

Lecture - 24

Moments-Skewness and Kurtosis --- old

Welcome to the next lecture on the course, descriptive aesthetics with R Software. You may recall that, in the last two lectures, we had discussed the concept of movements. And we discussed the raw movements, central movements and absolute movements. We also learned how to compute them, on R Software. Now, I am, going to introduce here, the concept of skewness and kurtosis, which are again, the two features of a frequency curve and our objective is that, to understand firstly, what are these features? Secondly how to quantify them? And thirdly how to compute them on the R Software? When we are trying to quantify them, then you will see that, we will need the definition of the movements and in particular the

central movements. And that was the reason, that I had, explained those concepts earlier. So now, let us start this discussion.

Refer Slide Time :(1:39)

Skewness

Literal meaning of skewness: Lack of symmetry

freq curve

Skewness gives an idea about the

- shape of curve obtained by frequency distribution (frequency curve) of data.
- nature and concentration of observations towards higher/lower values of variables.

more data

of person (f)

9:00 / 42:42

And first we try to understand, what is the skewness? The dictionary or the literal meaning of the skewness: is lack of symmetry. What does this mean? This symmetry is, talking about the symmetry of the, frequency curve or the frequency density; you have seen that, how we have computed the frequency table, from there we had drawn the frequency density curve. So, when we say that, this is the lack of symmetry. Then what is symmetry? Symmetry here is like this, so this is the basic meaning of symmetry. So now, I am saying that, this symmetry is lacking, when the symmetry is lacking, what will happen? Means in, ideal situation if I say, suppose if I say this is my symmetric curve, then the symmetry is disturbed mean, this curve will look like this or this curve will look like this. Now, what is the interpretation of these curves? Suppose I try to take an example, where we are counting, the number of persons passing through a place, where many, many offices are located. So, we know, what is the phenomena? The phenomena is like this; that usually the office will start at, nine o'clock, ten o'clock, in the morning. The traffic at that point will be very less say around 7 a.m., 8 a.m., in the morning. And then, the traffic will start increasing and then, it will increase say up to 10 o'clock 10:30 or say 11 o'clock in the morning and after that the traffic will decrease.

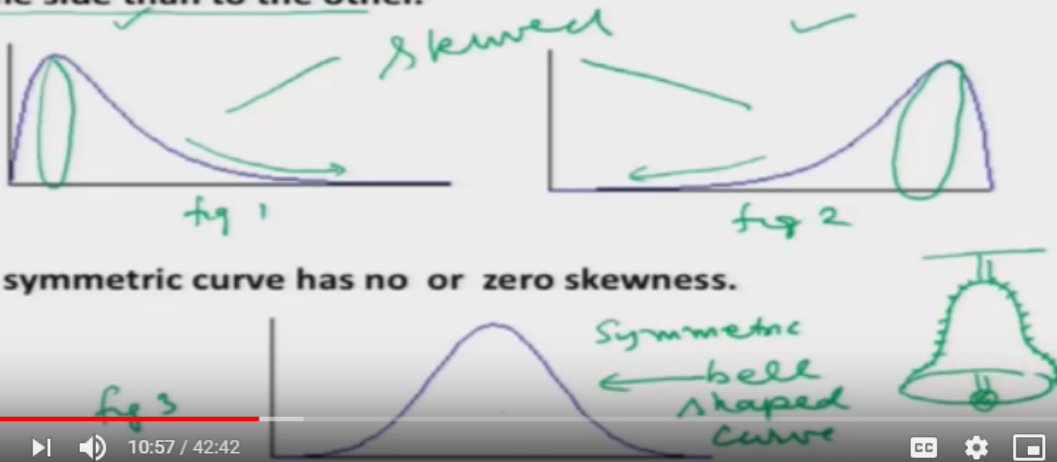
So, in case if I want to, show this phenomena through a curve, this curve will look like this, suppose if I say, this is the time here, I'd say here, 10 a.m.. And this is the time here, somewhere here say, 7 a.m... And this is that time here, up to here, say here 3 p.m. to 10:00 a.m., 11 a.m. and so on. And here is the number of persons passing through that point. So, I can say here that from 7 a.m. this frequency or the number of persons, this is very small, it starts increasing and then it keeps on increasing up to say 10 o'clock and then after that everybody comes in the office and then, there are less number of people, who are coming to office and then finally, see if you come up to here, 3 p.m. this number will decrease. Now, on the other

hand, the opposite happens in the third case, suppose if I try to mark these points, as here I said 12:00 p.m., 1 p.m., I'm up to here sometime here, say here 6 p.m. and then here 7 p.m. and say here 9 p.m. and here we try to count the, same did, same record, the same data that is the number of person, which is denoting the frequency. So now, what will happen once the office starts and offices from say 9:00 to 5:00 or say 9 a.m. to 6 p.m. or so, so usually in that marketplace are in the that place, where there are many, many offices, people will be working inside the office and then, in the evening when the office hours, closes then they will leave the office. So what will happen? Say from say 12 o'clock or 1 o'clock, the number of person passing through that point will be very less. And this number will start increasing say from, 4 p.m. 5 p.m. and it will be maximum say, between say 5 p.m. and 6 p.m. and once everybody has, left the office, then the number of person passing through that point will, sharply decrease and say at 7 p.m., 8 p.m. the number will be very, very less. So, now how to denote this phenomena through a curve, so this type of phenomena can be expressed by, this curve. So, initially at 12 o'clock the number is very, very less. And then, say around 5 p.m., 6 p.m., the number is increasing and then it is decreasing after, say 7 p.m. or so on. In both these cases, what is happening, you can see here, more data is concentrated here and the first figure and more data is concentrated on the right-hand side, in the last figure. So, these are the areas in red color, where more data is concentrated. Now, if you try to take that third figure, you can see here in this case, the curve is symmetric around the mid value. If you try to break the curve into two parts, from this point and if you try to fold it and if you keep it or dis thing, then this will look symmetrical. So what I'm going to say? Suppose the curve is like this and if I try to break it in the mid and if I fold it, then the curve will look the same. So, this is what we mean by symmetry. And this feature is missing in the, first and last curve. That if you try to break the first curve add to this point and the last curve at this point, where I am moving my pin, then this will not be symmetric. So, the objective here is, how to study this departure from symmetry, on the basis of given set of data, I would like to know, on the basis or given set of data. That whether the data is concentrated, on the left hand side more or more concentrated on the right hand side of the frequency curve. So, this feature is called as, 'Skewness'. And in order to quantify it, we have a coefficient of skewness. So now, I can say that here is skewness gives us an idea about the, shape of the curve which is obtained by the frequency distribution or the frequency curve of the data. And it indicates us, the nature and concentration of observation towards, higher or lower values of the variable.

Refer Slide Time :(9:03)

Skewness

A distribution is said to be 'skewed' if the frequency curve of the distribution is not bell shaped curve and it is stretched more to one side than to the other.



A symmetric curve has no or zero skewness.

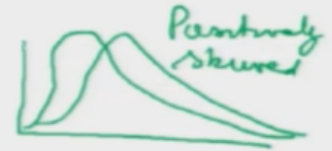
So, you can see here, I have plotted here, three figures. I can call it; say head figure 1, figure 2 and here that this is a figure 3. So, I will call this figure number three, as bell-shaped curve, why this is called bell-shaped? Have you ever seen a bell, the bell is like this and then, there is here is a ring. Right? So, you can see the structure of this curve here, this is symmetric, so that is why this structure shape in the figure 3 is called as, 'Bell Shaped Curve'. So, the bell shaped curve I will say, this is a symmetric curve. Now, when the symmetry is lacking, then the frequency curve will look like, as in Figure 1 of Figure 2. So, the curve is or the frequency curve is, more stressed towards right or towards left, this is indicating that more values are concentrated in this region, in Figure 1 and more values are concentrated in the, in this region in the figure 2. In these cases, we say that the curves are skewed. So, our frequency distribution is said to be skewed, if the frequency curve of the distribution is not the bell shaped curve and it is stress more on, one side then to the other. Now, how to identify it, because now we have two types of lack of symmetry, one in Figure 1 and 1 in Figure 2. So, we try to give it here a name.

Refer Slide Time :(11:01)

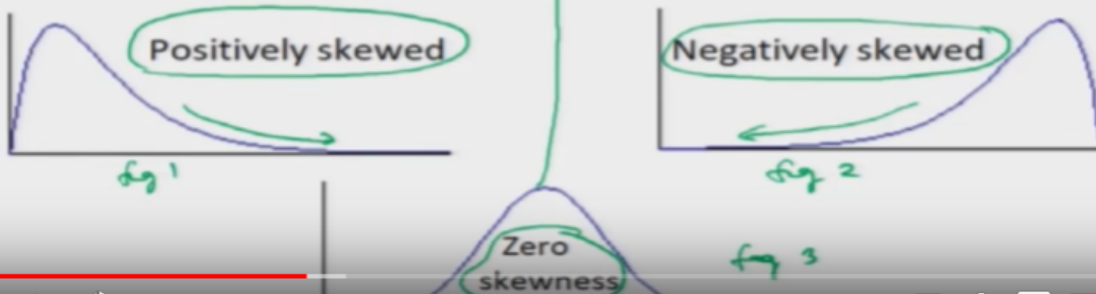
Skewness

Frequency distribution for which the curve has longer tail towards the

- right hand side is said to be positively skewed.
- left hand side is said to be negatively skewed.



A symmetric curve has no or zero skewness.



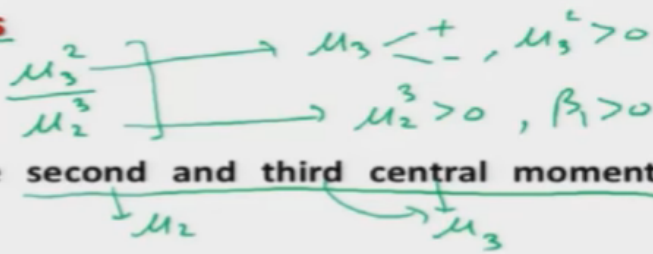
So let me rename the figure, this is Figure 1, this is Figure 2 and this is figure 3, from the last slide. So, now in the figure 1 you can see that the curve is more stress on the right hand side. So, when the curve is more stress on the right-hand side, this is called as, 'Positively Skewed Curve'. And similarly if the curve is more staged on the left-hand side, then this is called as a, 'Negatively Skewed Curve'. And in case of a symmetric curve, we assume that the curve is symmetric and we say that it has got, zero skewness. When we want to discuss the property of his skewness, we try to write whether the frequency curve is positively skewed, negatively skewed or it is symmetric. And this is how we try to express the, finding from the data. But, definitely there will be one thing, suppose if I try to take here, two curves, like this and like this, both the curves are, positively skewed. So, the finnex coefficient is both the curves are positively skewed, but their structure is different, one curve is lacking the symmetry more than the others. But, just by saying, less or more it will not help us we need to quantify it. So, our next objective is that, how to quantify this lack of symmetry. And in order to understand this thing, we have a concept of coefficient of skewness.

Refer Slide Time :(13:02)

Coefficient of Skewness

$$\beta_1 = \frac{\mu_3^2}{\mu_2^3}$$

where μ_2 and μ_3 are the second and third central moments respectively.



Another coefficient of skewness is

$$\gamma_1 = \pm \sqrt{\beta_1}$$



β_1 measures the magnitude only.

γ_1 gives information on magnitude as well as signs as positive (+) or negative (-).

And the definition of coefficient of skewness depends on, the second and third central movements of the data. So, you may recall that, we had denoted the second central moment by μ_2 and third central moment by μ_3 . So now, the coefficient of skewness is denoted by β_1 . And this is defined as, the square of third central moment, divided by the cube of second central moment. And this is called the, 'Coefficient of Skewness'. There is another coefficient of skewness, which is defined as, the square root of this β_1 and this is denoted by γ_1 . Now, what is the difference between the two measures of coefficient of skewness that is β_1 and γ_1 ? You can see here that in the case of β_1 , μ_3 can be positive or μ_3 can be negative. But, μ_3^2 will always be positive. And similarly, μ_2 will always be positive, so μ_2^3 will always be positive. So, this β_1 will always be positive. So, this will give us the information, on the magnitude of the skewness or the magnitude of the lack of symmetry. But, this β_1 will, not be able to inform us, whether the, the skewness is positive or negative. Where is, when we are trying to use the, next coefficient γ_1 , then what we try to do here: that γ_1 will also give us information about the sign. And when I try to combine, the information obtained by β_1 , then this γ_1 will, give us the information on the magnitude, as well as, the sign. Sign can be positive; sign can be negative, indicating the positive or negative say skewness. So, this is the basic difference between the two measures, β_1 and γ_1 . And you will see that in the R Software, R Software provides the value of γ_1 . And one thing now you have to notice here and I can explain you on this slide itself, that I have defined here, $\beta_1 + \gamma_1$, this is for the population. But, what happens to us,

Refer slide time :(15:55)

Coefficient of Skewness

Sample based coefficients of skewness are

$$\beta_{1s} = \frac{\left(\frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^3 \right)^2}{\left(\frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2 \right)^3}$$

$$\gamma_{1s} = \pm \sqrt{\beta_{1s}} = \frac{\frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^3}{\left(\frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2 \right)^{3/2}}$$

data set
 x_1, x_2, \dots, x_n
 μ_2
 μ_3
↓
Sample moments
2, 3

mag. &
sign

We have got a data set say X_1, X_2 and here see here X_n in this case, what we try to do? We try to compute the value of μ_2 and μ_3 , on the basis of data x_1, x_2, x_n or they are called as, 'Sample Movements'. So, we try to compute the sample movements, of order 2 & 3 and we try to replace, μ_2 and μ_3 by their sample movements. So, in this case, this β_1 I am trying to denote it by β_{1s} that means β_1 based on the sample values, becomes like this. So, that is the same thing, I simply have computed, the second and third, central moments and I have replaced him at in the definition of β_1 . Next the coefficient of γ_1 , now I am denoting by γ_{1s} , s means sample, this becomes here, this square root of β_{1s} and it is given by here like this simply, the square root of the β_{1s} . So, now this will give us the information, on the magnitude and sign and where this β_1 will, give us information only on the magnitude.

Refer slide time :(17:20)

Coefficient of Skewness

Interpretations:

- If $\gamma_1 = 0$, it means the distribution is symmetric.
- If $\gamma_1 > 0$, it means the distribution is positively skewed.
- If $\gamma_1 < 0$, it means the distribution is negatively skewed.

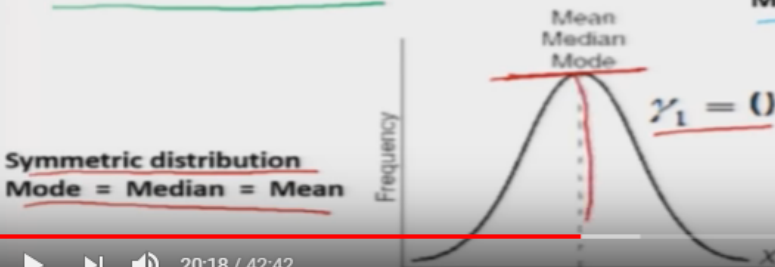
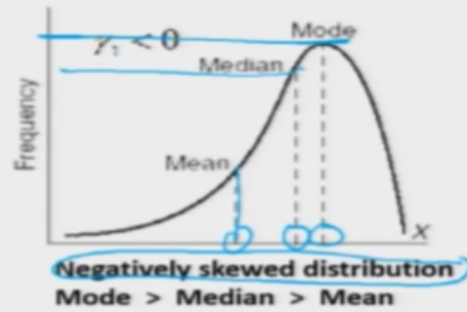
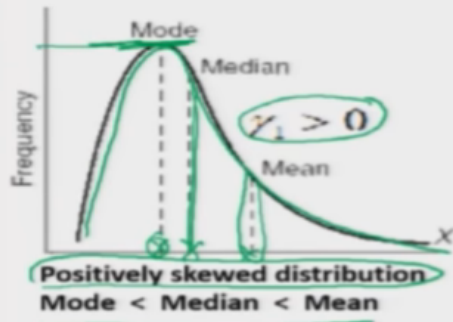
Same interpretations are considered for sample based coefficients of skewness.

- If $\gamma_{1s} = 0$, it means the distribution is symmetric.
- If $\gamma_{1s} > 0$, it means the distribution is positively skewed.
- If $\gamma_{1s} < 0$, it means the distribution is negatively skewed.

Now what is the interpretation? The interpretation goes like this. So, I'm trying to divide the interpretation, based on gamma 1 and gamma 1 s and both are actually the same? So, the first case is if I say gamma 1 is equal to 0, this means the distribution is symmetric, when I say gamma 1 is greater than 0 that is positive, then the distribution is also said to be positively skewed. And if gamma 1 is negative, then the distribution is negatively skewed. And the same continues, for the definition of gamma 1 s, gamma 1 s, s is 0, this means symmetry, if gamma 1 is positive that means that distribution is positively skewed, if gamma 1 is negative then the distribution is negatively skewed. So, you can see here, now that having the coefficient of skewness, it is not difficult to know, the feature of the frequency curve with respect to the symmetry and I can see whether my distribution is symmetric or not if not symmetric then it is positively skewed or negatively skewed. Now a simple question comes, what happens to, you're here mean median and mode in these different types of distributions, when that transmission is symmetric or positively or say negatively skewed. So, now I try to give you a brief information,

Refer slide time :(18:49)

Coefficient of Skewness



when we have a positively skewed distribution, in this case γ_1 will be equal to 0. Now if you recall what will be the hair mode, mode will be somewhere here, which is the maximum frequency value. So, corresponding to maximum frequency this will give me the value of your mode, median is the value which will try to divide the entire area under this curve into two parts. So, this will be somewhere here and the mean, will be somewhere here, yeah. So, in the case of positively skewed distribution, mode will have the highest value followed, by median and mean. So, mode will be smaller than median and median will be smaller than mean, the opposite will happen, when we have the negatively skewed distribution, in this case, the mode will be somewhere here, corresponding to this frequency. So, mode will be somewhere here, corresponding to X and similarly the median, will be corresponding, to this frequency and median will be somewhere here and the mean will be somewhere here. So, in this case, when we have with a negatively skewed distribution, then mode will be greater than median and median will be greater than mean. Well in the case of symmetric, distribution all the values of mean median and mode they are going to be the same see here and in this case γ_1 is equal to 0 and median mean median mode will be somewhere here.

Refer slide time :(20:19)

Coefficient of Skewness

Another coefficient of skewness is defined as

$$S_{sk} = \frac{\bar{x} - \bar{x}_{mode}}{\sigma_x}; -3 \leq S_{sk} \leq 3$$

$$S_{sk} = \frac{3(\bar{x} - \bar{x}_{median})}{\sigma_x}; -3 \leq S_{sk} \leq 3$$

$$\bar{x} - \bar{x}_{mode} = 3(\bar{x} - \bar{x}_{median})$$

- $S_{sk} > 0$ for positively skewed distribution
- $S_{sk} < 0$ for negatively skewed distribution
- $S_{sk} = 0$ for symmetric distribution

There are some more, coefficients of skewness, which have been given in the literature so, I will just briefly give you an idea. So, one measure of coefficient of skewness or one coefficient of skewness is based on, the mean and mode, which is given by mean - mode divided by standard deviation. So, Sigma X is giving the value of standard deviation. So, this is essentially the value, of say s what we have used in the earlier lecture. But I'm using here is say Sigma X, to denote it that this is standard deviation, because there's a standard notation in many, many books. And this is the same as this quantity, which is based on the mean and median because you may recall that we have a relationship that $\bar{x} - \text{mode}$, is approximately equal to 3 times $\bar{x} - \text{median}$ under certain conditions. So, these two measures lie between minus 3 and plus 3 and we say that if these coefficients are greater than 0 that means the curve is positively skewed, if they are negative that means the curve is negatively skewed. And if this coefficient is 0 that means the curve is symmetric.

Refer slide time :(21:40)

Coefficient of Skewness

Other variants of coefficient of skewness based on quartiles (Q_1, Q_2, Q_3, Q_4) and percentiles (P_{10}, P_{50}, P_{90}) are

$$S_{qsk} = \frac{(Q_3 - Q_2) - (Q_2 - Q_1)}{(Q_3 - Q_2) + (Q_2 - Q_1)}; \quad -1 \leq S_{qsk} \leq 1$$

$$S_{psk} = \frac{(P_{90} - P_{50}) - (P_{50} - P_{10})}{(P_{90} - P_{50}) + (P_{50} - P_{10})}; \quad -1 \leq S_{psk} \leq 1$$

Same interpretation

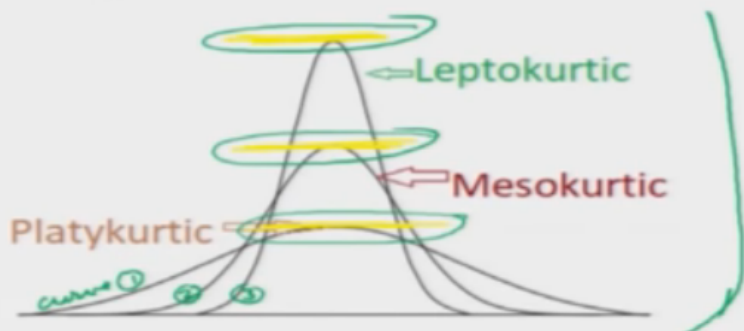
And similarly, we have two more measures, which are based on the definitions of quartiles and percentile. So, the coefficient of skewness based on quartiles, is given by like this q_3 , minus Q_2 , minus Q_2 , minus Q_1 and divided by q_3 minus Q_2 , plus Q_2 , minus Q_1 . And similarly the coefficient based on percentile, is given by this formula and both this coefficient, they lie between minus 1 and plus 1 and they have the same interpretation, as earlier means, positive value of this coefficient will indicate positively skewed, code negative value will indicate the negatively skewed, code and 0 value will indicate the symmetric curve.

Refer slide time :(22:22)

Kurtosis

Observe the following curve. The three curves are representing three frequency distributions.

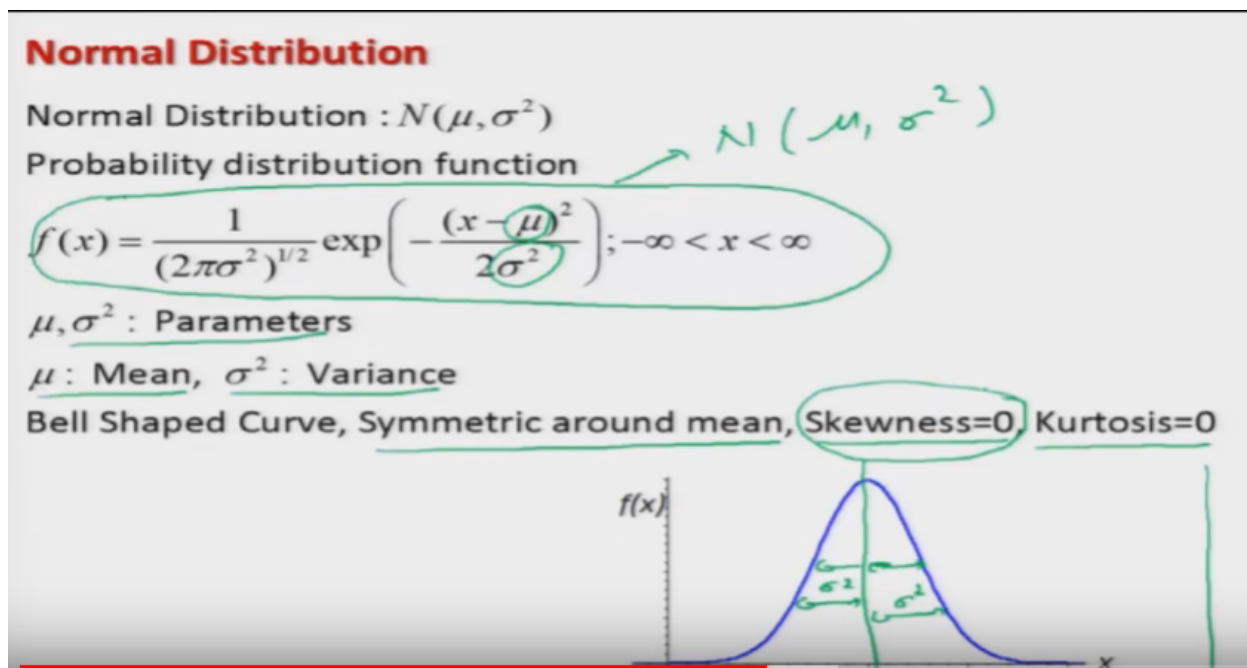
Peak of the curve



Kurtosis describes the peakedness or flatness of a frequency

Now after this I come on the next aspect, which is called, 'Kurtosis', please try to observe here, in this picture, I have made here three curves, duck car in the sail let me call it here's curve, one two and here three, I request used to you please try to observe, the hump of the curve, where is the hump of the curve, which is here I'm making my yellow color, this and by looking at these three curves, can we really say that what is the feature related to the peak of the curve? The peak of the curve 3, this is the highest, followed by the peak of the curve number 2 and followed by the peak of car number 1. So, the question is, is how to show this property of Peaks and how to quantify it? So, this property of kurtosis, this describes the wickedness or flatness of a frequency curve, flatness means, how flat is the curve at the peak. Now after this you can see here, from this curve that one of the curve has more peak and other curves have are, less Peaks. But how to compare it, what is more and what is less? So, what we try to do here, we try to measure the peakedness, with respect to the peakedness of normal distribution, what is the normal distribution? In statistics we have a probability density function, what will what we call as normal distribution or sometime it is called as, 'Gaussian Distribution'. So, before I try to give you an idea about, this peakedness let me try to give you the idea of normal distribution.

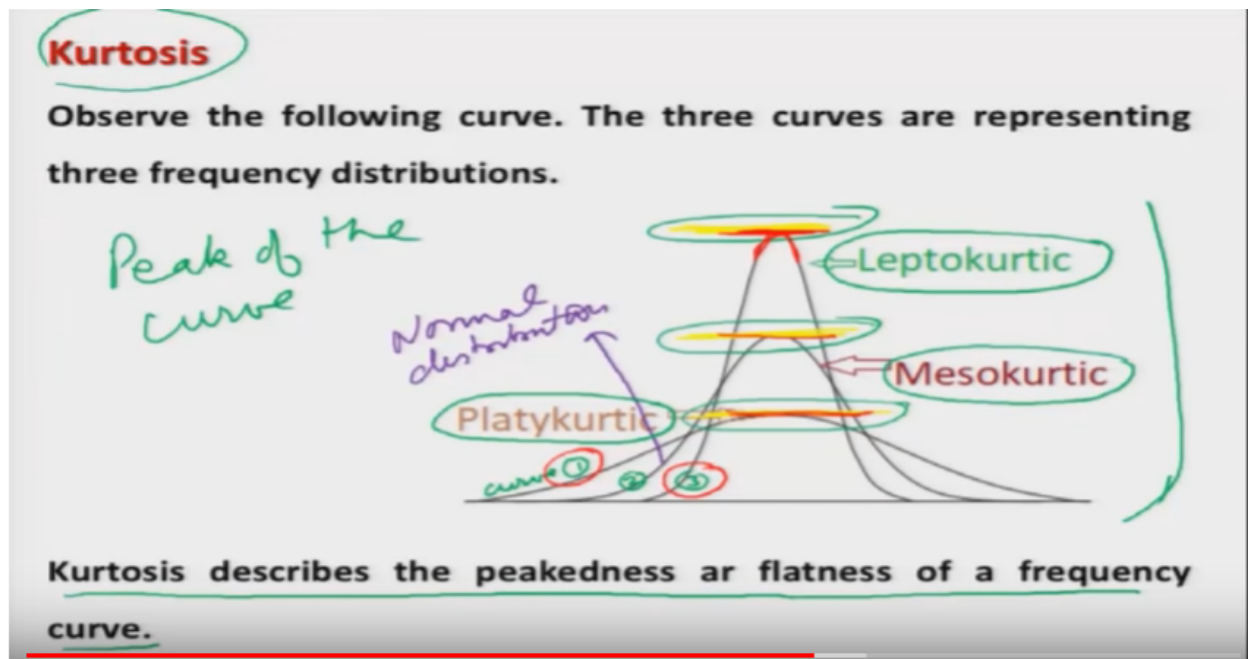
Refer slide time :(24:38)



So, the probability density function, of our normal distribution function, is given by like this. And this property density function is controlled by two parameters, mu and Sigma square. So, we denote this function, as say here n, which P is normal and the two parameters mu and Sigma square, which are the parameters of this, density function, here this mu is indicating, the mean and Sigma square is indicating the variance. So, if I try to draw this curve, this will look here like this. So, this value is indicating here, the mean and this is spread here, I don't the mean this is giving us the value of Sigma Square. And this

curve is actually, symmetric around mean. So, we try to compute, the coefficient of skewness and kurtosis in the case of normal distribution. And in this case the coefficient of skewness comes out to be zero and coefficient of kurtosis comes out to be zero. And this was the reason that we were trying to, conclude on the basis of coefficient of skewness, being zero positive or negative. And similarly we are going to do, with the case of kurtosis. So, now the curve of the normal distribution will have zero kurtosis. So, now we can compare the peaks, of other curves, with respect to the normal curve. And this is what we are doing in this picture,

Refer slide time :(26:21)



here if you try to see, the curve in the mid, curve number here tow this is the curve of normal distribution. And we are trying to compare the flatness or peakedness, with respect to the curve number two. So, now you can see here, first into the curve number three, this is here, which is here, the curve number three has got, more peak, then that curve number two. And similarly if you try to look in the curve number one, then curve number one has got a smaller peak, than the curve number two. So, what we try to do here? That all those curves which have got higher Peaks than the normal curve, they are called as, 'Leptokurtic', the peakedness of the normal distribution, is called as, 'Mesokurtic'. And the peakedness of those curves, which have got, lower peak than that of normal curve, this is called, 'Platykurtic'. And this is how we try to; characterize the less or more peakedness, with respect to the wickedness of the, normal distribution or the majority curve.

Refer slide time :(27:46)

Kurtosis

Shape of the hump (middle part of the curve or frequency distribution) of the normal distribution has been accepted as a standard.

Kurtosis examines the hump or flatness of the given frequency curve or distribution with respect to the hump or flatness of the normal distribution.

So, now I can explain you that the shape of the hump, which is the middle part of the curve or the frequency distribution, of the normal distribution, has been accepted as a standard one. And this kurtosis the property of kurtosis examines the humpable or the flatness of the given frequency curve or distribution, with respect to the hump or flatness of the normal distribution, which we have just understood.

Refer slide time :(28:15)

Kurtosis

Curves with hump like of normal distribution curve are called mesokurtic.

Curves with greater peakedness (or less flatness) than of normal distribution curve are called leptokurtic.

Curves with less peakedness (or grater flatness) than of normal distribution curve are called platykurtic.

And those curves, which have got the hump, like a normal distribution they are called, 'Mesokurtic', with greater peakedness or say less flatness, then that of normal distribution curve, they are called as, 'Leptokurtic', curves and those curves, which have got less peakedness or say greater flatness, then that of normal distribution, they are called as, 'Platykurtic', curves.

Refer slide time :(28:46)

Coefficient of Kurtosis

Karl Pearson's coefficient of kurtosis $\beta_2 = 3$ Normal dist.

$$\beta_2 = \frac{\mu_4}{\mu_2^2} \rightarrow f(\mu_2, \mu_4)$$

where μ_2 and μ_4 are the second and fourth central moments respectively.

$$\gamma_2 = \beta_2 - 3 \rightarrow R \text{ s/w}$$

γ_2 : mag.
 $\begin{matrix} > 0 \\ < 0 \end{matrix} \left. \vphantom{\begin{matrix} > 0 \\ < 0 \end{matrix}} \right\} \text{nature of hump}$

Now the Pearson's is how to quantify it. So, we have a coefficient of kurtosis and there are different types of coefficient of kurtosis, but here we are first going to consider, the Karl Pearson's coefficient of kurtosis, which is denoted by beta - that is the standard notation. And similar to the coefficient of skewness you can see here that this beta 2 is also depending on the mu 2 and mu 4, what are this mu 2 and mu 4? Mu 2 is the second central movement and mu 4 is the 4th central movement. And the coefficient of kurtosis is defined as the 4th central movement, divided by the square of second central moment, the value of beta 2 for a normal distribution; this comes out to be 3. So, what we try to do? That we try to define another measure, which is be tied to minus 3 and we denoted by here gamma 2. Now the advantage of gamma 2 is that that just by looking at the value of gamma 2, we will get the idea of the magnitude and if this is greater than 0, is smaller than 0 or equal to 0, will give us the idea about the, nature of hump. So, that is why we have two coefficient of kurtosis and in R software, this gamma 2 is produce in the outcome.

Refer slide time :(31:27)

So, now you can see here the same thing. So, I can see here for a normal distribution beta 2 is equal to 3 and gamma 2 equal to 0 and if beta 2 is greater than 3 or gamma 2 is greater than 0, then we say that the

curve is leptokurtic, if beta 2 is equal to 3 and or equivalently the gamma 2 is equal to 0, then we say the frequency distribution or the frequency curve is Mesokurtic. And if beta 2 is smaller than 3 and gamma 2 is smaller than 0, then we say that the distribution is platykurtic. So, you can see here that in the same figure that we had drawn, for the curves leptokurtic, we have this 4 major critique we have this and for platykurtic, we have this. So, this is about the coefficient of kurtosis.

Refer Slide Time :(31: 19)

Coefficient of Kurtosis

Few properties

- $\beta_2 \geq 1$
- $\beta_2 > \beta_1$
- $\beta_2 \geq \beta_1 + 1$

Some properties, which I'm not going to prove here, it is just trying to say that beta 2, the coefficient of kurtosis will always, be greater than or equal to one and coefficient of kurtosis beta 2, will always be greater than the coefficient of skewness, beta 1 and this beta 2 will always, be greater than or equal to beta 1 plus 1, these are some properties just for your information, I'm not going to use it here.

Refer Slide Time :(31: 46)

Coefficient of Kurtosis

Sample based coefficients of kurtosis are

$$\beta_{2s} = \frac{\frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^4}{\left(\frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2 \right)^2}$$

μ_2
 μ_4 } sample based

$$\gamma_{2s} = \beta_{2s} - 3$$

For leptokurtic distribution, $\beta_{2s} > 3$, $\gamma_{2s} > 0$

For mesokurtic distribution, $\beta_{2s} = 3$, $\gamma_{2s} = 0$

For platykurtic distribution, $\beta_{2s} < 3$, $\gamma_{2s} < 0$

And but, I would try to define the sample based coefficient of kurtosis, because there in the beta 2 and gamma 2, I simply use the population values. So now, I know, I don't have to do anything, I need here two central moments, mu 2 and mu 4, I will try to compute the sample based movements: that mean the value of mu 2 and mu 4, on the bases or given set of data and I will simply replace, them in the coefficient of kurtosis and I am denoting this coefficient of kurtosis, as beta 2's, right, as means sample and similarly, the coefficient of kurtosis, gamma 2 is now transformed to gamma 2 s and this is the same thing beta 2, s minus 3 and they have the same, interpretation as we have the case, in the beta 2 and gamma 2, for leptokurtic distribution, beta 2 s and gamma will be greater than 3 and gamma 2, s will be greater than 0, similarly beta 2 s will be 3 or gamma 2 s will be 0 in case of Mesokurtic and beta 2 s will be smaller than 3 or gamma 2 s will be smaller than 0, in case of platykurtic distribution. Right?

Refer Slide Time :(32: 57)

Skewness and Kurtosis

R Commands:

First we need to install a package 'moments'

```
> install.packages("moments")
```

```
> library(moments)
```

Sample based coefficient of skewness

```
skewness(x, na.rm = FALSE)
```

data vector

$$\gamma_{1s} = \frac{\frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^3}{\left(\frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2 \right)^{3/2}}$$

Sample based coefficient of kurtosis

```
kurtosis(x, na.rm = FALSE)
```

$$\gamma_{2s} = \frac{\frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^4}{\left(\frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2 \right)^2} - 3$$

x Numeric vector, matrix or data frame.

na.rm logical TRUE if missing values need to be removed

So now, we have a fair idea: that how to measure two types of characteristics, in a frequency curve, what is the symmetry and other is the peakedness and they are going to be quantified, by coefficients of skewness and kurtosis. So now, the next question is how to, compute them, in the R software. So, as we have seen: that in the case of, computing the moments, we need a package movements, so obviously, in order to compute the coefficients of kurtosis and coefficient of skewness, we need the information on the moments, mu 2, mu 3, and mu 4. So, we first need to install the package, movements and then based on that, we can compute the coefficient of skewness and coefficient of kurtosis. So, we try to understand it here. So, in order to compute the coefficient of skewness and kurtosis in R, first you need to install the package, movements by this command, install dot packages, inside the arguments, you have to write moments and then you have to load, this package movements. Right? And now, the sample-based coefficient of kurtosis, which was our gamma 1 s, this will be computed by the expression, a skewness, SKEWNESS and the data vector here X and yeah! If you want to use, the na dot RM, which means for the missing value, you are saying because here false: that means there are no, missing values and if there are missing value, we will simply write na dot R M is equal to true and similarly the sample based, coefficient of kurtosis, which we have defined as gamma 2 s, this will be computed by the command kurtosis, KURTOSIS all in small letters and inside the argument, same thing, like the data vector, X and if you do not have any missing value, then use na dot R M is equal to false and if there are missing values, then use na dot R M is equal to true. Okay.

Refer Slide Time :(35: 05)

Skewness and Kurtosis

R Commands:

When data is missing and data vector is `x.na`

Sample based coefficient of skewness

```
skewness(xna, na.rm = TRUE)
```

Sample based coefficient of kurtosis

```
kurtosis(x.na, na.rm = TRUE)
```

`x.na` Numeric vector, matrix or data frame containing NA values.

`na.rm` logical TRUE if missing values need to be removed

So, you can see here, this is not a difficult thing. So, yeah! If you have missing value and if you try to store those missing values inside the data vector X, NA then the command will become, skewness XN a and NA dot R M is equal to true, for computing the coefficient of skewness and the coefficient of kurtosis, in this case will be given by kurtosis, X dot na and na dot R M is equal to true. Right.

Refer Slide Time :(35: 31)

Skewness and Kurtosis

Example:

Following are the time taken (in seconds) by 20 participants in a race: 32, 35, 45, 83, 74, 55, 68, 38, 35, 55, 66, 65, 42, 68, 72, 84, 67, 36, 42, 58.

```
> time = c(32, 35, 45, 83, 74, 55, 68, 38, 35, 55, 66, 65, 42, 68, 72, 84, 67, 36, 42, 58)
```

```
> skewness(time)
```

```
[1] 0.05759762
```

> 0 Positively skewed

```
> kurtosis(time)
```

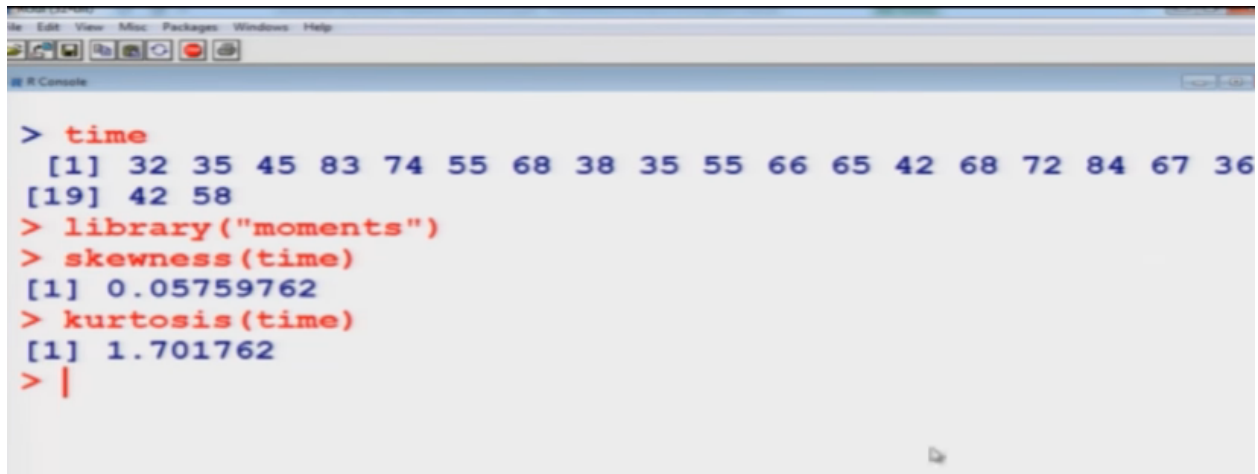
```
[1] 1.701762
```

> 0 leptokurtic
kurtosis > kurtosis(normal)

Now, I would try to take an example and show you that, how to measure this is skewness and kurtosis, in the data. So, I am going to use the same example, in which we had collected the timings of twenty

participants in a race and this data has been stored in Sider, variable check time. Now, after this, we simply have to use the command SKEWNESS, skewness and inside the argument give the, data vector and this is giving us the value, zero point, zero five seven five nine seven six two. So, this is indicating: that the skewness is, greater than zero. So, the frequency curve, in this case is, positively skewed and similarly you can see here, when you try to operate the kurtosis command, on the time vector, then it is giving to the value, one point seven and which is greater than zero. So, this is indicating that, the curve is leptokurtic, leptokurtic means, the hum, of this curve, is greater than the, hump of normal distribution. Now, I will try to assure you it on R software,

Refer Slide Time :(36: 56)



```
> time
 [1] 32 35 45 83 74 55 68 38 35 55 66 65 42 68 72 84 67 36
 [19] 42 58
> library("moments")
> skewness(time)
 [1] 0.05759762
> kurtosis(time)
 [1] 1.701762
> |
```

so you can see here, I have here that data on time, first I need to load the package, library, moments and I already, have installed this package on my computer. So now, I will need, the skewness of time, this comes out to be like this and kurtosis, of time. Right. You can see here, this is the same thing which you have just obtained and this is the screenshot of the same operation: that I just did.

Refer Slide Time :(37: 32)

Skewness and Kurtosis

Example: Handling missing values

Suppose two data points are missing in the earlier example where the time taken (in seconds) by 20 participants in a race. They are recorded as NA

```
NA, NA 45, 83, 74, 55, 68, 38, 35, 55, 66, 65, 42, 68, 72, 84, 67, 36,  
42, 58.
```

```
> time.na = c(NA, NA, 45, 83, 74, 55, 68, 38,  
35, 55, 66, 65, 42, 68, 72, 84, 67, 36, 42, 58)
```

```
> skewness(time.na, na.rm = TRUE)
```

```
[1] -0.0614137 < 0 negatively skewed
```

```
> kurtosis(time.na, na.rm = TRUE)
```

```
[1] 1.810021 > 0 leptokurtic
```

I know, I will take one more example, where I will show you that, how to compute the coefficient of skewness and kurtosis, when we have some data, missing. So, I will take the same example, in with the I have just removed to first to observation and I have replace it by na, na that means the data is missing and this data is stored in the data vector time dot, na and now, I will use the command skewness, on time dot na with the option any dot R M is equal to true: that means, please remove the,na value and then you try to compute the skewness. And similarly for the kurtosis, I will use the same command kurtosis, on the time dot na with then option n a dot R M is equal to true and this will give me the value of coefficient of kurtosis, when the two values are missing. So now, you can see here, the coefficient of skewness comes out to be negative, this is less than zero. So that means, the frequency curve based on the, the remaining observation, is now negatively, skewed. So, it this is indicating that, when the first two observations are deleted, then the nature of the skewness has changed, similarly for the kurtosis, this value is 1 point 8 1, which is greater than 0, this is again showing that, the curve is leptokurtic. So, this is indicating that, even after removing the first two observation, the nature of the curve, remains the same. And now, I will try to show you, this on the R software also,

Refer Slide Time :(39: 21)

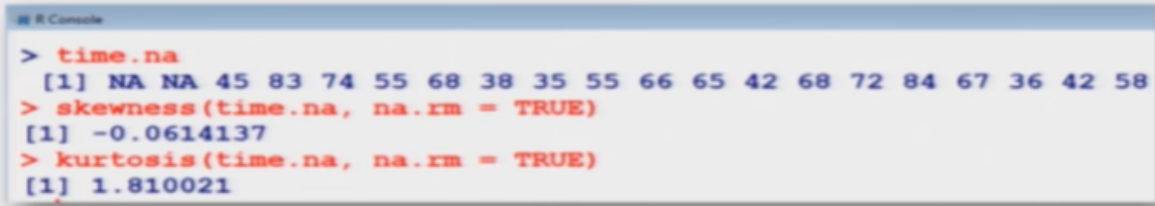
```
File Edit View Misc Packages Windows Help
R Console
> time
[1] 32 35 45 83 74 55 68 38 35 55 66 65 42 68 72 84 67 36
[19] 42 58
> library("moments")
> skewness(time)
[1] 0.05759762
> kurtosis(time)
[1] 1.701762
> time.na
[1] NA NA 45 83 74 55 68 38 35 55 66 65 42 68 72 84 67 36
[19] 42 58
> skewness(time.na, na.rm = TRUE)
[1] -0.0614137
> kurtosis(time.na, na.rm = TRUE)
[1] 1.810021
> |
```

so I already have stored this data time dot na on my computer. This is her like this, now I want to compute the skewness; this is coming out to be the same, what we have just observed? And if I try want to come compute the coefficient of kurtosis, this is here like this.

Refer Slide Time :(39: 39)

Skewness and Kurtosis

Example: Handling missing values



```
R Console
> time.na
[1] NA NA 45 83 74 55 68 38 35 55 66 65 42 68 72 84 67 36 42 58
> skewness(time.na, na.rm = TRUE)
[1] -0.0614137
> kurtosis(time.na, na.rm = TRUE)
[1] 1.810021
```

And then, the next slide is the screenshot of the same operation that we have done. Now, I would stop in this lecture and you can see that, we have discussed the coefficient of skewness and kurtosis, which are going to give you, two more pieces of information about your frequency curve, beside central tendency and variation. Now, you know how to find the behavior of the frequency curve, with respect to central tendency, variation, lack of symmetry and peakedness. And now you can see, just by looking the data, you were not getting all these things, but now, you know: that how to quantify these things and how the graphics in this case will look like, is if you try to plot the, frequency curves of the data, which I have

taken say time or say time dot na and try to see, whether the feature, of the curve is matching with the information given by the coefficient of skewness and kurtosis or not? And you will see that, it is matching. So, this is the advantage of these tools of descriptive statistics: that instead of looking at the huge, data sets, you simply try to, look into these values graphically, as well as quantitative way. And they will give you a very reliable information, but, you should know, how to use this information and how to interpret that data. So now, up to now, from the beginning I have used the tools, when we have data, only on one variable. So, we have discussed the univariate tools, of descriptive statistics. Now, from the next lecture, I will take up the case when we have more than one variable and P and in particular I will consider two variables. So, when we have the data on two variables, they also have some hidden, properties and characteristics, so how to quantify them? And how to have the information graphically? These are the topics which I will be taking from the next lecture. So, you practice these tools, try to understand it and enjoy it. And I will see you in the next lecture. Till then. Good bye.