Essentials of Data Science with R Software - 1
Professor. Shalabh
Department of Mathematics & Statistics
Indian Institute of Technology, Kanpur
Lecture No. 28
Skewness and Kurtosis

Hello friends welcome to the course Essentials of Data Science with R Software 1 in which we are trying to understand the concepts of probability theory and statistically inference. So, in the last lecture, you can recall that we had a considered the concepts of moments and we had considered the interpretation of the first two moments, which are coming from a raw moment as well as central moments.

Now, we are going to consider here two more aspects of the data. One is Skewness and another is Kurtosis and these are the two very peculiar properties of our data and the question is how to find them, how to measure them quantitatively and then how to interpret them and these two properties they can be obtained by the third and fourth central moments.

So, we once again go back to the definition of the expectation of the function of random variable and we tried to choose two different functions and we tried to see that how they can be used in finding out the two peculiar properties of data which are skewness and kurtosis. So, what are these two properties if I ask you a very simple question suppose you try to count the number of vehicles which are crossing a particular point between say 8am and say 3pm on a particular day. We know the office time is something around 9 to 10 am.

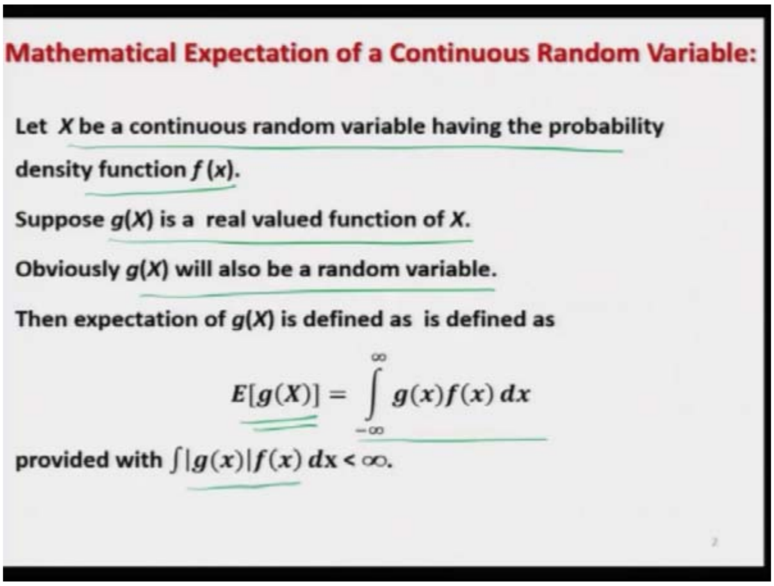
So, what will happen that at say 8 o'clock, the number of vehicles passing through that point will be very, very less and as the time goes from 8 to 9, the number of vehicles will increase and they will become maximum at say, between 9 and 10 and then after that the number of vehicles passing through that particular point will decrease and say at 3 o'clock the number of vehicles will become very, very less.

Now, if you try to plot this phenomena, what do you think how the curve look like, curve will go like this, very sharply, but after that it will decrease sharply, but it will go up to a long distance. So, what will happen that between 8 o'clock in the morning and 11 o'clock, the peak will be very high and this span will be very less, but compared to from 11 o'clock till 3 o'clock, this curve will go down quite slowly and the same thing can, you can also think say between say 11 o'clock till say this 7pm.

So, in this case the curve is not symmetric around particular value, but the data is concentrated more on one side of the curve, either that is left hand side or say right hand side. So, how to take care of this picture, how to take care of this graphic or how to take care of the properties of this phenomenon which is visible inside the graphic and similarly, if you try to plot a curve, the curve will have a hump.

Now, this hump can be higher this hump can be lower with something. So, how to judge this property of the frequency curve, these are the properties which are related to the properties of skewness and kurtosis. So, let us try to begin our lecture and try to see how we can handle these two situations.

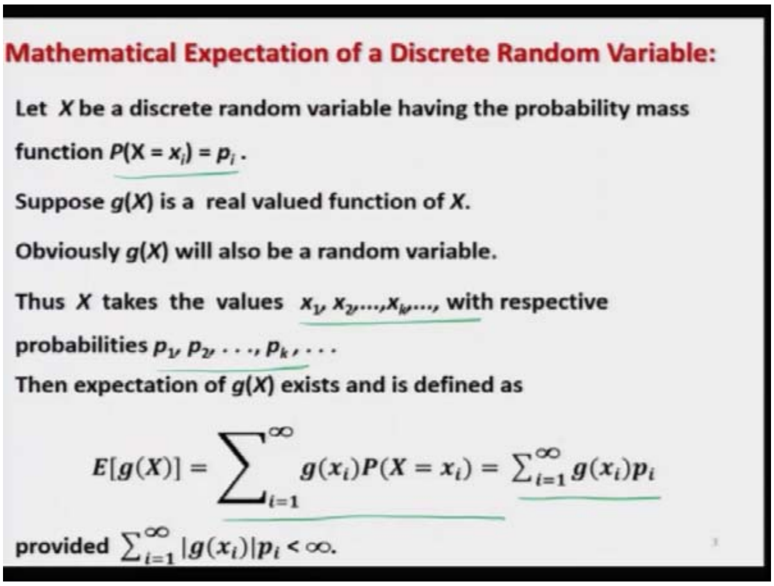
(Refer Slide Time: 04:26)



So, now we come to a slides and just for a quick review, we had considered earlier that let X be a continuous random variable whose probability density function is f(x) and suppose g(X) is a real valued function of the random variable X, so g(X) will also be a random variable and the expectation value of a g(X) is defined as $E[g(X)] = \int_{-\infty}^{\infty} g(x)f(x)\,dx$

and under the condition that this integral exists that is the $\int |g(x)|f(x)\,dx < \infty$.

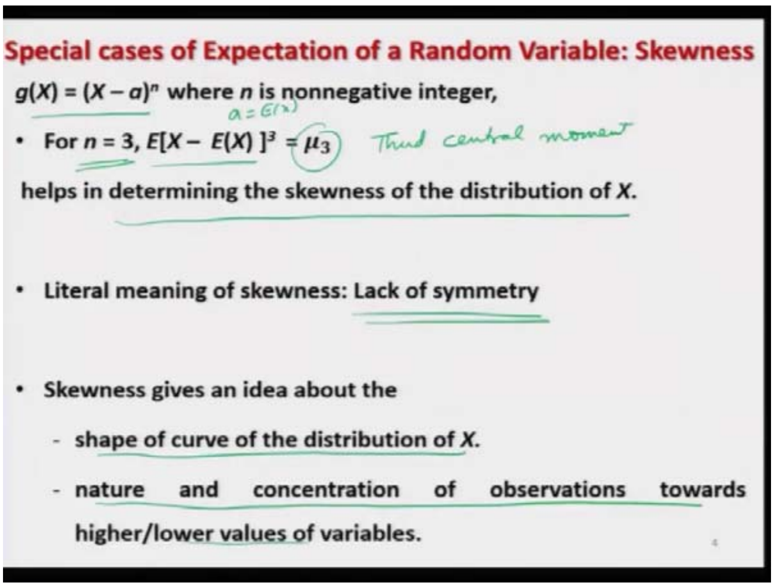
(Refer Slide Time: 05:07)



And the same definition can be given for a discrete random variable X whose probability mass function is given by say probability x equal to xi is equal to pi. So, in this case also the random variable takes the value x1, x2, …,xk … with the respective probabilities p1, p2,…, pk,… and gX is also here a random variable.

So, in this case, the $E[g(X)] = \sum_{i=1}^{\infty} g(x_i)P(X = x_i) = \sum_{i=1}^{\infty} g(x_i)p_i$ and this value can be obtained provided $\sum_{i=1}^{\infty} |g(x_i)|p_i$.

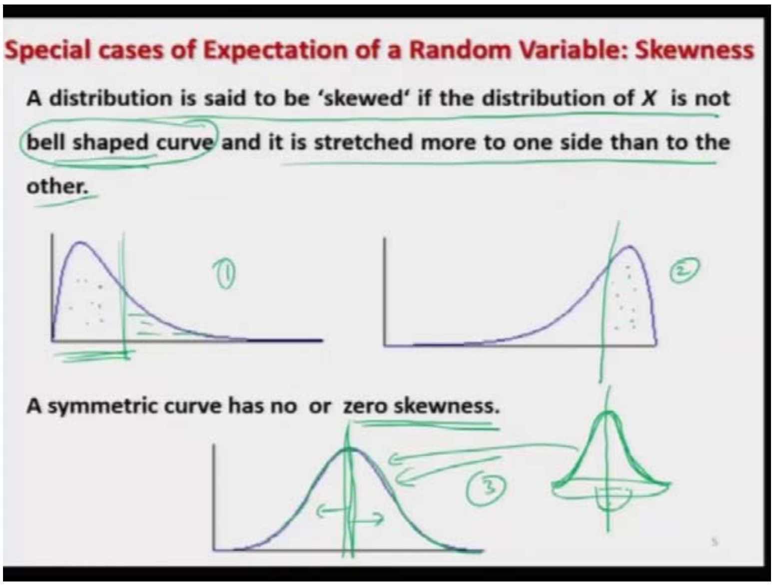
(Refer Slide Time: 05:56)



Now, we try to take the choice of g(X) as X minus a raise to power of your n where n is some non-negative integer and we try to choose here n is equal to 3. So, the expected value of this function will become here say $E[X - E(X)]^3$ and where I am trying to consider this a is equal to expected value of X obviously. So, this is essentially the third central moment.

So, this third central moment which is indicated by μ_3, this helps in determining the skewness of the distribution of X. The literal meaning of a skewness is the lack of symmetry and skewness gives an idea about the shape of the curve of the distribution of X and it also gives us an idea about the nature and concentration of observation towards the higher or lower values of the variables.

(Refer Slide Time: 07:04)

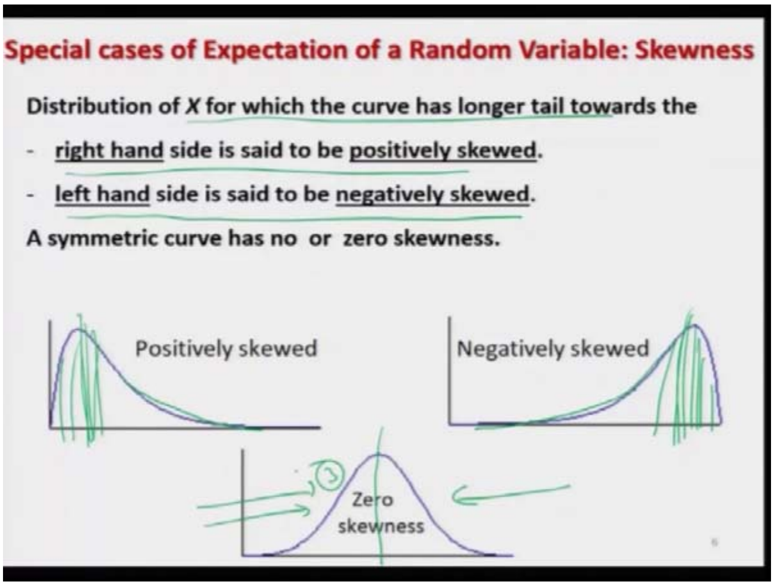


How? For example, if you try to consider here 3 possible curve, let me call it here is 1, 2 and here 3. So, you can see here first in the account number 3 that the curve is starting from here and it is going up to a maximum frequency and then it is decreasing. So, you can see here that this curve is symmetric around this line and you can see here that the distribution of the data or the probabilities on the left hand side of this line and right hand side of this line is nearly the same.

So, we call that this is a symmetric curve and in this case, the curve has got zero skewness and in case if you have curve like as in the figure number 1, you can see here that in this part, the concentration of the values is much more here, then in the remaining part. So, in this case more data is concentrated on the left hand side. And in the figure number 2, the opposite happened that more data is concentrated on the right hand side of the curve.

So, in these two figures 1 and 2, we say that the distribution is skewed. So, one can say that a distribution is skewed if the distribution of X is not bell shaped curve and it is stress more to one side then to the other. What is here a bell shaped curve? You can see here in this case, symmetric that is the figure number 3, do you remember that, how other bell looks like? The bell looks like this one. So, this structure of the bell, this is resembling with this curve and we say that if we try to dissect the belt from the middle, this will be a symmetric structure on both the sides.

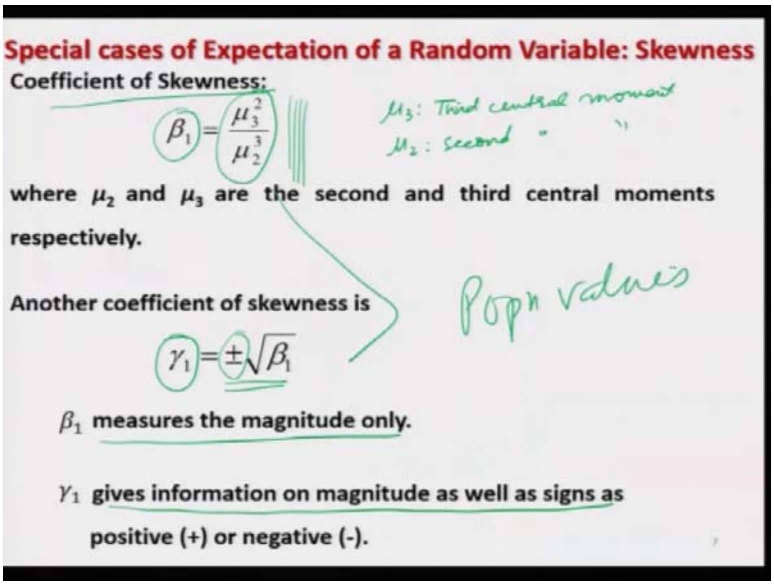
(Refer Slide Time: 09:09)



So, now the distribution of X for which the curve has longer tails towards the hand side. This is called as positively skewed. You can see here that there is a tail is here longer. So, this type of curve where you have more concentration on the left hand side here, this is called as positively skewed curve and the opposite where the distribution of X has a longer tail towards the left hand side is said to be negatively skewed.

You can see here in this case, there is more concentration on the hand side of the curve and the tail here is longer. So, this is called as negatively skewed curve and then if you try to look into this curve here, this is symmetric, so it has got zero skewness. So, this is the property of the skewness that is the lack of symmetry. So, in this case in the figure number this here 3 the curve is symmetric. So, there is no skewness and in curve number 1 and 2 here the curve is not symmetric. So, we call that the curve is skewed.

(Refer Slide Time: 10:27)

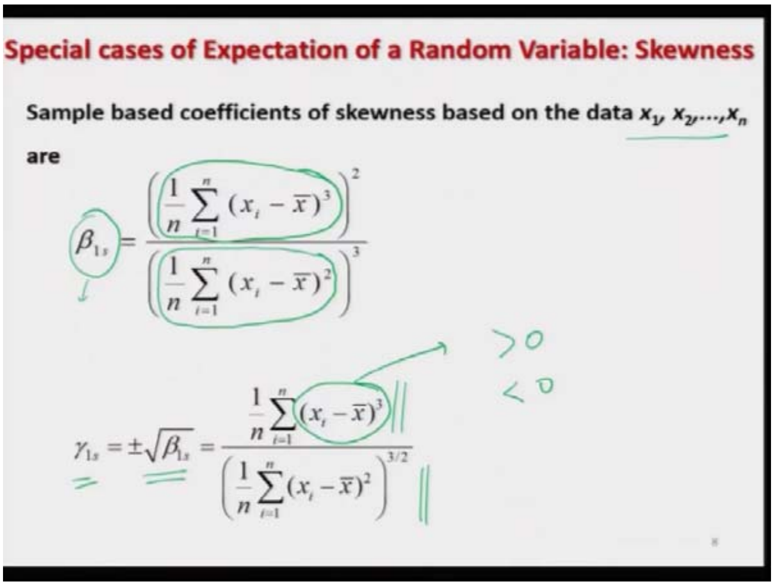


Now, the question is this how to measure this property. In order to measure this lack of symmetry, we have a coefficient of Skewness which is defined as $\frac{\mu_3^2}{\mu_2^3}$, So, μ_3 is your third central moment and μ_2 is your here second central moment, which is actually your here variance and the symbol to indicate this coefficient of skewness is β_1. So, β_1 is defined here as say the square of the third central moment divided by the cube of the second central moment.

So, another coefficient of Skewness is given by the square root of this β_1. So, this can be plus minus square root of β_1 and that is indicated by γ_1 and this is also a measure of skewness and the difference between β_1 and γ_1 is the following. That β_1 measures only the magnitude you can see here this is going to be a positive quantity always.

Whereas, in case of γ_1 this can choose a value say plus or minus. So, this γ_1 gives information on magnitude as well as the signs whether the curve is positively skewed or negatively skewed. Now, the question is this β_1 and γ_1 what you have used here they are based on the population values.

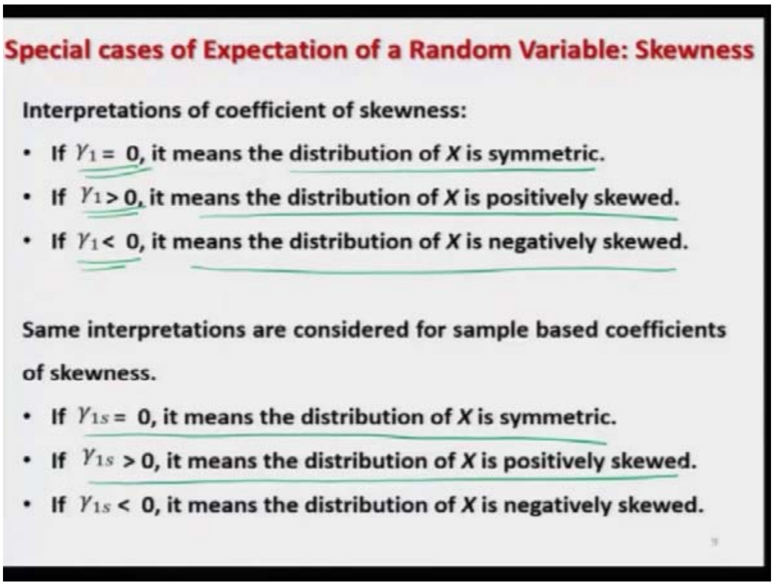
(Refer Slide Time: 12:12)



Now, you are going to use them on the basis of a sample. So, once again we assume that we have got our sample $x_1, x_2, \ldots, x_n$ on a random variable x and now, what we try to do, we try to compute that third and second central moments on the basis of the given cut off data. So, this μ_3 is computed $\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^3$ and variance is the same what you computed earlier as a same sample variance and we try to compute the sample version of the coefficient of Skewness β_1.

So, we indicated by β_{1s}, s mean sample and similarly, for the γ_1 the sample version of γ_1 is given by the square root of this β_{1s} which is here like this one. So, that can take the positive value and negative value you can see here because this is here at cube value, so this can be greater than zero this can be less than zero as well as here this can also be equal to zero, but this sign is going to indicate us whether that skewness is positive or negative.

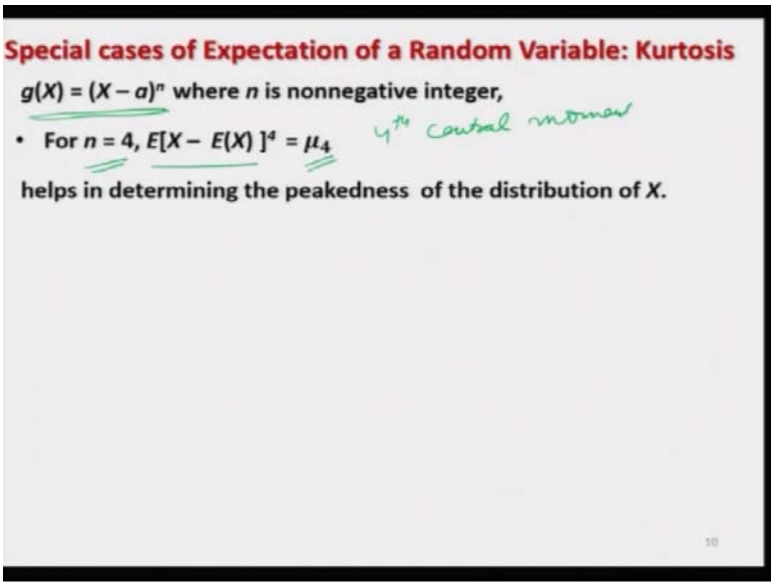
(Refer Slide Time: 13:29)



So, what is the interpretation of this coefficient of Skewness? So, in case if we say that γ_1 is suppose zero, this means the distribution of X is symmetric, if γ_1 is positive, it means that the distribution of X is positively skewed and if γ_1 is less than zero, it means the distribution of X is negatively skewed.

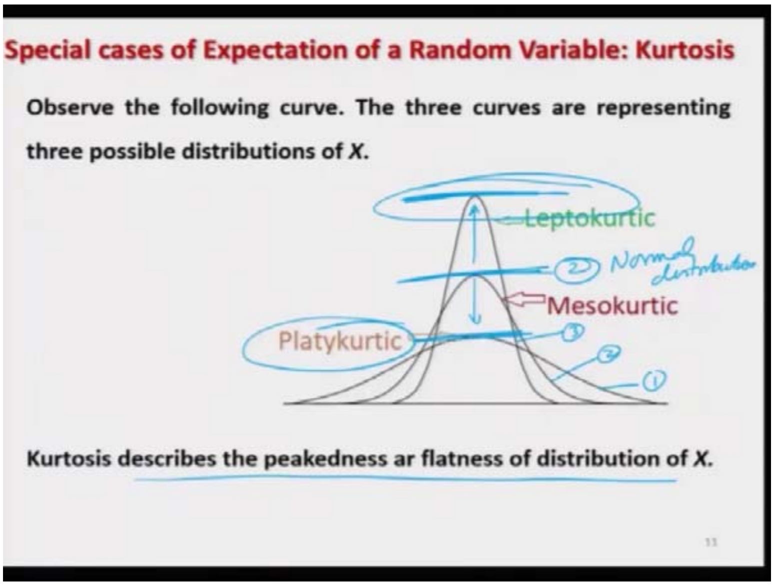
And the same interpretation is continued for our sample base coefficient of Skewness that if γ_{1s} is equal to zero this means that the distribution of X is symmetric and if γ_{1s} is greater than zero, it means that distribution of X is positively skewed and if it is negative, this means direct distribution of X is negatively skewed. So, this is what we tried to do.

(Refer Slide Time: 14:15)



And now, instead of clicking here n equal to 3 now, in case if I try to consider here g(X) and if I try to take n equal to 4, then we have here the fourth central moment, how this will be $E[X - E(X)]^4$, which is indicated by here μ_4 and this quantity helps in determining the peakedness of the distribution of the random variable X.

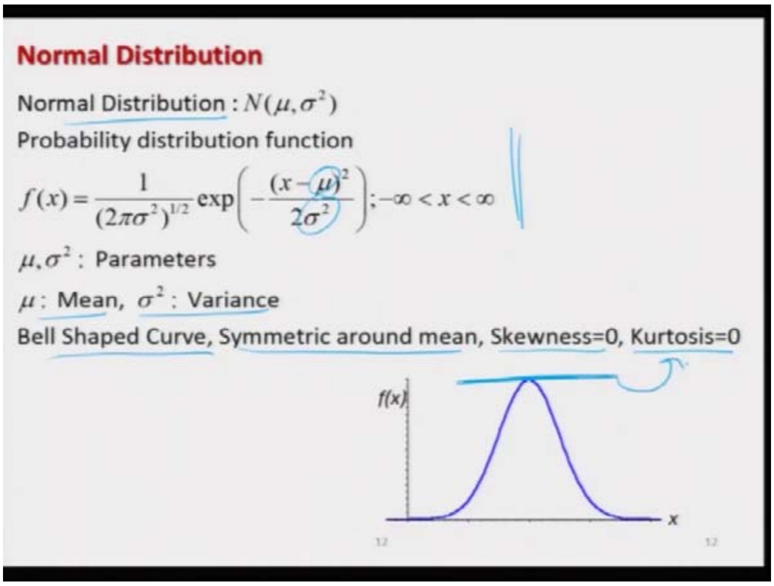
(Refer Slide Time: 14:47)



Now, suppose we try to consider here these three curves, let them call it a curve number 1, curve number here 2 and here, curve number here 3 and try to observe this part here, this hump, you can see here that if you try to compare the hump of curve number 1 and 3, then the hump of the curve number 1 is smaller than the hump of the curve number 2 and the hump of the curve number 3 is higher than the hump of the curve number 2. So, we try to compare the humps with respect to the curve number 2, the question is how this curve number 2 will come into existence, this I will try to address.

So, in case if you try to see if the hump of the curve is just like as in the curve number 2 here, like this one, then the curve is called as Mesokurtic and if a hump has the higher hump than the Mesokurtic curves hump, then it is then the curve is called as Leptokurtic and in case if the hump of the curve is lower than the hump of the curve of Mesokurtic curve, then it is called as Platykurtic. So, this is the property of the kurtosis, which describes the peakedness or flatness of the curve of the distribution of X or the distribution of X.

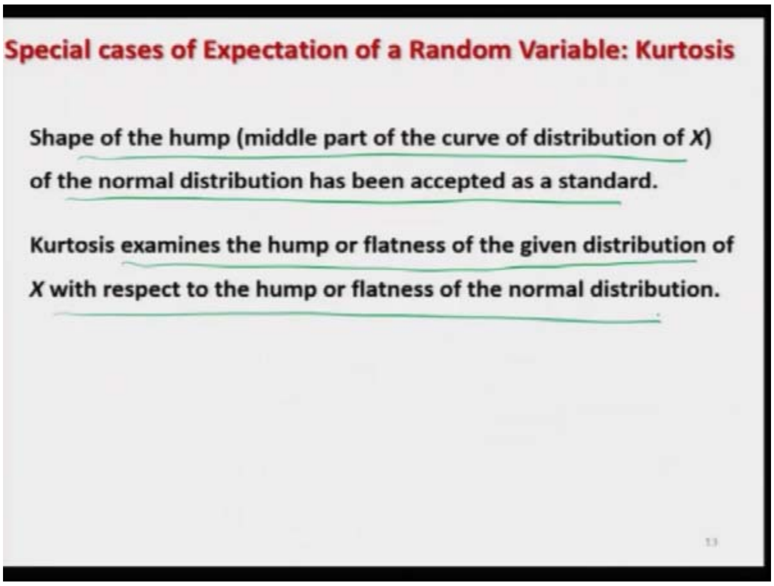
(Refer Slide Time: 16:25)



So and now, I am just trying to give you here an idea that we have a probability density function which is normal distribution whose probability density function is given like this, whose parameter here are here μ and σ^2, μ gives the idea about the mean and σ^2 gives the idea about the variance and this is a bell shaped curve, it is symmetric around mean and it has skewness zero and kurtosis zero. What is this normal distribution? This I will try to explain in much detail in the forthcoming lecture, but at this moment you can just take it on its face value.

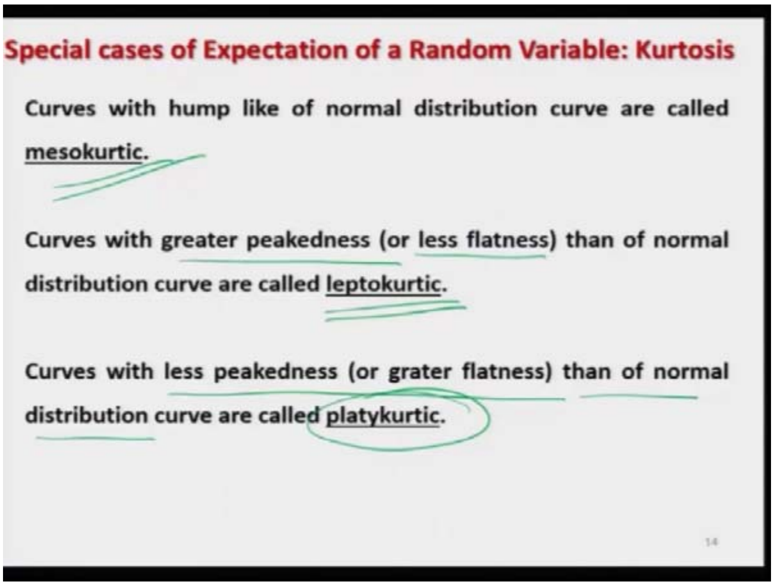
So, whatever is the coefficient of skewness and kurtosis for this normal distribution, they are taken as the standard one. So, this is a symmetry curve, so skewness is zero and whatever it is hump that we assume this is equal to zero and then we try to compare the hump of this curve number 1 and 3 from here. So, the hump of the curve number 2 is coming from normal distribution that you have to just understand at this moment and after that, I will try to explain you in more detail.

(Refer Slide Time: 17:34)



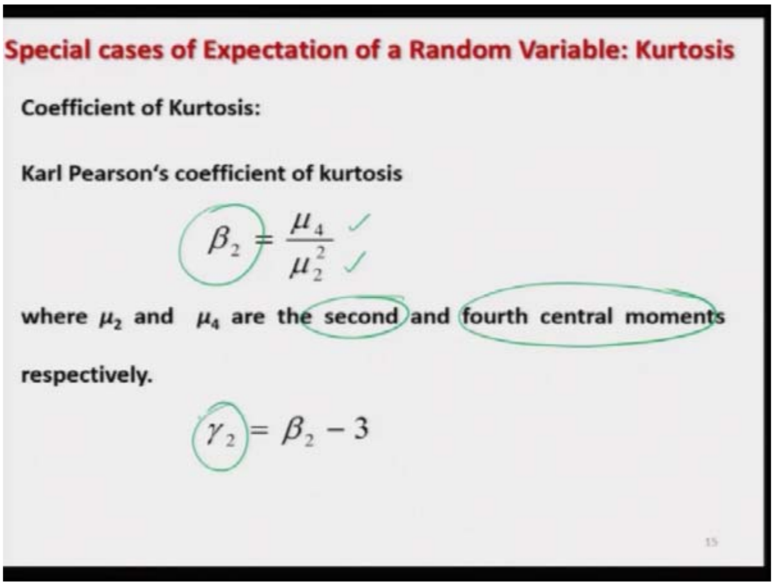
So, the shape of the hump or the middle part of the curve of the distribution of X of the normal distribution has been accepted as a standard one and the kurtosis examines the hump or flatness of the given distribution of X with respect to the hump or flatness of the normal distribution.

(Refer Slide Time: 17:55)



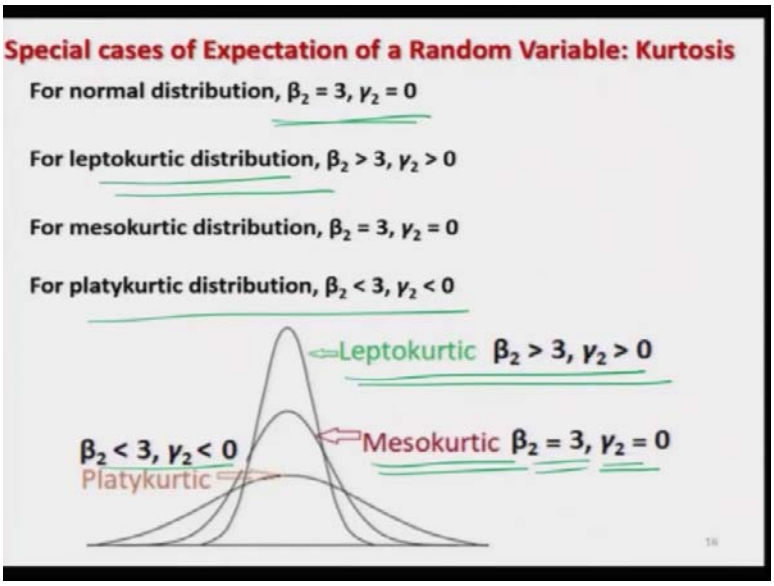
So, curves with hump like of normal distribution they are called as mesokurtic, curves with greater peakedness or less flatness than of the normal distribution curve they are called as leptokurtic and those curves which has got less peakedness or greater flatness, than of the normal distribution they are called as platykurtic. This is the basic definition of the kurtosis and this is how we try to divide the curve into these three categories.

(Refer Slide Time: 18:24)



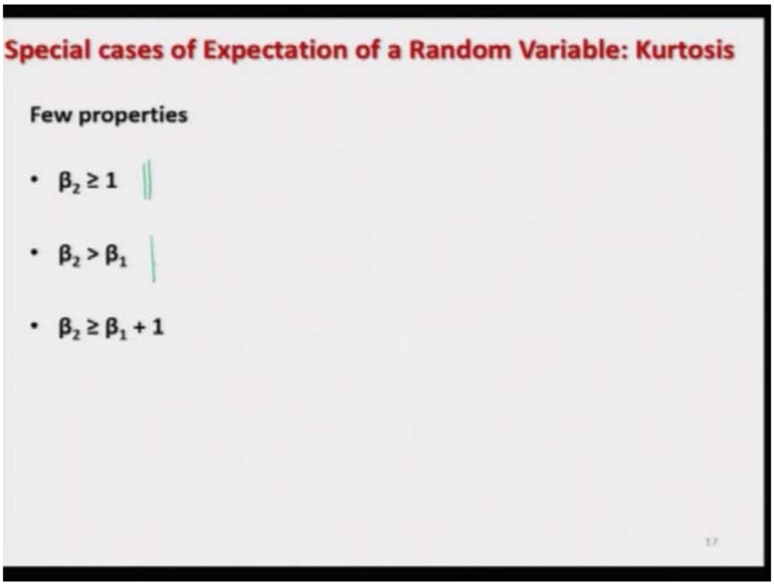
Now, the question is how to measure this property. So, for that we have coefficient of kurtosis and the Karl Pearson's coefficient of kurtosis is defined here like this $\frac{\mu_4}{\mu_2^2}$. So, now you can see here, μ_4 is the fourth central moment and μ_2 is the second central moment. So, you can see here that the fourth moment is now trying to help you in determining the property of kurtosis and this is the beta coefficient of kurtosis and similarly, we have defined here γ_2 also, $\gamma_2 = \beta_2 - 3$. So, this is called a gamma coefficient of kurtosis.

(Refer Slide Time: 19:09)



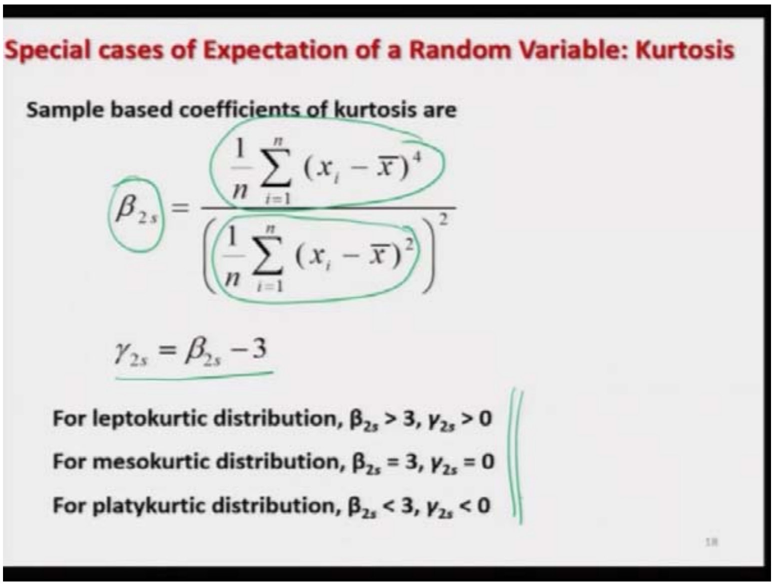
So, now the interpretation based on the value of β_2 and γ_2 is like follows, for the mesokurtic distribution, we have β_2 equal to 3 and γ_2 equal to zero like this here and for leptokurtic distribution, you can see here this is here β_2 greater than three and γ_2 greater than zero, which is positive and for the platykurtic distribution, we have β_2 less than 3 and γ_2 less than zero. So, we simply try to compute the values of γ_2 and β_2 and based on that we try to take a conclusion whether the given distribution is leptokurtic, mesokurtic or platykurtic.

(Refer Slide Time: 19:54)



Just for the information. These are some properties of this coefficients Skewness and Kurtosis that β_2 is always greater than equal to 1, β_2 is greater than β_1 and β_2 is greater than or equal to β_1 + 1. Well, I am not going to use it here.

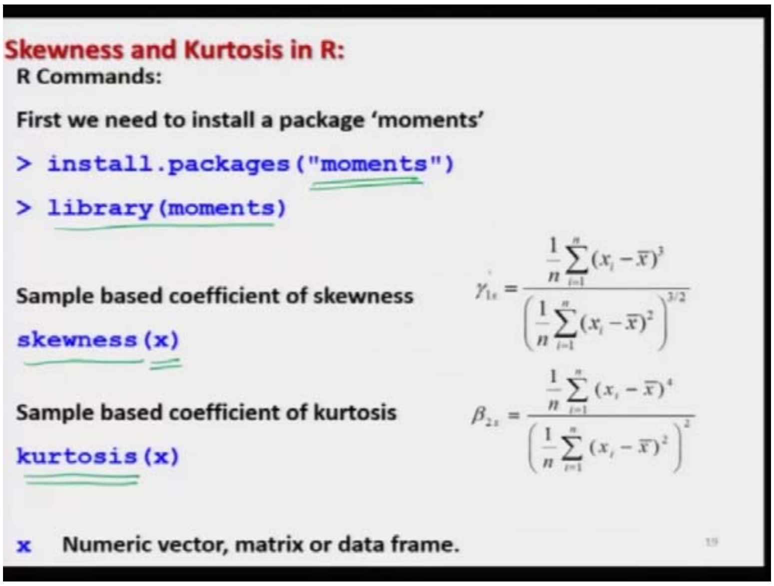
(Refer Slide Time: 20:06)



But once again, the same question comes, these are β_2 and γ_2 they are defined for the population values and we want to compute them on the basis of sample. So, what we try to do, we simply try to compute the fourth central moment based on the data set $x_1, x_2, \ldots, x_n$ and we try to compute the variance or the second central moment based on that data set x1, x2,…, xn and then we try to compute the β_2 and I am indicating it by β_{2s}, s means sample based.

And similarly, the γ_2 can be defined as γ_{2s}, which is β_{2s} minus 3 and the interpretation of leptokurtic, mesokurtic or platykurtic is going to be the same as in the case of γ_2 and β_2, there is no change.

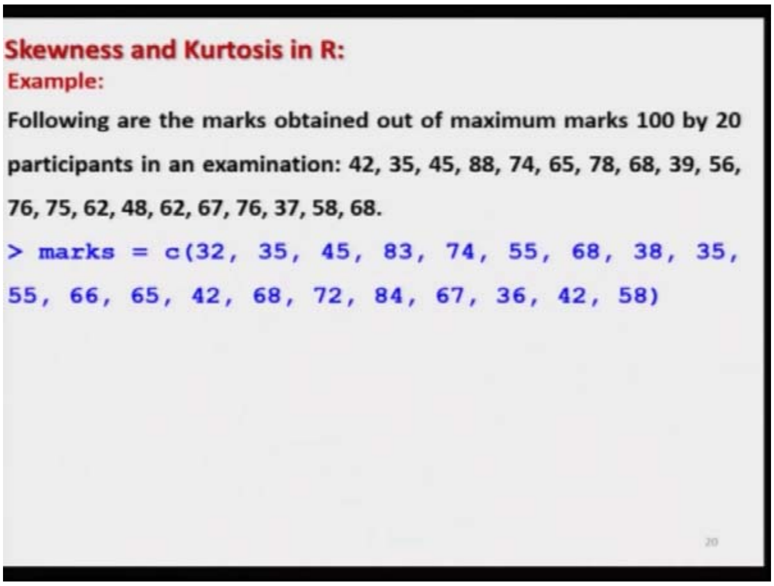
(Refer Slide Time: 20:55)



So, now the question is how you can compute these coefficient of skewness and kurtosis in the R software. So for that we need here a special package moments that is the same package that we had used earlier also, while computing the moments and then we tried to load it using the command library. Now, the sample based coefficient of skewness can be obtained by the command skewness, s k e w n e s s, all in lowercase alphabets and inside the parentheses you have to write the data vector.

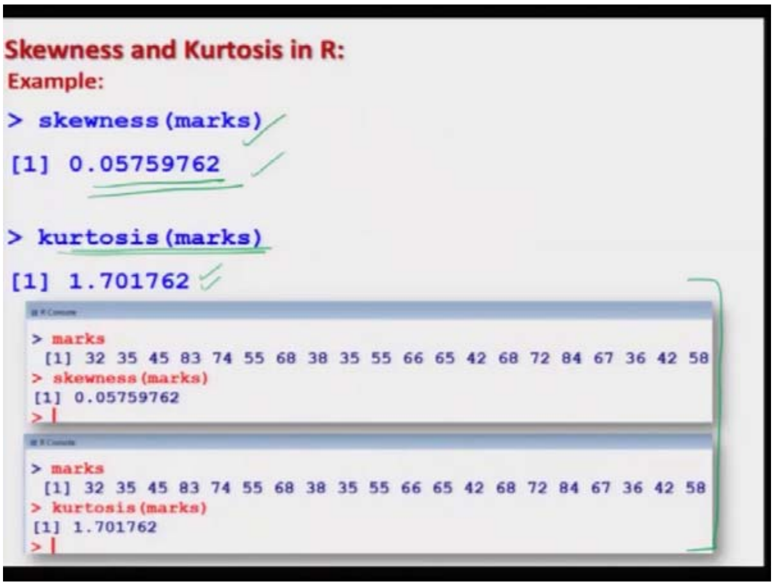
And similarly, the sample base coefficient or kurtosis can be obtained by the command kurtosis k u r t o s i s is and inside the parentheses write down the data vector x, this data vector can be a numeric vector matrix or a data frame also. But what you have to keep in mind that R computes the gamma coefficient for the coefficient of skewness and it computes the beta coefficient for the coefficient of kurtosis that you have to keep in mind while interpreting the data.

(Refer Slide Time: 21:59)



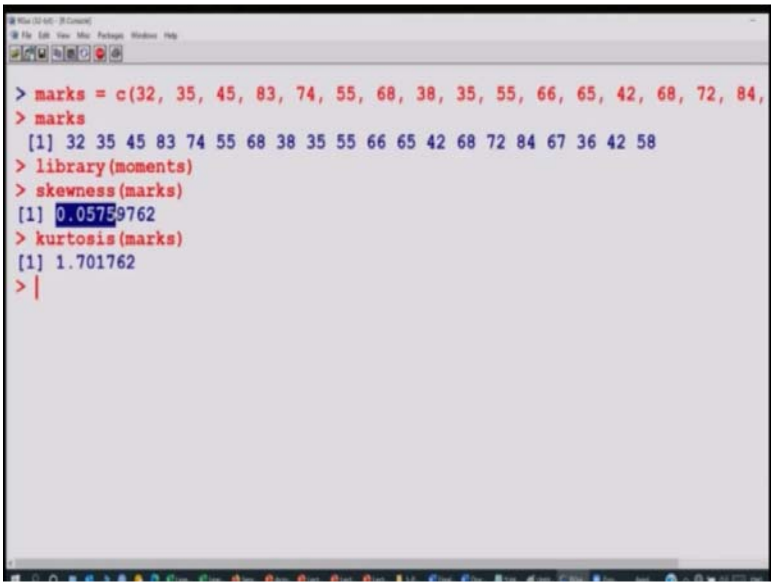
So, now the life becomes very simple, if I can show you that how are you going to compute it? So, we try to consider here the same example that we consider earlier in the computation of moments that we try to consider the marks of 20 participant in an examination, the marks are obtained out of 100 and these are the marks which are stored in a data vector here, marks.

(Refer Slide Time: 22:27)



Now, we try to find out the coefficient of skewness and kurtosis using the command is skewness and kurtosis, so skewness of marks give will give you the value 0.05 and the value of the kurtosis here is 1.70. So, now using the rules, what you have learnt here, in this case, you can just take a proper conclusion, whether your distribution is positively skewed negatively skewed or leptokurtic, mesokurtic or platykurtic whatever you want, you can do it and these are the screenshot of the same operation. So, but now let me try to show you these things on the R console, so let us try to come to the R console.

(Refer Slide Time: 23:13)



So, first let me try to enter here the data. So, you can see here these are the marks and now I have to first upload the library moments, this package is already there in my computer. So, I am not installing it, but it is, but it may not be on your computer. So, you please install it first. And now if you want to see what is the skewness, you can see here a skewness of the data set marks and it will come out to be here like this and if you want to know the kurtosis that it is coming out to be here, like this kurtosis of marks 1.70.

So, you can see here the value of this skewness is 0.05. So, you can see that the structure of the data is nearly symmetric and similarly you can take the conclusion about the kurtosis of the distribution also. So, now we come to an end to this lecture and now you can see here that we have, we started with the definition of moments and initially I am sure that it must have looked like we are going to discuss about the theoretical property, there is some theoretical expectation.

But now you can see that these moments are really going to help us and they are going to give us different types of information which are hidden inside the data which you possibly cannot see from your eyes. You can think that you have got a say 20,000 observations or 1 million observation and you have no idea that what is the central tendency of the data, variance of the data is skewness kurtosis etcetera.

So, by using just these tools, you can get a fairly good information about the structure of the data and the different properties which are hidden inside the data. The question comes here, why we have done up to here moments 4, what is the interpretation of fifth, sixth, or say higher order moments?

The fact is this up to now, we know only about the interpretation given by the first four moments what means, other higher order moments are going to indicate us that is not known to us. But computation of higher order moments is not difficult at all and people are looking forward that that how this fifth and sixth or say higher order moments are going to reveal some other properties of the curve of the distribution of X or the distribution of X.

So, my objective here is this, why do not you take a small data set and try to compute this mean, variance, skewness, kurtosis etc, try to plot it and try to see whether these concepts are matching with the real data or not and how are you going to interpret it that is the most important part. The day you can interpret the data correctly, you can read what data is trying to inform you through this values you can be a very good data scientist that is my promise to you and that is what we are actually doing.

So, I stop in this lecture and you try to practice it and I will see you in the next lecture with more topics till then, goodbye.