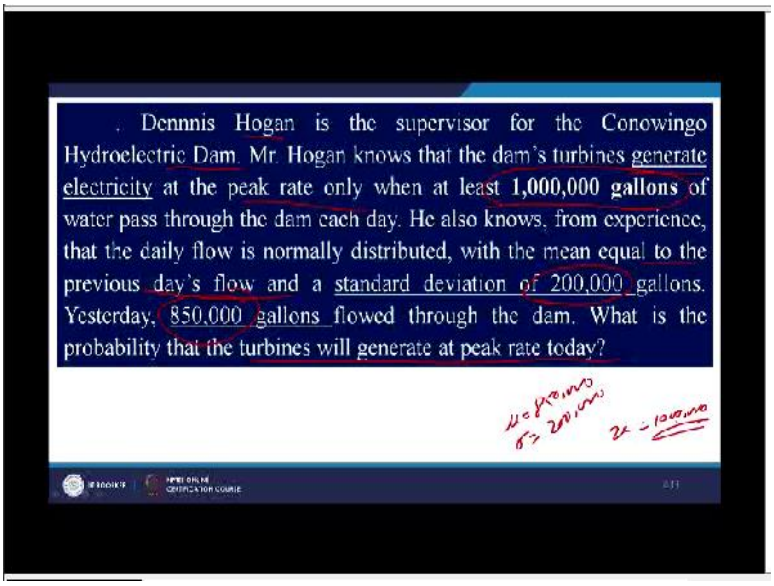
Business Statistics
Prof. M. K. Barua
Department of Management Studies
Indian Institute of Technology-Roorkee

Lecture-20
Evaluating Normality, Exponential Probability Distribution

Good morning friends I welcome you all in this session, as you are aware in previous session we discussed continuous probability distribution and the distribution which we discussed was normal probability distribution, as I said one of the widely used distributions. And we have calculated probabilities in upper tail in lower tail and we also calculated probability between any 2 points under the curve.

(Refer Slide Time: 01:12)



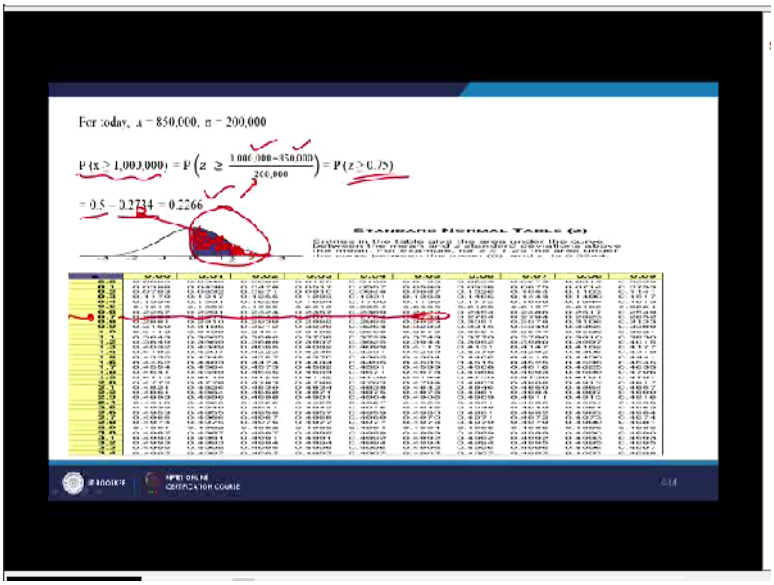
Let us look at one more example quite a simple one, so a supervisor whose name is Dennis Hogan working at hydroelectric dam. So, he knows that the dams turbines generate electricity at the peak rate when at least 1 million gallons of water pass through the dam each day. Otherwise the electricity would not be generated at its peak rate. He also know from his experience that the daily flow of water is normally distributed with the mean equal to the previous days flow.

So, from experience he knows that the mean would be the one it would be equal to whatever water was flown in yesterday right. Now a standard deviation of 200000 gallons and yesterday 850,000 gallons flowed through the dam. Now what is the probability that the turbines will

generate at its peak rate today, so what information you have been given its mew right what is mew it is mean of yesterday's flow right.

So, this 850 is not it what about standard deviation is 200,000 right and what is x value over here just 1 million right. So, x=1000 right or 1 million right.

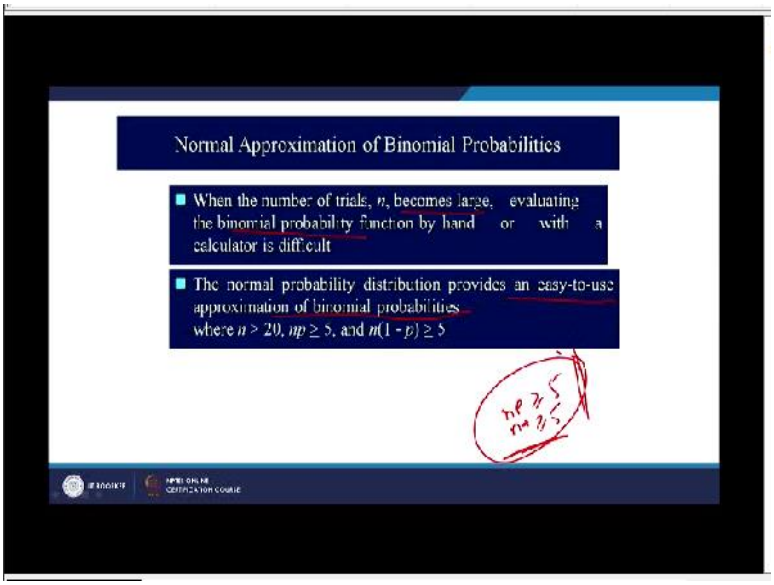
(Refer Slide Time: 02:55)



Now let us look at first of all you should convert this X value into Z value right, so probability that the electricity would be generated can be calculated like this. So, this x-μ/σ, so Z value is equal to 0.75 now look at this table and what is the area when Z value is 0.75. So, 0.7 this is 0.7 and 5 is, so this is 4**,** so this 0.27 let me do it. So, this is 0.2734 this value. Now, so we know that this area is 0.2734 and we also know that this entire area is 50% or 0.5.

So, what we are interested in we are interested in finding this area right, this area and how we can find it since we know this area we know the entire area, so we subtract 0.2734 from 0.5 so, you will get answer is 0.2266 right.

(Refer Slide Time: 04:29)

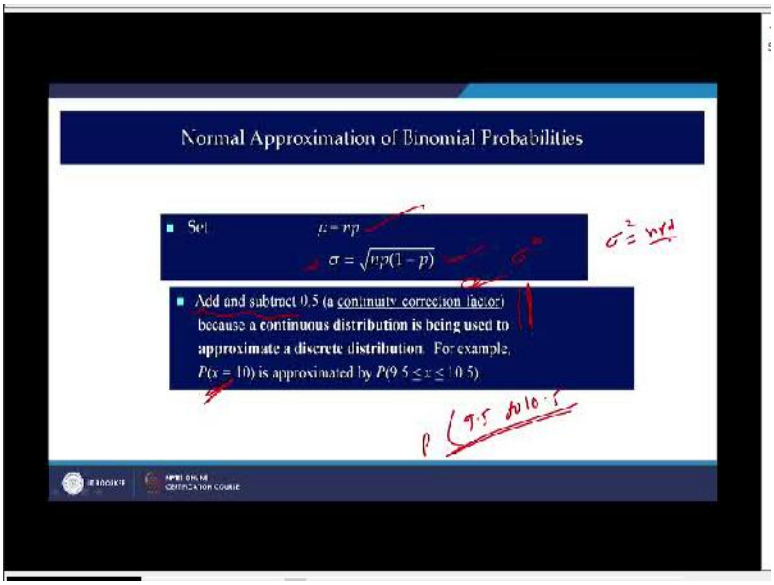


Now let us move on to the approximation of binomial distribution we have seen earlier that we can use poisson distribution instead of binomial distribution and both of them are discrete distributions. But poisson distribution can also be used instead of normal distribution, so let us look at in fact normal distribution can also be used for binomial distribution is not the poisson distribution for binomial estimate.

It is the normal distribution for binomial distribution, so when the number of trials n becomes large evaluating binomial probabilities are quite tedious. So, there is a solution to this particular problem, so normal probability distribution provides an easy use to approximate binomial probabilities. So if np and nq both are greater than or equal to 5 we can use normal distribution instead of binomial distribution.

So, just remember these two conditions and what were the conditions when we used poisson distribution for binomial distribution the n was equal to or greater than 20 and p was less than or or equal to 0.05 right. So these are two different conditions for approximation of binomial to normal right.

(Refer Slide Time: 06:13)



Now this is the mean we know that it is mean of binomial distribution np and npq right. So, this is this is standard deviation is under root of or variance is npq right. So, variance you can variance you can write like this and npq right is not it. Now since this is a case wherein normal distribution is continuous and binomial is discrete. So, there is a need to add and subtract 0.5, which is also known as continuous correction factor.

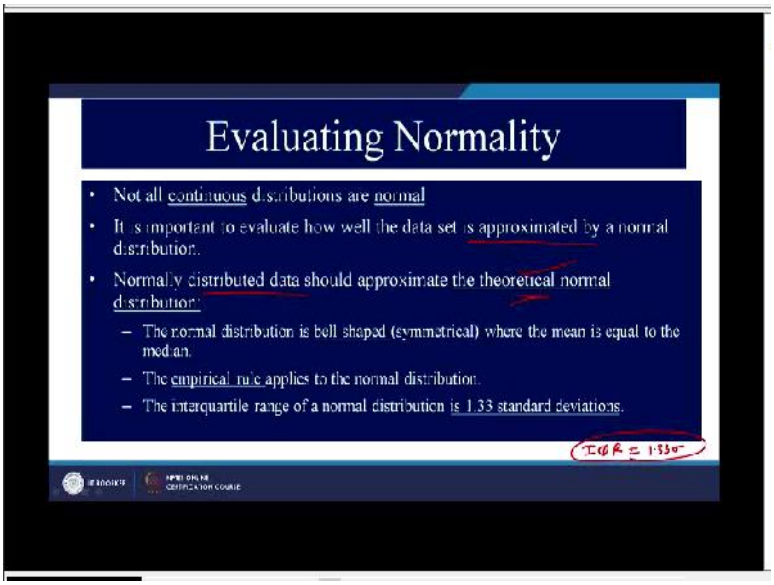
Because we are using continuous distribution to approximate discrete distribution. So, for example if p(x=10). So, rather than going for probabilities of p(x=10) we will subtract and add 0.5 from 10, so it would be 9.5 to 10.5 right. So we calculate probability in this right, so x has to be less than or equal to 10.5 but more than or equal to 9.5 right. So, this is something which you need to remember continuous correction factor ok.

Now whenever you get some data either from primary source or from secondary source you should check it for normality. Because this is one of the assumptions to carry out several statistical analysis, for example if you are going for let us say regression then some assumptions are to fulfilled and one of them is that the data should be normal, since we are discussing normal probability distribution.

So, you should know how to find out whether data normal or not, so we should evaluate normality of the data set right. So, first of all you should remember that there are several

continuous distributions, but all of them are not normal. So, to have normal distribution certain conditions are to be fulfilled by the data set, so it is important to evaluate how will the dataset is approximated by normal distribution.

(Refer Slide Time: 08:34)

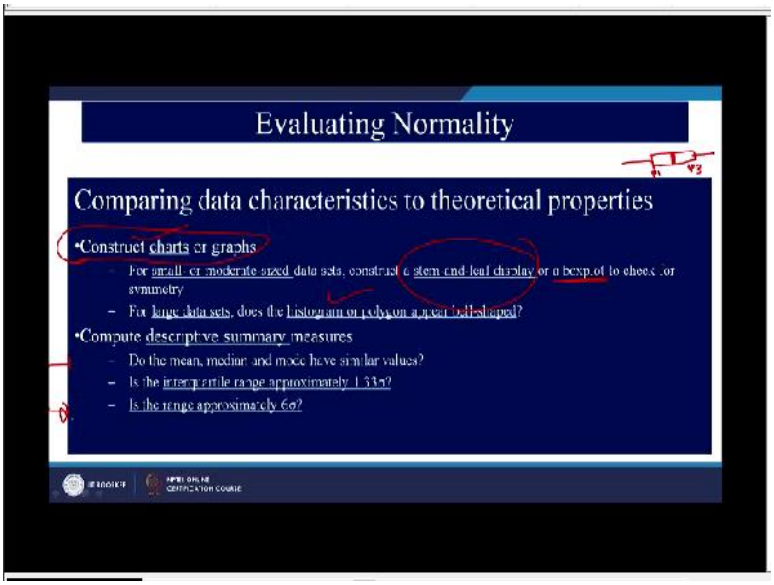


Normally distributed data should approximately theoretical normal distribution. So, whatever is normally distributed data that should fit some theoretical distribution some theoretical points, we all know that normal distribution is a bell shaped curve, it is a symmetric curve, where mean mode and median are equal right. And we also know that the empirical rule applies to normal probability distribution.

So, empirical rule is what is within 1 sigma limit you have got approximately 60% of data points, within 2 sigma limits is 95, and within 3 sigma limits. It is approximately 99.72 right and one of the other characteristics of normal distribution would be iq or should be 1.33 of standard deviation. So, whatever is a IQR interquartile range this should be equal to 1.33 of standard deviation of data. Then you may say that the data are normally distributed.

So, these are some of the conditions and there are many more conditions. So, let us look at data characteristics to theoretical properties. So, first of all to ensure or to check whether data are normally distributed or not you construct a graph or a chart.

(Refer Slide Time: 10:11)

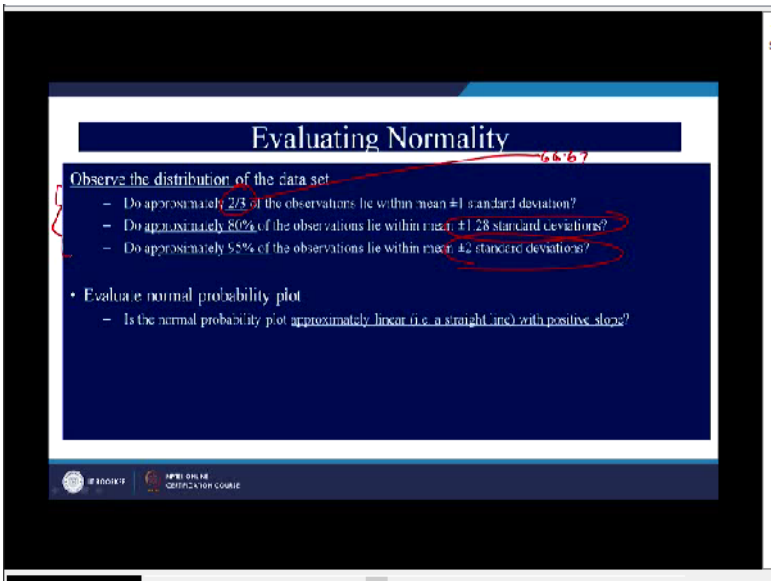


So, if data size is small or small rate then it is always good to construct stem and leaf diagram or box plot to check the symmetry right. And we know that what is stem and leaf diagram and we also no boxplot right, in boxplot you it looks like this is not it is what we have discussed earlier. So, if this point is and this is known as median right this q1, q3 right this q1, q3 right q1, q3 and this is median value right.

So, if this median value is in between here then will say the distribution is normal right or it is not q. So, if data set is quite large than rather than going for stem and leaf diagram and boxplot it is always suggested to go for histogram or polygon. So that will give you an idea whether histogram is taking bell shaped curve or not. So, once so this is these are vary you know visual ways of identifying whether data are normally distributed or not.

But a better solution would be to compute descriptive statistics of the data set, so if you get mean, mode median equal then of course it would be a normally distributed data if IQR is equal to 1.33 standard deviation it would be normal data. If range is equal to 6 standard deviation, range is difference between maximum to minimum, so this will also give you an idea about normality of data.

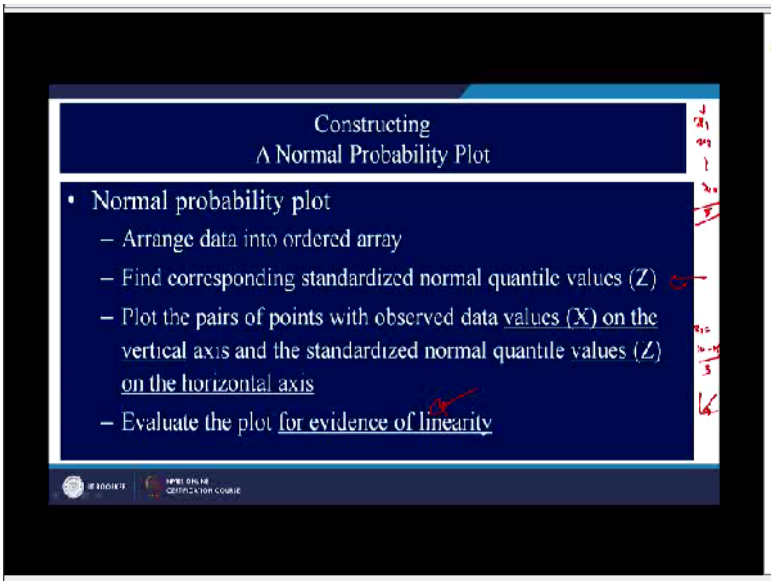
(Refer Slide Time: 12:06)



Now after looking at descriptive statistics and graphs and charts you should observe the distribution of data set. So, if you think that approximately 2/3 of observations are it is approximately 66.66 or 66.67% of observations lie within 1 standard deviation. We will say it is a normal data, otherwise 80% of observations if lie within it is a 1.28 standard deviation, then it would be again normal data, approximately 95% of the data lie within 2 standard deviation.

So if you fulfill all these 3 conditions you can say that data normally distributed, now apart from these methods there is something normal probability plot. So, if you draw normal probability plot will see how to draw normal probability plot and if it is linear then will say it is a normally distributed set of data.

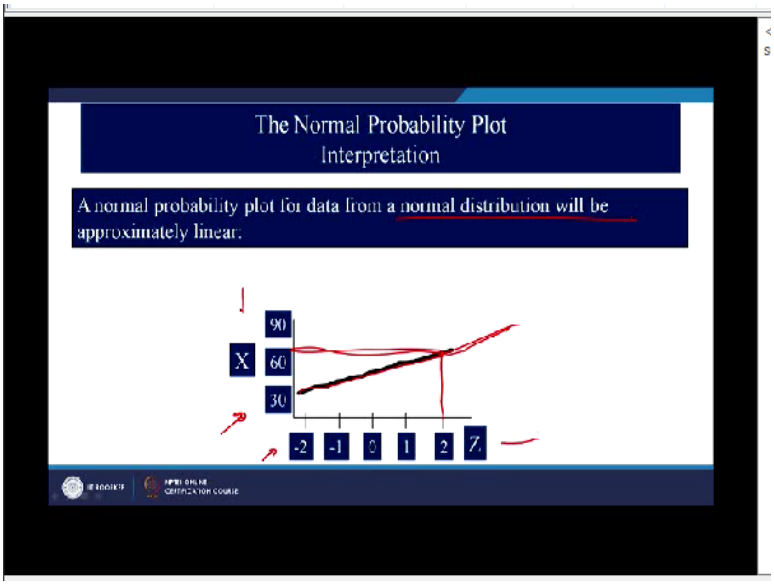
(Refer Slide Time: 13:36)



So for normal probability plot arrange data into array right, find corresponding Z values, so you have got let say X1, X2 and so on right let say X10 right. So, corresponding to these X1 and X10 you can easily calculate Z1 and Z10, so how would you do that because you can easily calculate mean right and standard deviation isn't it.

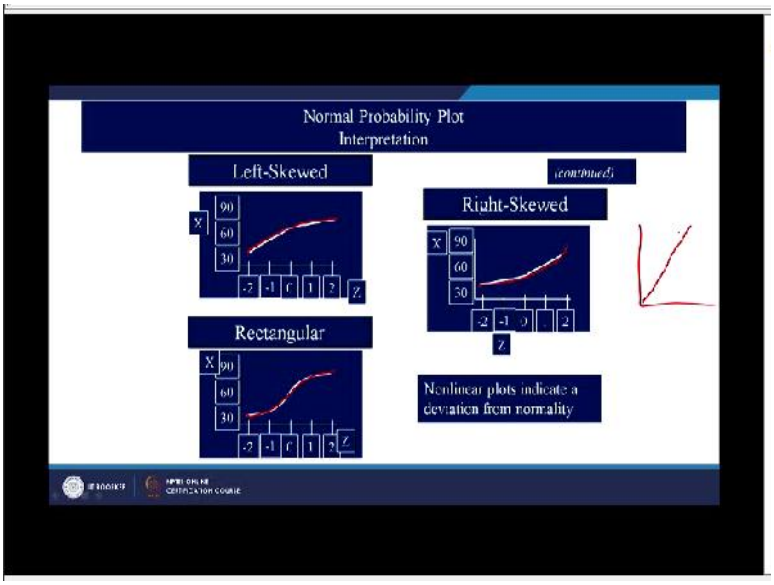
So, how would you calculate Z 1 for X1, so it would be let say Z1 = x-μ right. So, X1 is let say 10 right 10 let say mean is 15 and standard deviation is 3. So, this how you can calculate Z1 and Z10 ok, so plot the pair of points observed. So, in other words let me plot it in this way, so plot X values on vertical axis and Z values on horizontal axis. So, if you see straight line then it would be normally distributed data set. So, if it will evaluate the plot for evidence of linearity if it is straight line then it would be straight line means a line like this right having 45 degree angle from X-axis right.

(Refer Slide Time: 15:12)



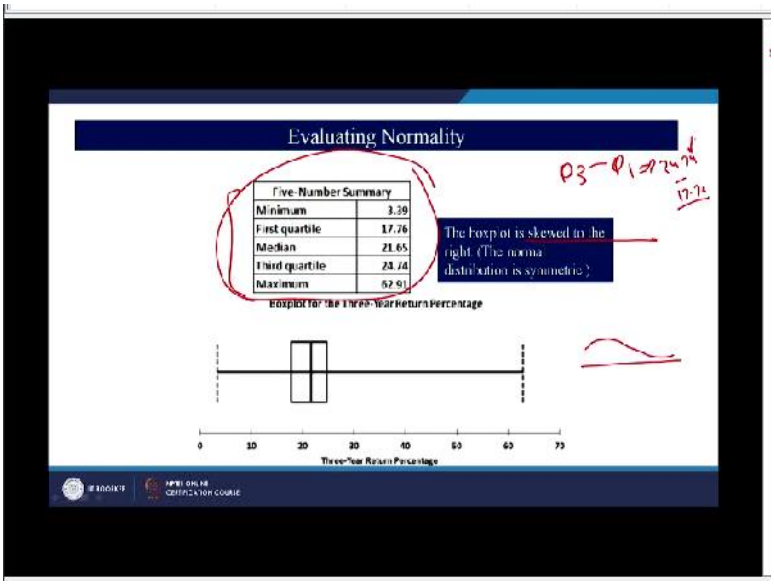
So, let us look at, so this is an example a normal probability plot for data from normal distribution will approximately big linear. So, just see this is quite a linear one isn't it, so for X=30, Z is -2 and for some other value of X let say this 60 or 70 right, Z is +2. So, you can draw X and Y, X and Z values on vertical axis and horizontal axis.

(Refer Slide Time: 15:56)



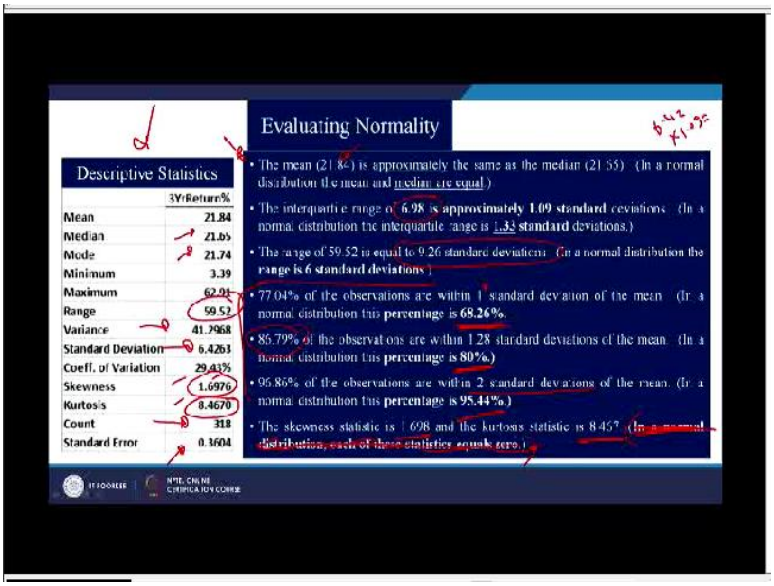
So, these plots will give you an idea how data are skewed whether left skewed or right skewed. So, a plot like this is nothing but a left skew data right and a plot like this right skew data right. And this is a kind of a rectangular plot, but how normally distributed plot would look like if data normally distributed it would be look like, it will look like this is not it a linear line ok.

(Refer Slide Time: 16:40)



So, if you have got a data set then you can evaluate again for normality, so first of all you just come up with descriptive statistics. So, this is nothing but 5 number summary, so you can calculate mean, mode, median and standard deviation then variance then IQR and so on right. So, this is a boxplot now this one is skewed to the right, so this skewed to the right means this is this kind of distribution ok. So, we will see whether the data set from which we have found out this descriptive statistics was normal or not right.

(Refer Slide Time: 17:38)



So, let us look at this, so you calculated all these descriptive statistics, so mean was 21.84 median mode minimum value maximum value range is this right. Variance is this under root of this variances standard deviation skewness, kurtosis there were 318 data points and standard

error is 0.36 right. So, will see whether the data set from which we obtained this descriptive statistics was normal or not right.

So, if you look at this the mean is 21.84, median 21.65 and 21.7, so all are almost equal right. And we know that in a normal distribution all these are equal right, so this first one right IQR is approximately 1.09 what is IQR here IQR is not given. So, you need to calculate IQR is what it is Q3-Q1 is not it, so Q3 is 24.74 - Q1 is first quarter is 17.76 right. So, if you subtract 17.76 from 24.76 will get IQR right.

So, IQR would be 6.9 or approximately 7 right, so if IQR is 1.33 of standard deviation will say that it is a normal distribution. So, it is not exactly 1.33 standard deviation but is 1.09 standard deviation, so deviation is what it is 1.42 right. So, 1 this 6.42 right 6.42 into 1.09, so you will get IQR is 6.92 ok. Range is equal to we also know that the range has to be equal to 6 standard deviation is it.

So, it is not exactly the range is equal to 9.26 standard deviation not 6 standard deviation in this particular data set. Let us look at the other criteria, so if we have 68 26% of data points .within 1 sigma limit will say let us normal But here in this case it is not 68 but it is 77.04%, again 80% of data should be between 1.28 standard deviation that is what we have seen in one of the slides earlier is not it.

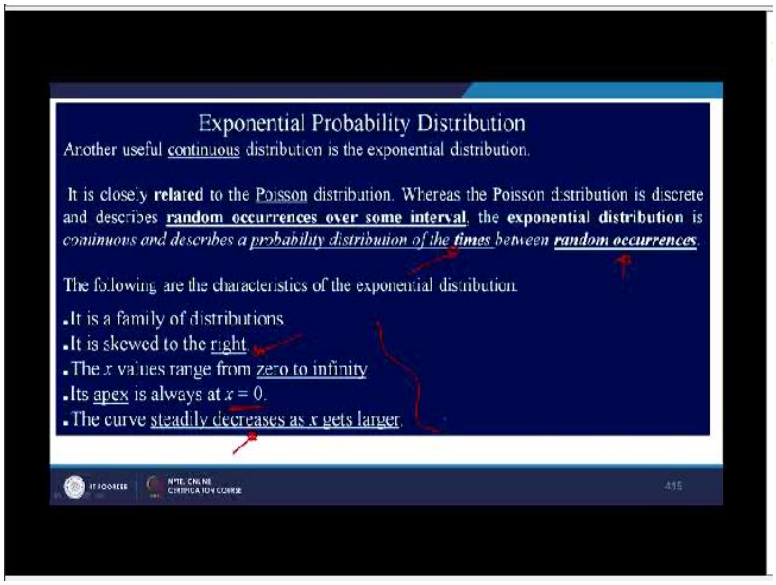
Just what have seen 80% of data within 1.28 standard deviation isn't it. It is there in our example no it's not 80% but it is it is 86 approximately 87% right. So for is within 2 sigma limits or within 2 standard deviation if you them 95.4% of data but here it is 96.86% of data points and skewness is 1.69, kurtosis is this. So, in normal distribution each of these statistics equal to 0. In fact we will look at this particular part and little later.

So for the time being you just do not look at this right, so skewness in this particular case is 1.69, this is skewness and kurtosis is 8.46. Because we will look at what should be the values of skewness and kurtosis for normality, will look at in some other session right for the time being

do not consider this ok. Now will move on to the third continuous distribution which is exponential probability distribution.

So, we have seen already two continuous probability distribution, first one was uniform and the second one was normal. So, this another continuous distribution is closely related to poisson distribution, what we have seen poisson distribution, In Poisson distribution be calculated probability of happening of an event within given interval of time. But here it is just convert to that, here what we try to do we try to find out probability distribution of the times between random occurrences.

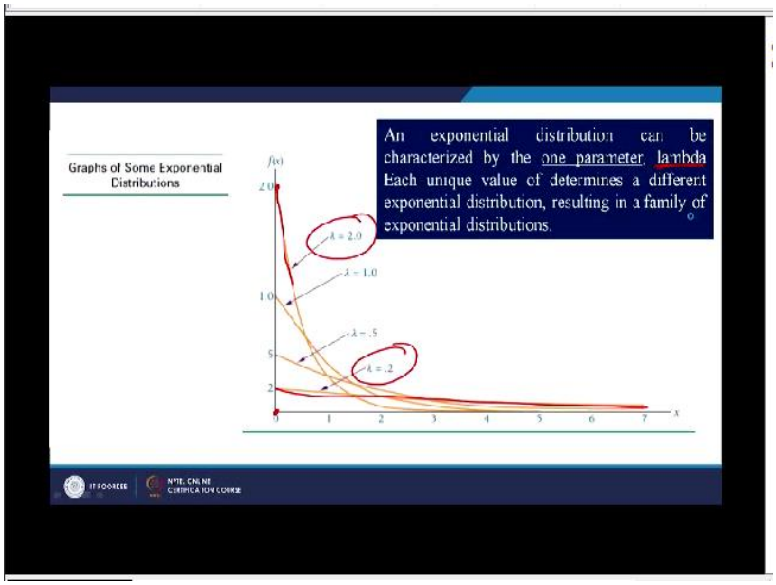
(Refer Slide Time: 23:10)



So, it has got close relationship with Poisson distribution and we can describe the exponential distribution with several characteristics. So, first of all exponential distribution is a family of curves, so the shape of the curve would depend on lambda or is the mean right mean arrival rate. It is skewed to right, it is always skewed to right it is never skewed to left the X value range from 0 to infinity.

And the apex value or the peak value is always at X = 0 of the peak value of the curve, the curve is steadily decreases as X gets larger so will see all these points in next slide.

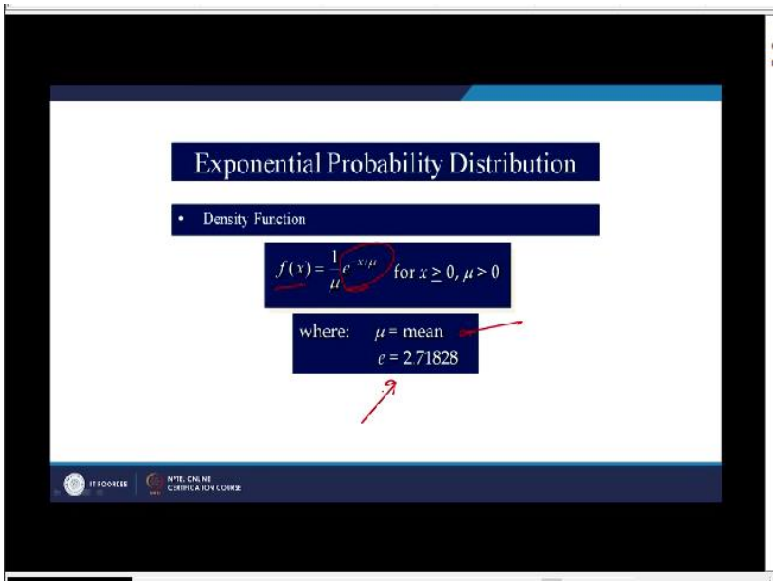
(Refer Slide Time: 24:15)



So, let us look at this graph of exponential distribution, so as I said it is a family of curves, so these are different curves. The peak value of this curve is here and this peak value is when X=0 this what we have said in previous class right in previous slide. So, apex its apex is always at X=0 right, so when you increase X from 0 to some positive value then you can see that this λ values decreases right.

So, this is exponential distribution for λ = 2 and this one is for λ is equal to 0.2 right. So, an exponential distribution can be characterized by one parameter that is λ right, each unique value determines different exponential distribution this what we have seen in resulting in a family of exponential distribution.

(Refer Slide Time: 25:26)



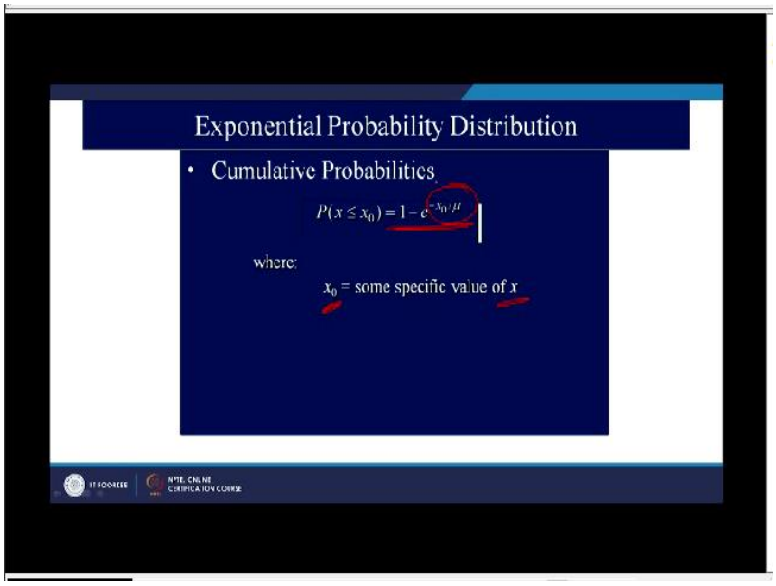
Exponential probability distribution is useful in describing time it takes to complete a task, but we have seen in Poisson distribution what is the probability of happening something in a time interval. But here what we would calculate what is the how much time a task text to complete it. So, the exponential random variable can be used to describe exponential distribution.

So, let us look at some of the examples or applications of exponential probability distribution. The exponential random variable can be used to describe these processes let say time between vehicles arriving at a toll booth isn't it, time required to complete the questionnaire or distance between major defects in highway right. So, these are couple of applications wherein you can use exponential probability distribution.

The exponential probability density function $f(x) = \frac{1}{\mu} e^{-x/\mu}$ for x ≥0, μ>0

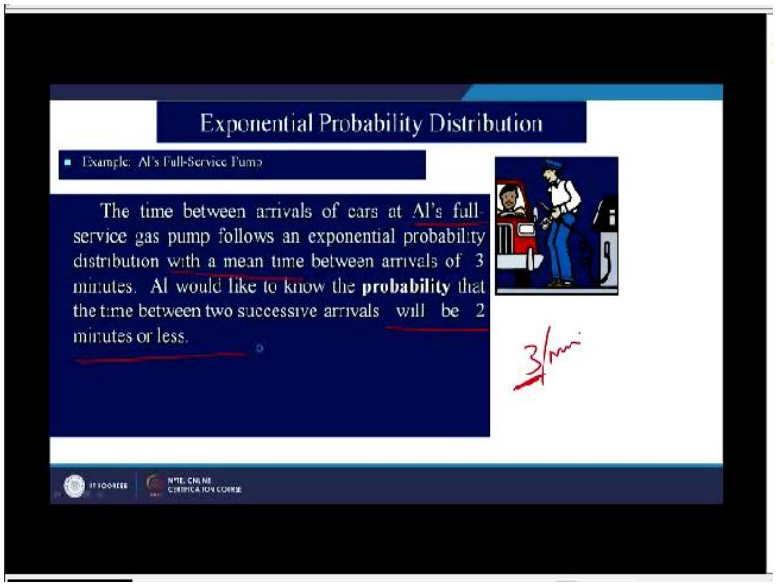
So, here μ is μ right and e = 2.71.

(Refer Slide Time: 27:12)



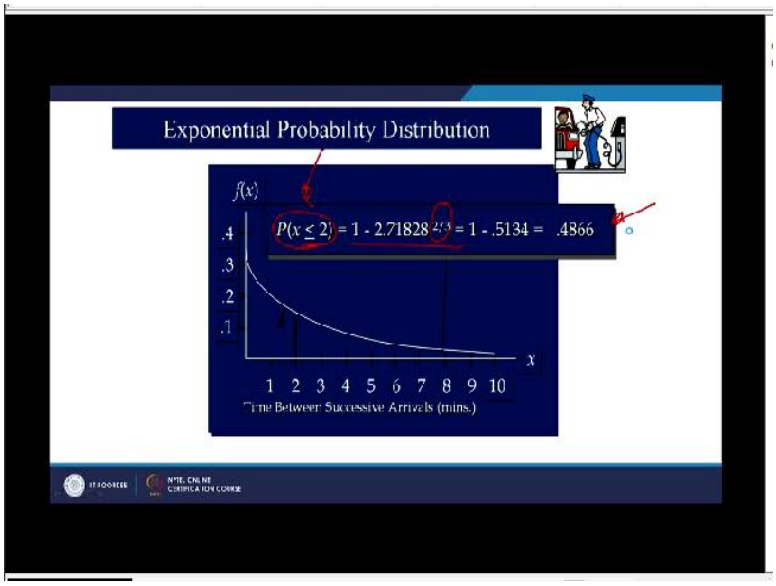
We will work out an example calculate, so this how you can calculate cumulative probabilities for some specific x0 is some specific value of x right, so this 1, so 1-e to the power this, ok.

(Refer Slide Time: 27:31)



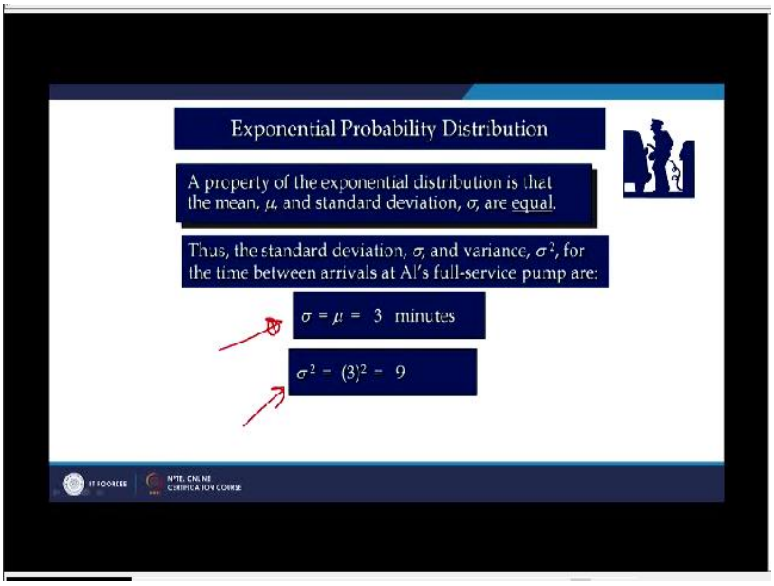
Let us look at an example Al's fuel service pump, so the time between arrivals of cars at petrol pump or gas pump follows an exponential probability distribution with mean of 3 arrivals per minute, so number of vehicles coming at a gas station or a fuel station is 3 per minute. Now, we would like to know the probability that the time between 2 successive arrivals will be 2 minutes or less than 2 minutes ok.

(Refer Slide Time: 28:16)



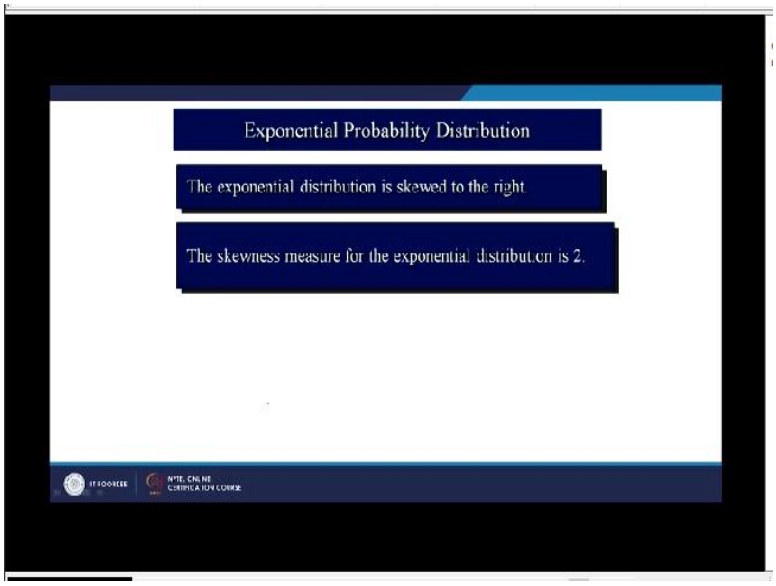
So, let us look at this, so the probability that the arrival would be 2 minutes or less than 2 minutes right equal to or less than 2 minutes would be 1-e to the power x/μ right, here μ is 3. So, this is the answer 0.4866.

(Refer Slide Time: 28:40)



So one of the properties of exponential probability distribution is that the mean and standard deviations are equal. So, in this particular case the mean would be mean and standard deviation would be 3 and of course variance would be 9.

(Refer Slide Time: 28:59)



We know that the exponential distribution is skew to right we have already seen this and skewness measure of exponential distribution is 2. So, with this let me summarize what we did in today's session, we have seen couple of examples on normal probability distribution and we have also seen how to evaluate normality of data set. So, there are certain parameters on which you should check your data set for normality, the example which we have seen in which most of the criteria were not fulfilling.

So, we would say that the hour data set was not normally distributed and we have also seen exponential probability distribution wherein we want to find out the distribution of probabilities of times rather the probability of occurrence of an event within time intervals. So, thank you very much for this particular session, will have a next chapter in next class, thank you.