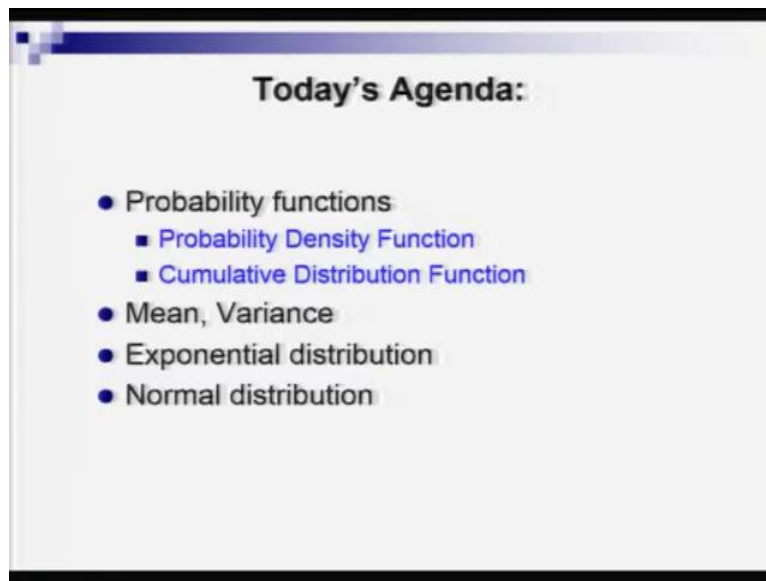


Applied Statistics and Econometrics
Professor Deep Mukherjee
Department of Economic Sciences
Indian Institute of Technology Kanpur
Lecture 06

Continuous Random Variables & Probability Distributions

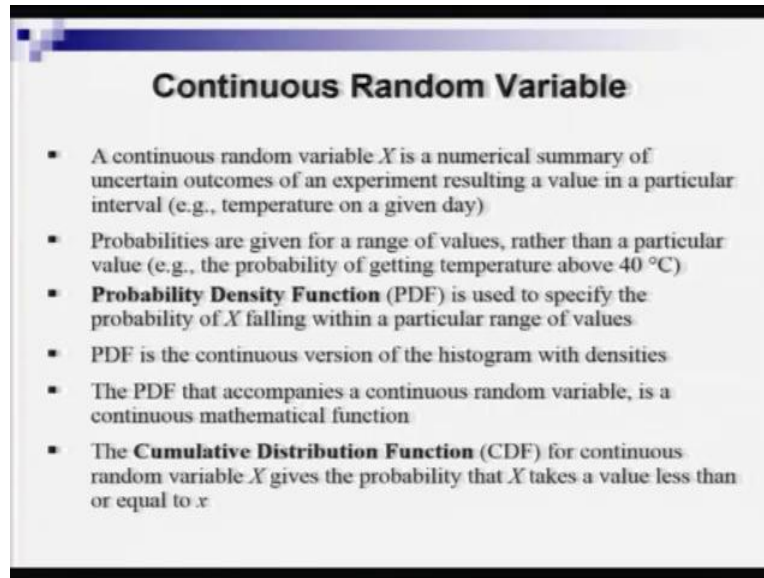
Hello, welcome back to the lecture series on Applied Statistics and Econometrics. Today, we are going to discuss the case of continuous random variable and associated concepts. So, let us first look at that today's agenda items.

(Refer Slide Time: 00:30)



So, as we are going to introduce the continuous random variable, the associated concepts of probability density function and cumulative distribution function will also be presented to you, and then we are going to look at how to compute mean and variance of a continuous random variable. Then, we are going to end today's lecture with two examples of very popular continuous random variables, namely exponential distribution and normal distribution.

(Refer Slide Time: 01:02)



Continuous Random Variable

- A continuous random variable X is a numerical summary of uncertain outcomes of an experiment resulting a value in a particular interval (e.g., temperature on a given day)
- Probabilities are given for a range of values, rather than a particular value (e.g., the probability of getting temperature above 40°C)
- **Probability Density Function (PDF)** is used to specify the probability of X falling within a particular range of values
- PDF is the continuous version of the histogram with densities
- The PDF that accompanies a continuous random variable, is a continuous mathematical function
- The **Cumulative Distribution Function (CDF)** for continuous random variable X gives the probability that X takes a value less than or equal to x

So, to define a continuous random variable, one can go ahead like the following, continuous random variable X is a numerical summary of uncertain outcomes of an experiment resulting a value in a particular interval. So, it is important to note that, on the contrary to discrete random variables, continuous random variable does not take a particular value, we say it takes a value in a particular interval, so note the difference, it is very important. So, the basic feature of a continuous random variable is that it is one realization in a (continuum) from a continuum on a real line, an example could be temperature on a given day.

So, here as we are not talking about individual values of continuous random variable probability is also to be found not for one particular value, it will be calculated for a range of values. So, if we go back to the example of temperature on a given date, we can say that what is the probability of getting temperature above 40 degrees Celsius or getting temperature between 30 degree and 40 degrees Celsius. So, in the case of discrete random variable, we have seen that the probabilities are to be calculated from the probability mass function.

Here, we introduce the concept of probability density function for the case of a continuous random variable. So, it gives us the probability of X falling within a particular interval or range of values. Note that a PDF is the continuous version of the histogram with densities. So, what do I mean by that? So, if you remember the histograms and if you can relate the histogram to discrete random variable, for one particular value of discrete random variable, you see one spike or one bar.

Now, if there are only 3 or 4 or maybe 10 such values taken by the discrete random variable, you observe 10 spikes and each spike actually gives you the relative frequency or the

probability of one particular outcome of the event, but if you have an interval from where your realizations of random variable is coming, then you can generate in finite number of such bars or many large number of bars or spikes in that diagram, hence, the gap the width of a particular bar is going to be so small that as N tends to infinity, it implies that the number of values becomes very large, then, you can approximate those collection of bars by a continuous line and you get PDF by doing so.

So, the PDF that accompanies a continuous random variable, is a continuous mathematical function. In this regard, let me also introduce the concept of the cumulative distribution function. So, for a variable X , it gives the probability that X takes value less than or equal to X a particular value. Now, it is not very different from the case of discrete random variable. Let me move on.

(Refer Slide Time: 04:32)

PDF & CDF

- ♣ Properties of PDF: (i) $f_X(x) \geq 0$ and (ii) $\int_{-\infty}^{+\infty} f_X(x) dx = 1$
- ♣ CDF is given by: $F_X(x) = \int_{-\infty}^x f_X(t) dt$
- ♣ Properties of CDF:
 - As $x \rightarrow -\infty$, the value of CDF approaches 0: $\lim_{x \rightarrow -\infty} F_X(x) = 0$
 - As $x \rightarrow +\infty$, the value of CDF approaches 1: $\lim_{x \rightarrow +\infty} F_X(x) = 1$
 - CDF is a non-decreasing function: $F_X(x_2) \geq F_X(x_1)$ if $x_2 > x_1$
 - If CDF is differentiable, its first derivative gives the PDF: $f_X(x) = \frac{dF_X(x)}{dx}$
- ♣ The probability that continuous random variable X lies in the interval $(x_1, x_2]$ is

$$P(x_1 < X \leq x_2) = F_X(x_2) - F_X(x_1)$$

Now, we are going to look at features of PDF and CDF, some of the nice properties that they have. So, first, we will look at the properties of PDF and note that PDF the as denoted by f of x is always positive number the lowest possible number is 0, it cannot be negative that is same that we have seen in the case of discrete random variable. Now the second number important property of the PDF is that, when an integration is taken from minus infinity to plus infinity, these PDF results into 1.

So, in this course, I am not going to explain how you can compute the value of this indefinite integral, I assume that you have done a little bit of calculus in class 11th and 12th level so that you can handle indefinite integrals. But in a nutshell, what does this expression say, if you want to express these mathematical expressions in terms of layman's language, what to

say? So, when you have the probability density function drawn on a piece of paper in the x and f_x quadrant, then the area below the curve, should be equal to 1. That is what it means. Now, let me move on to the CDF.

Now, we have expressed CDF in language in the last slide, now, it is CDF using mathematical symbol. So, again the mathematical symbol or mathematical representation of CDF is given by another integral, and this time it ranges from minus infinity to the particular value at which the CDF is to be calculated. Now, we are going to look at the properties of the CDF. First, as x extends to minus infinity the value of CDF approaches 0.

So, this is a limit statement, then as x tends to plus infinity the value of CDF approaches 1 this is another limit expression, then the third property says that CDF is a non-decreasing function that means that if we have two values x_2 greater than x_1 , then capital F of x_2 is greater than or equal to capital F of x_1 .

Now, this property is very handy when we are going to compute the probability of the continuous random variable whose values lie in a particular interval, but we will come back to this computation feature later, let me finish the properties first with the last one and that is the differentiability property. So, if CDF is differentiable, then its first derivative gives the PDF. So, you take the first derivative with respect to the random variable of capital F, you get the small f .

So, now, coming back to the computation of probability when the continuous random variable x lies in an interval small x and small x_2 . So, here you please pay attention to the sign or notation that I am using. So, I am using the first bracket or parentheses against the x_1 value. So, that means that x_1 is not included in the interval. So, x_1 value is not to be taken by the continuous random variable whereas, x_2 ends with a third bracket and that means that x_2 value is inclusive in the interval.

So, the random variable x can take the value x_2 . So, in that case you have this probability of x taking value between x_1 and x_2 is basically the difference between two CDF values. So, the first CDF number is going to be calculated at the highest value which is x_2 , so, F of x_2 and then F of x_1 needs to be calculated as well and then you need to subtract F of x_1 from F of x_2 .

(Refer Slide Time: 08:53)

Mean & Variance

- Mean: $E(X) = \int_{-\infty}^{\infty} xf(x)dx$
- Change of variable formula for expectation: $E[g(X)] = \int_{-\infty}^{\infty} g(x)f(x)dx$
- Variance: $\text{Var}(X) = E[(X - E[X])^2]$
 $\text{Var}(X) = E[X^2] - E[X]^2$
- Change of variable formula:
 $E[rX + s] = rE[X] + s$
 $\text{Var}(rX + s) = r^2\text{Var}(X)$
- Two distributions with identical mean values but different variances

The slide includes a graph of a probability density function $f(x)$ versus x . The graph shows two normal distribution curves centered at the same point on the x-axis. The taller, narrower curve represents a smaller variance, while the shorter, wider curve represents a larger variance. A blue arrow points from the text 'Two distributions with identical mean values but different variances' to the graph.

Now, we are going to look at the mean and variance of the random variables. So, mean as I have defined earlier also it is an average value of the distribution. So, this is a central point of the frequency distribution or probability distribution and in the case of a continuous variable, you need to take an infinite, indefinite integral, where the range is from minus infinity to plus infinity and you need to take the independent integral of X times fx.

If you want to look at this formula from discrete point of view, suppose you have many, many values of X and the corresponding probability values and you multiply each X value with the corresponding probability value and then you are summing them up. Then, we would like to present a very interesting feature of expectation that is the change of variable formula for expectation.

So, sometimes functional transformation of the random variable is taken. Now, if that happens, then how the expectation is going to change. So, here we suppose that some functional transformation has been taken for the continuous random variable X. So, we are interested in expectation of g of X. And then you see the previous formula changes marginally in place of X, you just write g of X and it remains the same otherwise.

Now, we are going to look at the variance of the continuous random variables denoted by Var X. And we have discussed this formula in the previous lecture and it remains the same. So, here I am showing you two formula. One is the first expression and that is the fundamental expression for variance and then if you simplify, you can get the second expression which is very handy for computation or calculation in practice.

Now, we are going to look at the change of variable formula again for mean and variance. And this time, we are going to talk about two interesting things. So, when we talk about a changing variable, or in the distribution that can happen in two ways. One, there could be a change in the location parameter, or there can be no shift of origin in the variable itself. For example, you have some marks in some mathematics test or something and you feel like the score is too low.

So, you want to give 10 points bonus to the students. So, if you add 10 to all existing courses, then this is going to be defined as X plus 10 as a new variable, or you can also call it a new variable Z . Then, the other way to, marks obtained is by multiplying the existing or obtained actual marks by some number, say 1.1. So, if you multiply 1.1 with the existing marks, then that means that you are increasing each number each obtain number by 10 percent.

So, if you do these kinds of adjustments to existing values of continuous random variable X , then how will that impact your expected value and the variance, that is what we are going to see next. So, first the mean. So, you look at the transformed variable which is given by small r times X plus s . So, s is the change of origin. And small r is the change in scale. So, if that happens, you see, the transformed variable says it.

Now, we will have a mean of small r times expectation of X . So, that is the expected value or the mean of the original variable plus the change in origin which is s . Now, if we look at the variance, which is the last red box in the slide, you see that the change of origin does not have any impact on the variance. But now, these variants of X is increased as it is multiplied by r square. So, variance increases.

Now, we are talking about different values of mean and variances, what impact does it have on the shape of the distribution or on the distribution diagram. So, here I am showing you a diagram, where you have X values plotted along the horizontal axis and along the Y - axis, I am plotting the probabilities and you see that I am showing you two curves both with. So, these are the distributions of the same random variables here.

So, it has got, they have got the same mean, but suppose, I scale the variable by one number say positive number r and that will increase the variation. So, it will not impact the mean of the distribution, but it will definitely have an impact on the variance spread will be much more. So, the curve will have more mass towards the tails, both upper and lower. Now, we are going to look at one example of the continuous random variable concept and that is through the example of exponential distribution.

(Refer Slide Time: 14:54)

Exponential distribution

- In a Poisson process, events occur at random and independent of the past, but with a known long term average rate λ per time unit
 - Example: Customers arriving at a store
- The exponential distribution is the probability distribution of the time between events in a Poisson process
 - Example: The amount of time (beginning now) until a cyclone strikes a coastal district of West Bengal

♣ PDF: $f(x) = \begin{cases} \lambda e^{-\lambda x}, & \text{if } x > 0 \\ 0, & \text{otherwise} \end{cases}$ $\lambda > 0$ is a parameter of the $\text{Exp}(\lambda)$ distribution, often called the rate parameter

♣ CDF: $F(x) = \begin{cases} 1 - e^{-\lambda x}, & \text{if } 0 \leq x < \infty \\ 0, & \text{otherwise} \end{cases}$

♣ Mean: $E(X) = \mu = \frac{1}{\lambda}$ ♣ Variance: $\text{Var}(X) = \sigma^2 = \left(\frac{1}{\lambda}\right)^2$

♣ Median or 2nd Quartile, the "middle" value of the random variable, can be found from $F(x) = 0.5$. In this case, $Q2 = \ln(2)/\lambda$.

So, an exponential distribution is linked with the Poisson process or Poisson distribution that we have studied earlier. So, Poisson process and Poisson distribution are not the same thing. So, let me introduce Poisson process separately before we discuss exponential distribution. In a Poisson process events occur at random and independent of the past, but with a known long term average rate lambda per time unit. So, an example could be customers arriving at a store.

So, it can happen that manager of a store is interested to know what is the what is the rate at which the customers arrive at his or her store and he also wants to compute the probability that if one particular customer has arrived, what will be the next time another customer will come? So, what is the gap between these two customers visiting the store? So, from this Poisson process, exponential distribution is proposed and it gives the probability distribution of the time between events in a Poisson process.

So, let me look at an example here. So, an exponential distribution can model the amount of time until a cyclone strikes a coastal district of any state in India say West Bengal. So, with this example and description, let me have a discussion on the PDF and CDF and mean and variance measures of exponential distribution. So, here note that the PDF is looking simple as I have told you that, for a continuous random variable, a PDF is given by a mathematical function or continuous function.

And here this function f_x is defined in the following way. So, if x is positive, then for that, f_x takes value lambda times e to the power minus lambda times x , and if x is 0 or negative, then if f_x takes 0 value. Now, what is this lambda parameter? So, lambda is a parameter of the

exponential lambda distribution, often called the rate parameter. Now, next is CDF, I have already told you that to get CDF, you need to integrate the PDF over its full range of values. So, that is the way you get CDF.

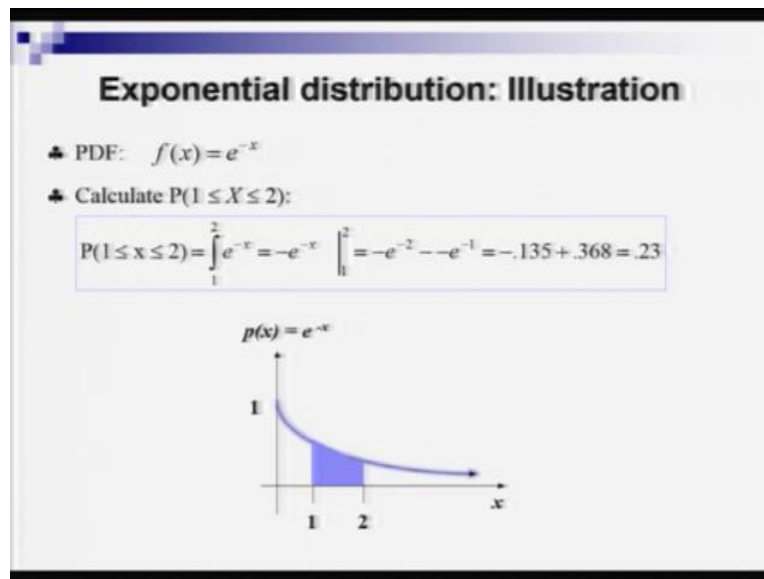
Now, if you take that pain of calculating the integral, then you get capital F_x , and which is again defined in two parts. So, if x takes positive value, so if x is bounded between 0 and plus infinity, then F_x is equal to 1 minus e to the power minus lambda times x , otherwise F_x takes 0 value. Now, it is time to have a look at the mean and variance as I have promised earlier, that I am not going to bother you with lots of computation, because that is not the focus of the course. So, I have already shown you the expression, how to compute mean and variance of continuous random variables.

So, you can follow those steps. And you can check whether by applying this formula to these F_x of exponential distribution, you can get the mean equal to 1 over lambda or not. Similarly, please check whether you can derive variance of X equal to 1 over lambda square or not. Now, in this slide, the last topic we are going to cover is about median. So median, or the second quartile of the continuous random variable X is the middle value of the variable and it can be found from the CDF.

So, you have to do a very simple thing, you have to equate the expression of capital F_x to 0.5 and then solve for the value of x and this is the way you can get the value of median. So, in this particular example, Q_2 value or the median value is \log , natural log of 2 divided by lambda. Now, if I ask you to calculate the third quartile or the fourth quartile from any continuous random variable, then the same strategy can be adopted.

So, you have to first look at the F_x or the CDF of the variable and then you have to equate that to 0.25 for the first quartile, and 0.75 to compute the third quartile value. Now, we are going to look at the exponential distribution using a very simple illustration. Hopefully this illustration will make things a bit more clear to you.

(Refer Slide Time: 20:26)



So, here we take a very simple and this is the simplest possible form of the PDF that one can assume for exponential distribution, small $f(x)$ equal to e to the power minus x . And given that PDF, let me show you how we can calculate the probability that the random variable x takes some value between 1 and 2. So here in the box below, I am showing you how you can compute the definite integral in this case, and then how to arrive at this probability figure. Hopefully, you can figure it out from this example. And, you can calculate the other probabilities.

So, the graph below shows you how to mark the probability in the PDF diagram. So, along the X-axis, I am plotting the values of the random variable x , and along the Y-axis, I am plotting the value of the PDF or the probability of x . And note that this downward sloping curve is my exponential distribution, and it has got an intercept of 1 along the Y-axis. And now, please note that if I take a different form of exponential distribution, the intercept value will change, it is taking the value of 1 because I have chosen this particular form of exponential distribution.

Now, I take two values 1 and 2. And I am interested to calculate the probability of the random variable taking value between these two numbers. And we have already calculated the number which is 0.23. But now, I am showing you here how to mark that area in the diagram. So, here you see that area under the curve, or below the curve, and bounded by two perpendicular lines at the values 1 and 2, gives you the area, and that area is marked with blue color for your identification.

So, now, I am going to introduce the case of normal distribution as an example of continuous random variable and associated distribution. And this concept of normal distribution is probably the most important distribution in the Applied Statistics and Econometric research work. So, before I show you the formula, and boring mathematical expressions, again, let me go back to history. Because as normal distribution is very popular and very useful, it is not a bad idea to spend couple of minutes on the evolution of normal distribution.

So, way back in 18th century, in France, there was a mathematician, and his name was Abraham de Moivre. So, he was interested in statistics of gambling. And often he was consulted by the gamblers and people who used to take risky ventures to find expected values and these and that, what is the probability that they will gain or they will win in a particular bit, and stuff like that. So, de Moivre was pretty much interested in this computation of probabilities and expected values.

So, de Moivre started thinking that if he repeats the experiment of coin tossing, a fair coin tossing for many, many times, how the shape of the frequency distribution or histogram is going to change, because of course, if you say toss a coin, say for only 3, 4 times, you are going to expect only 3, 4 bars or spikes in the quadrant. But if you do it, say for 60 times or 100 times, then there will be more spikes or more bars, each bar showing the relative frequency of the particular value, but then how the curve overall is going to look like.

So, he did some mathematics and he found that if this random experiment of coin tossing is repeated many, many times then the bars become much more closer and the width falls and then as the gap, gap between these two bars or many bars are reducing, then you can approximate the movements of these bars by drawing a continuous curve on the top of these bars. And then this is the way, he accidentally invented the idea of normal curve what we call a normal curve.

Of course, he did not talk about the normal distribution, he did not provide any PDF or CDF as theoretical statistics, talks about, for that we had to wait for 3 more decades. In Germany, famous mathematician and statistician, Carl Friedrich Gauss was working with astronomical data, and he was interested to find some pattern in the measurement error in astronomical calculations, and then he figured out that these observational errors or measurement errors are taking almost, a similar kind of shape this normal curve as proposed by de Moivre, and then he got interested, and he did some mathematics and statistical deviations. And finally, he proposed what is today we call a normal distribution.

(Refer Slide Time: 26:38)

Normal distribution

- ♣ Normal p.d.f. is given by: $f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$
 - Random variable X has infinite range
 - Two parameters: μ and σ are the mean and standard deviation
 - $\pi = 3.142$ and $e = 2.718$ (mathematical constants)
- ♣ Examples:
 - In finance, changes in the logarithm of exchange rates, price indices, and stock market indices are assumed normal
 - Physiological measurements, such as blood pressure of adult humans
- ♣ If a variable's relative frequency distribution is shaped roughly symmetric like a bell, then the variable approximates a normal distribution
- ♣ Percentages for a normally distributed variable are equal to the areas under the associated normal curve

So, with this brief history stuff, let us come back to statistics. So, the normal PDF is given by $f(x)$, and that is a very clumsy and complicated notation compared to the previous notation of exponential but try to remember it because, it is one of the most useful distributions for the course, and not only for this course, but also in practical life. So, it is given by 1 over square root of 2 pi times sigma and then this ratio is multiplied with e to the power minus half times x minus mu divided by sigma whole square.

Now, e of course, we all know it is a mathematical constant taking value 2.718 and pi is also a mathematical constant in all known to us and that takes value 3.142 So, (()) (27:35) this, we have two other symbols in the expression and they are mu and sigma. So, of course, mu probably you can, guess this is the mean, because we have used this symbol mu to represent the expected value or the mean of the random variable and then sigma of course, probably you can again guess that it is the square root of the variance of a random variable, yes, you are absolutely right, if you have guessed so.

So, the normal distribution is defined over two parameters mu and sigma compared to the exponential distribution, which had only one parameter lambda. Now, again, the random variable x, there has to be a range to be defined. So, that we can write the PDF and CDF mathematically. Now, in contrast to exponential distribution, which is defined over only positive values of x, x here can take any value between minus infinity and plus infinity. So, it has got infinite range.

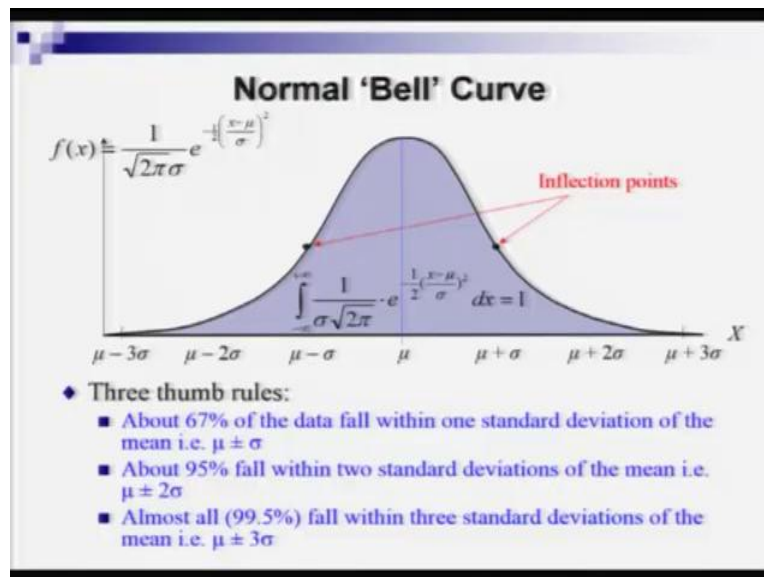
Now, what could be the examples, so, let us start with finance. So, in finance, changes in the logarithm of exchange rates, price indices, and stock market indices are assumed to be

normal. And there are physiological measurements like blood pressure of adult human that also approximately follow normal distribution. Now, what is the linkage between these normal distributions and what we have done before in terms of the frequency distributions of other discrete random variables and all.

So, first, of course, as de Moivre said that, if you repeat one experiment, which gives you only two outcomes, if you repeat that for 100 times the gap between the histogram or frequency diagram reduces and then you get a smooth kind of a curve enveloping these bars. So, now, if this frequency diagram is a relative frequency distribution diagram and if it is roughly symmetric like a bell, then the underneath variable is said to be follow a normal distribution approximately.

So, the percentages or the probabilities for the normally distributed variable in that case are equal to the area under the associated normal curve which is fitted over these bars. But that is in layman's term. Of course, if I asked you to calculate probability and all you cannot give me a rough measure, you have to give me a concrete measure. So, to get the concrete measure of probability, let us look at the next slide.

(Refer Slide Time: 30:30)



So, here, we want to start where we ended the last slide. So, in the last slide, we ended by coining this not coining by emphasizing this term normal curve. And we also said that if the curve looks more or less like a bell in a symmetric one, then you can call that will the underlying random variable may follow a normal distribution. So, here I am showing you a pure bell curve, which is a pure and perfect normal distribution. So, here along the X-axis, I

am measuring the values of the continuous random variable x , and along the Y-axis, I am plotting the small f_x , which are the probability values coming from the PDF.

And note that mean is μ , and that is at the exact center of this bell-shaped curve. So, mean partitions the area under the bell curve into two equal parts are two equal halves. And each, then we will give us the equal area. Now, note that if we take an integration, so, if we now talk about the area under the normal curve, then that can be computed mathematically by these integrals. So, here I am showing you that expression, so, you have to do nothing you have to take that f_x form and then you have to integrate that thing and the integration range will be from minus infinity to plus infinity.

And as usual as the x which follows a normal μ - σ square distribution can take any number in between minus infinity and plus infinity, the range has to be of that sort and then the area or the value of the indefinite integral will be exactly equal to 1. So, if I am talking about this particular property of normal distribution, that it is a perfect symmetric curve, it is a bell shaped curve, then mean μ is the central point of the distribution. And on the left-hand side, we have 50 percent of the mass.

So, if I now change the range of this integration from minus infinity to μ , then we get the value of this integral of 0.5. And then if we change the range again and this time if I now say that my integration will run from μ to plus infinity, then again, the value of the integral is going to be 0.5. So, the area around or in or at the left-hand side and at the right-hand side of the expected value or the mean is equal at 0.5.

Now, note that there are two points on this normal bell curve, and I have indicated them to be the inflection points. So, what do I mean by the inflection points? Now, if you inspect this behavior or the curvature of the normal bell curve, you will see that the slope of the curve is first increasing and after some point of time, it will increase, but it will not increase at an increasing rate as it was doing before, and it will still increase but it will then increase at a decreasing rate.

So, the slope, the slope value will fall. And that juncture, when this movement, or this shift happens in the value of the slope from an increasing slope to you get to see the decreasing slope, that is the point of inflection. And note that it is very easy to find out the inflection points because the first inflection point, which is in the left-hand side of the curve that is to be found by $\mu - \sigma$. That is the first σ difference from the mean value.

So, this is the difference of standard deviation from mean or you can say that this is the deviation and then for the inflection point in the right-hand side of the curve, you see the corresponding x value is $\mu + \sigma$. So, again, it is a deviation that you can see, so, the deviation is by 1 sigma. Now, I would now like to draw your attention to a set of thumb rules, which are very useful in practice. And as I read these thumb rules and talk about them, you may feel like that this is very similar to what we have seen in the case of Chebyshev theorem. Yes, I mean, these results are very much similar to what Chebyshev has told us.

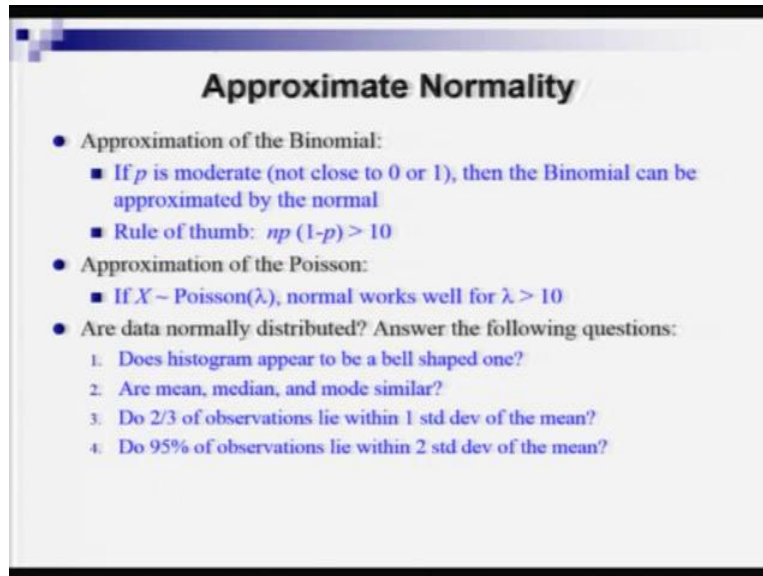
So, the first result says that about 67 percent of the data fall within one standard deviation of the mean, which is $\mu \pm \sigma$. So, if I draw two parallel lines or perpendicular lines at values $\mu - \sigma$ and $\mu + \sigma$ the inflection points, then 67 percent of data will fall in that bounded region. So, that implies that continuous random variable x , if it follows a normal μ σ^2 distribution, then you observe that a value of x will fall between $\mu - \sigma$ and $\mu + \sigma$ is 0.67.

So, if I now want to say this thing again in terms of the integral, then I can take you back to that mathematical formula that you see inside the normal bell curve. And you see, I just have to change the range of the integration here. So, the now it will become a definite integral. So, the lower range value will be $\mu - \sigma$ and then the upper range, I have to erase plus infinity and then I will replace that with $\mu + \sigma$ and then if you calculate the value of the integral it will give you roughly 0.67.

So, I can do this trick, I can now change this minus infinity and plus infinity again to $\mu - 2\sigma$ and $\mu + 2\sigma$ and then the integration will give me a value of 0.95. So, I can say that approximately 95 percent value of the variable, 95 percent of the data fall within two standard deviations of the mean it implies $\mu \pm 2\sigma$ and I can do the same thing for $\mu \pm 3\sigma$ also and in that case, I see almost all values fall in that range.

So, this is not the end of normal distribution, there are many more things that we should learn about normal distribution for practical purposes. But, to end the lecture, I would like to focus on two important things, which are not very technical. First of all, I would like to say is there any link between Binomial, Poisson, and Normal? There is and normal distribution can be used as an approximation to these distributions in special cases. So, that, we are going to see how and the last thing that, I would like to discuss in today's lecture is, how do I know if I am given a data set that this variable can be approximated by a normal distribution or not.

(Refer Slide Time: 38:57)



Approximate Normality

- Approximation of the Binomial:
 - If p is moderate (not close to 0 or 1), then the Binomial can be approximated by the normal
 - Rule of thumb: $np(1-p) > 10$
- Approximation of the Poisson:
 - If $X \sim \text{Poisson}(\lambda)$, normal works well for $\lambda > 10$
- Are data normally distributed? Answer the following questions:
 1. Does histogram appear to be a bell shaped one?
 2. Are mean, median, and mode similar?
 3. Do 2/3 of observations lie within 1 std dev of the mean?
 4. Do 95% of observations lie within 2 std dev of the mean?

So, first I will start with approximation of the binomial distribution. So, there are two results that I would like to mention without showing any proof, these are important results to remember. So, if p is the probability of success in a binomial case is moderate, not close to 0 or 1, then the binomial can be approximated by the normal. So, the other rule of thumb says that the figure which is n times p times 1 minus p , if it is greater than 10 then you can approximate binomial distribution by a normal distribution.

Similarly, in the case of Poisson, if the lambda parameter or the rate parameter takes a value higher than 10, then you can approximate that by normal distribution. Now, I will end with answering this question, is my data normally distributed, can I use normal distribution as an approximation to the data that I am looking at by looking at the histogram or the relative frequency distribution?

So, here you have to answer the following four questions; does your histogram appear to be a bell-shaped one? If yes, then that is a good sign. Are mean median and mode similar or very close? If the answer is yes, then yes, of course, it is a good sign and you can think of normal approximation, then there are two other checks. So, do two-third of observations lie within one standard deviation of the mean?

And the final one is do 95 percent of the observations lie within two standard deviation of the mean? So, if you get answer, yes, for all of these questions from your data, then feel free to go for normal approximation. So, we will end here in the next lecture, we are going to continue our discussion on normal distribution. See you then. Thank you.