1. Introduction to Skewness

Welcome to the series of E-learning modules on Skewness & Kurtosis. In this module, we are going to cover the definition, calculation, types and measures of Skewness & Kurtosis.

By the end of this session, you will be able to:

- Explain the concept of Skewness & Kurtosis
- Explain the types of Skewness & Kurtosis
- Explain the measures of Skewness & Kurtosis

We have discussed the frequency distribution in detail. It may be recapped here that the frequency distribution differs in three ways, namely - average value, variability or dispersion and shape.
In our previous modules, we have studied that the measures of central tendency which helps in understanding the average value and the measures of dispersion, which gives a view about the variability, or dispersion of the frequency distribution.

In our discussion today, let us focus on the shape of the frequency distribution and the averages. The two comparable characteristics of the shape of a frequency distribution are skewness and kurtosis, which help in revealing the entire story of a distribution.

Figure 1

Look at the diagram, we have represented two set of data in the histogram form and have calculated the mean and the standard deviation. We have found that both the data are having the same mean 15 and the same standard deviation 5 but they differ widely on their overall appearance as seen. Figure (a) is a diagram that looks alike and is called a symmetrical distribution whereas figure (b) is having an asymmetrical distribution and is said to have a skewed distribution.

Let us now understand the meaning of a skew.

Skew is the indicator of lack of symmetry in a distribution. When the mean, median and mode of the distribution do not have the same value in a distribution, we call it as a skewed distribution.

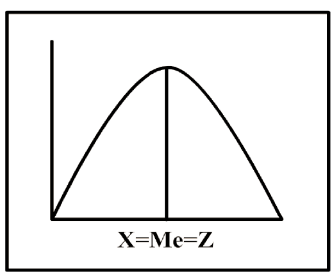
“When a series is not symmetrical it is said to be asymmetrical or skewed”- Croxton & Cowden
“Skewness refers to the asymmetry or lack of symmetry in the shape of a frequency distribution” – Morris Hamburg
“A distribution is said to be skewed when the mean and the median fall at different points in the distribution and the balance is shifted to one side or the other- left or right.”- Garett

Let us discuss the types of Skewness.
Here, let us recall that the mean is a center of gravity or balance point of a data, the median is a value that divides the data into equal areas and the mode is the value indicating the largest frequency that is the maximum ordinate of the distribution.

While studying the shape of the distribution we need to know that the direction of the skewness which is determined by ascertaining whether the mean is equal, greater or lesser that the mode. Based on which we have two types of distribution. They are Symmetrical distribution and asymmetrical distribution. The asymmetrical distribution is further classified into positively skewed and negatively skewed distribution.

When the values of the mean, median and mode coincide or are equal in a distribution then, it clearly indicates that the spread of the frequencies are same on both the sides of the center point of the curve and is called as a symmetrical distribution.

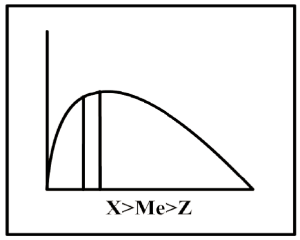
Figure 2



A distribution in which the mean, median and mode are different is called an asymmetrical distribution. Based on the values of the mean, median and mode the asymmetrical distribution is classified as positively skewed distribution and negatively skewed distribution.

A distribution is said to be positively skewed when the value of the mean is greater than the median and the mode and the value of the mode is least. The median lies in between the mean and the mode.

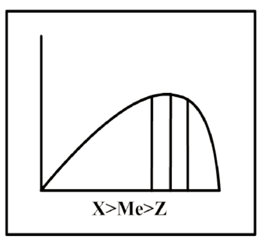
Figure 3



In such a case, the frequency distribution is elongated to the right that is, having a longer tail to the right of the center of gravity. This indicates that the frequencies are spread out over a greater range of values on the high value end of the curve (right hand side) than they are on the low value end.

A distribution is said to be negatively skewed when the value of the mean is least and that of the mode is greatest. The median lies in between the mean and the mode.

Figure 4



In such a case, the frequency distribution is elongated to the left that is having a longer tail to the left of the center of gravity. This indicates that the frequencies are spread over the lower range of values on the low value end of the curve (left hand side) than they are on the high value end.
While studying a frequency distribution to ascertain whether the distribution is skewed or not the following test may be applied.
If a distribution satisfies the following conditions then, it indicates that the distribution has skewness present in it as seen in an Asymmetrical distribution

- The value of mean, median and mode do not coincide
- The sum of the positive deviation from the median is not equal to the sum of the negative deviation from the median
- Quartiles are not equidistant from the median
- Frequencies are not equally distributed at points of equal deviation from the mode
- Finally, when the data is plotted on the graph they do not give the normal bell shaped form

2. Measures of Skewness

Now, let us discuss the Measures of skewness.
The measure of skewness is the indication about the direction of the distribution, extent of the asymmetry in a distribution and allowing the comparison of two or more data series. There are two kinds of measures of skewness, namely Absolute measure and relative measures.

An absolute measure of skewness is calculated on the basis of the absolute terms: mean, median, mode and quartiles.
When Skewness is calculated using mean and mode then skewness is the difference between the mean, and mode it is symbolically represented as absolute skewness is equal to mean minus mode.

When skewness is calculated from the lower quartile, upper quartile and median then, it is given by the formula:
absolute skewness is equal to the upper quartile plus the lower quartile minus twice the median.

Advantage of using the absolute measure is that the difference between mean, median and mode is used to measure the skewness, which is alike in a symmetrical distribution. In addition, the mean moves away from the mode in an asymmetrical distribution indicating whether the skewness is positive or negative.

The disadvantage of absolute measure is that, it would be expressed in the unit of value of the distribution. In addition, the comparison between series of different units is not possible as distributions vary greatly and the difference between the mean and the mode are in absolute terms. This could be considerable in some series and small in other series though the frequency curve of both the distribution is similarly skewed.

Let us discuss the relative measure of skewness.
As a description of one distribution alone, its interpretation of measure of skewness as slight, marked or moderate skewness is necessarily vague. When the absolute measures are expressed in relation to some measures of the respective distribution, the measure would be relative and can be used for comparison directly.

There are four measures of relative skewness, which are more frequently used and are known as the coefficient of skewness. The four measures are:

1. Karl Pearson's Measure
2. Bowley's Measure
3. Kellys's Measure
4. Moment's Measure

Before discussing the measures in detail, let us understand what the property of a good measure of skewness is.

1. It is a pure number, that its value should be independent of the units of the series and of the degree of variation in the series.
2. If it has a zero value, then the distribution is symmetrical.

3. It should have some meaningful scale of measure, so that we can easily interpret the measured value.

Let us now discuss the various measures of skewness.
Karl Pearson's coefficient of skewness is also known as Pearsonian coefficient of skewness denoted as Sk_p. It is based upon the difference of the mean and the mode of the distribution divided by the standard deviation of the distribution to give a relative measure.

The formula for this calculation thus becomes Sk_p that is Karl Pearson's coefficient of skewness is equal to mean minus mode divided by standard deviation. The value given should lie between ± 1. This equation gives both the direction as well as the extent of skewness.

In a situation where the mode becomes ill defined or indeterminate, we use the empirical formula Mode = 3Median – 2 mean to calculate the mode substituting this in the previous formula.

Skp= mean –mode/standard deviation
= mean –(3 Median -2 mean)/standard deviation
= 3(median –mean)/standard deviation

Now, this formula is equal to the previous formula.

Thus,
3(Mean – Median)/Standard deviation = mean –mode/standard deviation
Theoretically, the value of this coefficient varies between ±3, but generally, the value does not exceed ±1.

3. Bowley's and Kelly's Coefficient of Skewness

Bowley's coefficient of skewness is based on the quartiles. The quartiles that we consider are the first quartile (Q_1), median (Q_2) and the third quartile (Q_3). It is denoted as Sk_B. In a symmetrical distribution the first and the third quartile are equidistant from the median.
In an Asymmetrical distribution, the third quartile is the same distance over the median as the first quartile is below it. So, we calculate the skewness as the upper quartile minus the median which is equal to median minus the lower quartile. So, by equating we get that the skewness is equal to upper quartile plus the lower quartile minus twice the median is the absolute measure.

A distribution is said to be positively skewed when the upper quartile is farther from the median than the lower quartile is from the median. A distribution is said to be negatively skewed when the lower quartile is farther from the median than the upper quartile.

The formula for this calculation thus becomes Sk_B that is Bowley's coefficient of skewness is equal to upper quartile plus the lower quartile minus twice the median divided by upper quartile minus lower quartile. The value given should lie between +/- 1. In case of open ended series as well as where extreme values are found in the series, this measure is particularly useful.

Let us discuss the Kelly's Coefficient of skewness.
Generally, when we are considering the data we would like to cover the entire data in measuring the skewness and more interested in the extreme values. Kelly has developed a measure of skewness which is based on deciles and percentiles by taking any two values that are equidistant from the median. It is denoted as Sk_k.

The following formulae are used for measuring the skewness upon the 10^{th} and the 90^{th} percentiles or the 1^{st} and 9^{th} deciles.

$Sk_k = P_{10}+P_{90}-2Q_2/P_{90}-P_{10}$ for percentiles
$Sk_k = D_1+D_9-2Q_2/D_9-D_1$ for deciles

This measure of skewness is best used when we have to calculate skewness based on percentiles and deciles. However, this method is not popular in practice.

A measure of skewness is obtained by making use of the second and the third moments about the mean. When the method of moments is applied ß1 is used as relative measure of skewness.
ß1 is defined as beta one is equal to square of the third moment divided by the cube of the second moment.

In a symmetrical distribution ß1 shall be zero. The greater the value of ß1 the more the skewed the distribution. But this measure cannot tells us about the direction of the skewness i.e. whether it is positive or negative. To overcome this draw back we calculate the Karl Pearson's gamma one.

Gamma one is defined as the square root of beta one and the sign of the skewness will depend upon the value of mu three. If mu three is positive we will get positive skewness and if mu three is negative we will get negative skewness. It is advisable to use gamma as a measure of skewness.
Gamma one is equal to square root of beta one which is equal to mu three divided by mu two cube which is equal to mu three by standard deviation.

4. Introduction to Kurtosis

Now that we have understood the concept of Skewness, let us move further and discuss the concept of Kurtosis.

If a distribution is symmetric, the next question is about the central peak: is it high and sharp, or short and broad? You can get some idea of this from the histogram, but a numerical measure is more precise.

Kurtosis is a Greek word which means “bulginess”. In statistics we use kurtosis to help in identifying the degree of flatness or peakedness of frequency curve in the region about the mode.

Let us differentiate the flatness or peakedness of frequency curves with the help of these images. The first image of platypus shows the platy or flatness in kurtosis and the second image of kangaroos show the lepto or peakedness in kurtosis.

The degree of kurtosis of a distribution is measured relative to the peakedness of normal curve. This concept of kurtosis is rarely used in elementary statistical analysis.

Balanda and MacGillivray say the same thing in another way: Increasing kurtosis is associated with the “movement of probability mass from the shoulders of a distribution into its center and tails”.

In other words the measure of kurtosis tells us the extent to which a distribution is more peaked or flat topped than the normal curve. We can define “Kurtosis is the degree of peakedness of a distribution, usually relative to a normal distribution”.

The height and sharpness of the peak relative to the rest of the data are measured by a number called kurtosis. Higher values indicate a higher, sharper peak; lower values indicate a lower, less distinct peak.

There are three types of Peakedness of curve Leptokurtic, Platykurtic and Mesokurtic.

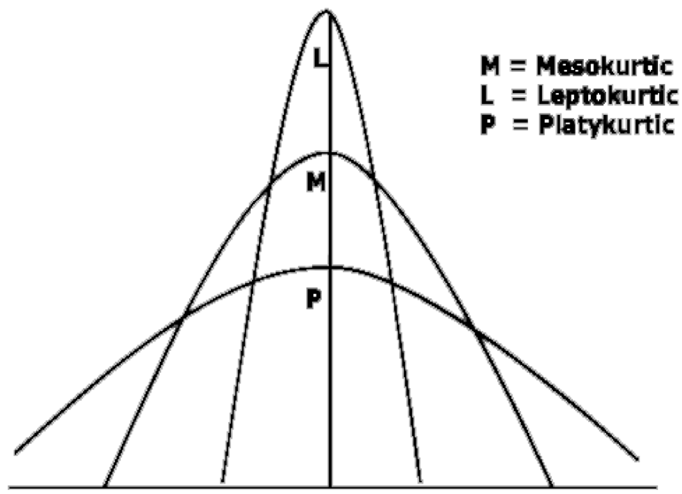
When a curve is more peaked than the normal curve, it is called a leptokurtic curve. In a leptokurtic curve the items are closely bunched around the mode.

When a curve is flat topped than the normal curve, it is called a platykurtic curve. A platykurtic curve indicates the flat toppedness or the kurtosis of excess.

When the curve is a normal curve we call it as mesokurtic curve.

The following diagram helps in understanding the shape of three different curves mentioned. It clearly indicates that these curves differ widely in their convexity.

Figure 5

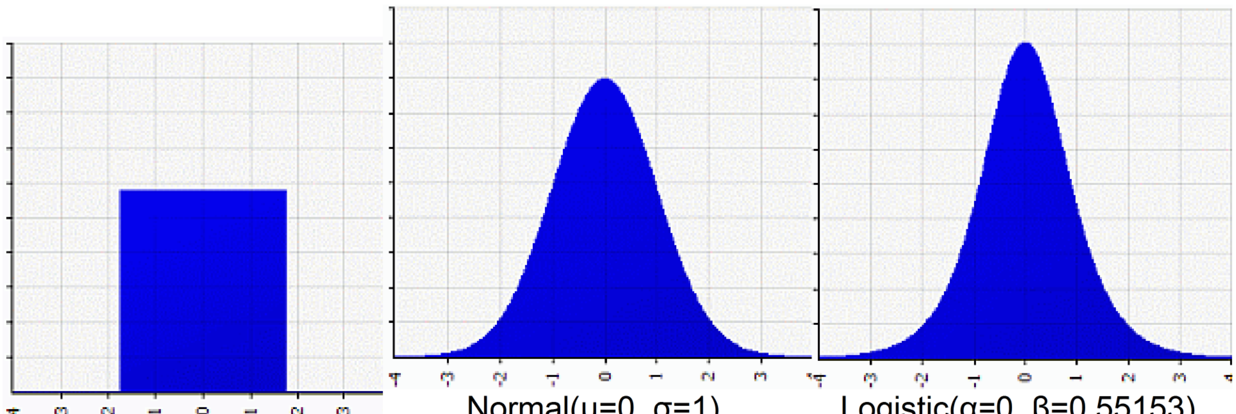


The curve M is a normal curve and is called mesokurtic.
The curve L is more peaked than M and is called the leptokurtic curve. The leptokurtic curve has a narrow central portion and a higher tail than the normal curve. The leptokurtic is compared to the kangaroos noted for leaping.

The curve P is less peaked (flat –topped) than curve M and is called platykurtic. The platykurtic curve has a broader central and lower tails. The platykurtic is compared to a squat of a platypus with shorter tails.

Visualizing the Kurtosis
Kurtosis is unfortunately harder to picture than skewness.
Let us look at the three pictures. All three of these distributions have mean of 0, standard deviation of 1, and skewness of 0, and all are plotted on the same horizontal and vertical scale. Look at the progression from left to right, as kurtosis increases.

Figure 6



Uniform(min=−√3,

Normal(μ=0, σ=1)
kurtosis = 3, excess = 0

Logistic(α=0, β=0.55153)
kurtosis = 4.2, excess = 1.2

max=√3)
kurtosis = 1.8, excess = −1.2

Moving from the illustrated uniform distribution to a normal distribution, you see that the “shoulders” have transferred some of their mass to the center and the tails. In other words, the intermediate values have become less likely and the central and extreme values have become more likely. The kurtosis increases while the standard deviation stays the same, because more of the variation is due to extreme values.

Moving from the normal distribution to the illustrated logistic distribution, the trend continues. There is even less in the shoulders and even more in the tails, and the central peak is higher and narrower.

5. Computing of Kurtosis and Skewness

Computing of Kurtosis:
You may remember that the mean and standard deviation have the same units as the original data, and the variance has the square of those units. However, the kurtosis has no units: it's a pure number, like a z-score.

The reference standard is a normal distribution, which has a kurtosis of 3. In token of this, the excess kurtosis is presented: excess kurtosis is simply kurtosis−3.
The moment coefficient of kurtosis of a data set is computed almost in the same way as the coefficient of skewness is computed. The most important measure of kurtosis is the value of the coefficient ß2. It is defined as beta two is equal to the fourth moment divided by the square of the second moment. The greater the value of beta two, the more peaked is the distribution.

- For a normal curve, the value of beta two is equal to three and is called the mesokurtic curve.
- When the value of beta two is greater than the normal curve then the curve is more peaked than the normal curve and is called the leptokurtic curve.
- When the value of beta two is less than three then the curve is less peaked from the normal curve and is called the platykurtic curve.

Sometimes we also derive the beta two through gamma two and are used as a measure of kurtosis. Gamma two is defined as gamma two is equal to beta two minus three. As, the value of gamma two will be zero for a normal curve. If gamma two is positive the curve is leptokurtic, and if gamma two is negative the curve is platykurtic.

Let us take an example to understand the calculation of the coefficient of skewness and kurtosis based on the following data:

Table 1

X	4.5	14.5	24.5	34.5	44.5	54.5	64.5	74.5	84.5	94.5
F	1	5	12	22	17	9	4	3	1	1

Solution:
As we are going to calculate the coefficient of skewness and kurtosis from the moments let us calculate the moments for the given data.

First step let us prepare the various measures needed for calculation.

1. In this table the first column indicates the variable.
2. The second column is the frequency of the data and its total is taken as 75.
3. The third column is the deviation of the variable from the mean divided by 10 the width

of the interval.
4. Column 4 indicates the product of the deviation with the frequency and the total is minus 30.
5. Column 5 indicates the product of the frequency with the square of the deviations having a total of 224.
6. In column 6 we calculate the product of the cube of the deviation with the frequency and have a total of minus 6.
7. Finally the last column is the product of the fourth power of the deviation and the frequency with a total of 2072.

Calculations of the moments from the arbitrary mean be: mu one dash is equal to sigma of frequency into deviation divided by N is equal to minus 30 divided by 75 is equal to minus 0.4. Similarly we calculate the mu two dash as 2.99, mu three dash as minus 0.08 and mu four dash as 27.63.

For the calculation of the central moments we need the second, third and fourth moment. Mu two is equal to mu two dash minus mu one dash square is equal to 2.99 minus of minus 0.4 square is equal to 2.99 minus 0.16 is equal to 2.83.

Similarly when we calculate the third and the fourth moment we get the third moment as 3.38 and the fourth moment is equal to 30.295.

Calculation of the skewness and the kurtosis
Skewness is equal to beta one is equal to minus mu three square divided by mu two cube is equal to 3.38 square divided by 2.83 cube is equal to 11.424 divided by 22.665 is equal to 0.504

Kurtosis is equal to beta two is equal to mu four divided by mu two square is equal to 30.295 divided by 2.83 square is equal to 30.295 divided by 8.01 is equal to 3.782

Skewness is 0.504 and hence the distribution is asymmetrical and the curve is leptokurtic as the kurtosis is above 3.

Let us take another example.

The first four central moments of distribution are 0, 2.5, 0.7 and 18.75. Comment on the skewnss and kurtosis of the distribution.

Solution:
We are given that mu one is zero, mu two is 2.5, mu three is 0.7 and mu four is 18.75.

Testing the skewness:
We know that the skewness is measured with the beta one coefficient.
Beta one is equal to mu three square divided by the mu two cube which is equal to 0.7 square divided by 2.5 cube therefore beta one is equal to 0.031.
Since beta one is equal to 0.031, the distribution is slightly skewed that is it is nit perfectly symmetrical.

Testing the Kurtosis:

For testing the kurtosis we use the beta two coefficients.
Beta two is equal to mu four divided by mu two square which is equal to 18.75 divided by 2.5 square is equal to 3. Since beta two is equal to 3 the distribution is mesokurtic.

Here's a summary of our learning in this session, where we have understood:

- The concept of Skewness & Kurtosis
- The calculation of Skewness & Kurtosis under various averages for a frequency distribution
- Measures and Application of Skewness & Kurtosis