What Is Skewness?

Skewness is a statistical measure that assesses the asymmetry of a probability distribution. It quantifies the extent to which the data is skewed or shifted to one side.

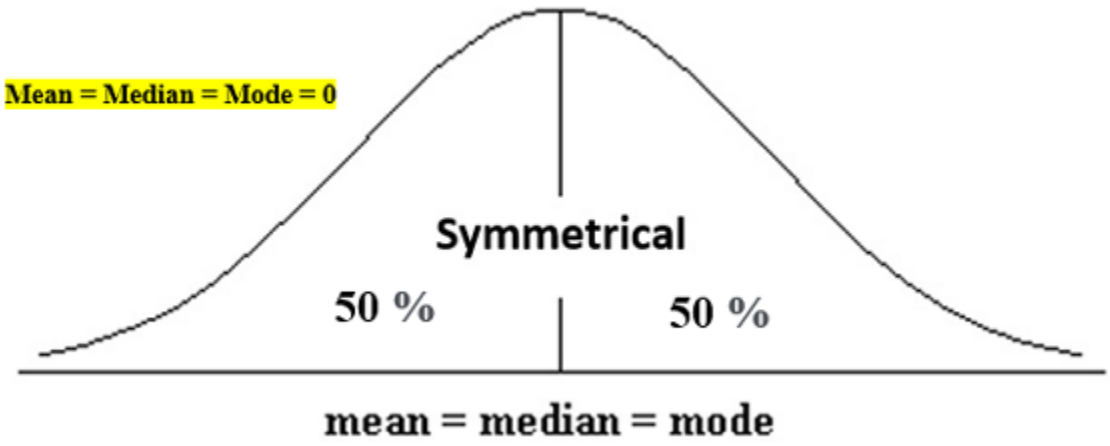
Positive skewness indicates a longer tail on the right side of the distribution, while negative skewness indicates a longer tail on the left side. Skewness helps in understanding the shape and outliers in a dataset.

Depending on the model, skewness in the values of a specific independent variable (feature) may violate model assumptions or diminish the interpretation of feature importance.

A probability distribution that deviates from the symmetrical normal distribution (bell curve) in a given set of data exhibits skewness, which is a measure of asymmetry in statistics.

A skewed data set, typical values fall between the first quartile (Q1) and the third quartile (Q3).

The normal distribution helps to know a skewness. When we talk about normal distribution, data symmetrically distributed. The symmetrical distribution has zero skewness as all measures of a central tendency lies in the middle.

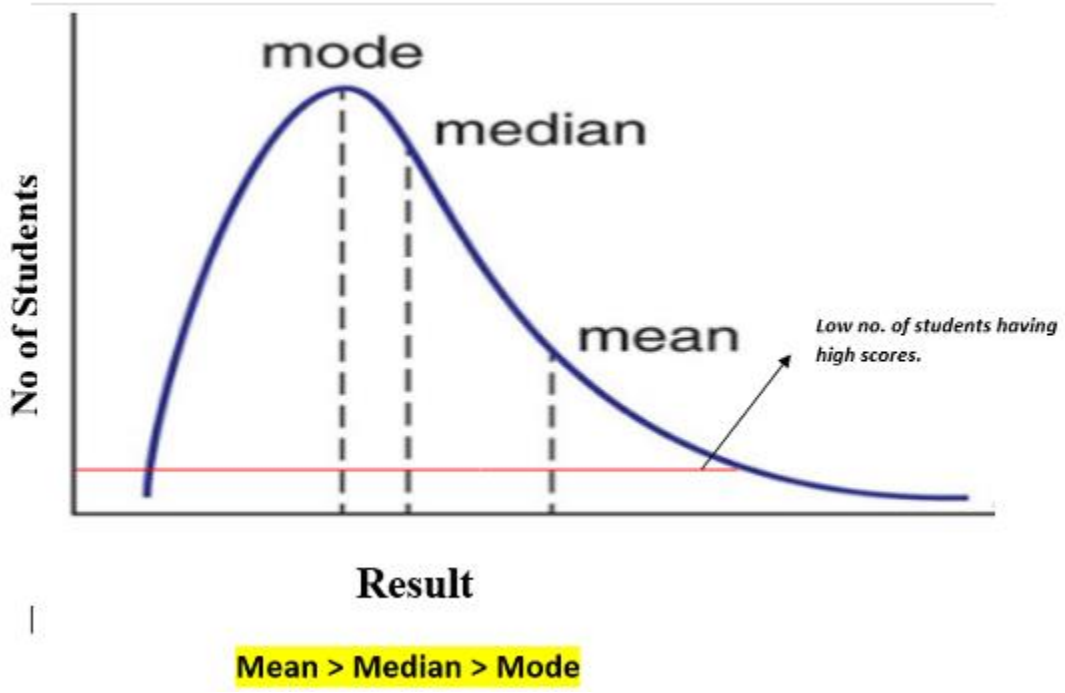


In a symmetrically distributed dataset, both the left-hand side and the right-hand side have an equal number of observations. (If the dataset has 90 values, then the left-hand side has 45 observations, and the right-hand side has 45 observations.). But, what if not symmetrical distributed? That data is called asymmetrical data, and that time skewness comes into the picture.

Types of Skewness

Positive Skewed or Right-Skewed (Positive Skewness)

In statistics, a positively skewed or right-skewed distribution has a long right tail. It is a sort of distribution where the measures are dispersing, unlike symmetrically distributed data where all measures of the central tendency (mean, median, and mode) equal each other. This makes Positively Skewed Distribution a type of distribution where the mean, median, and mode of the distribution are positive rather than negative or zero.



In positively skewed, the mean of the data is greater than the median (a large number of data-pushed on the right-hand side). In other words, the results are bent towards the lower side. The mean will be more than the median as the median is the middle value and mode is always the most frequent value.

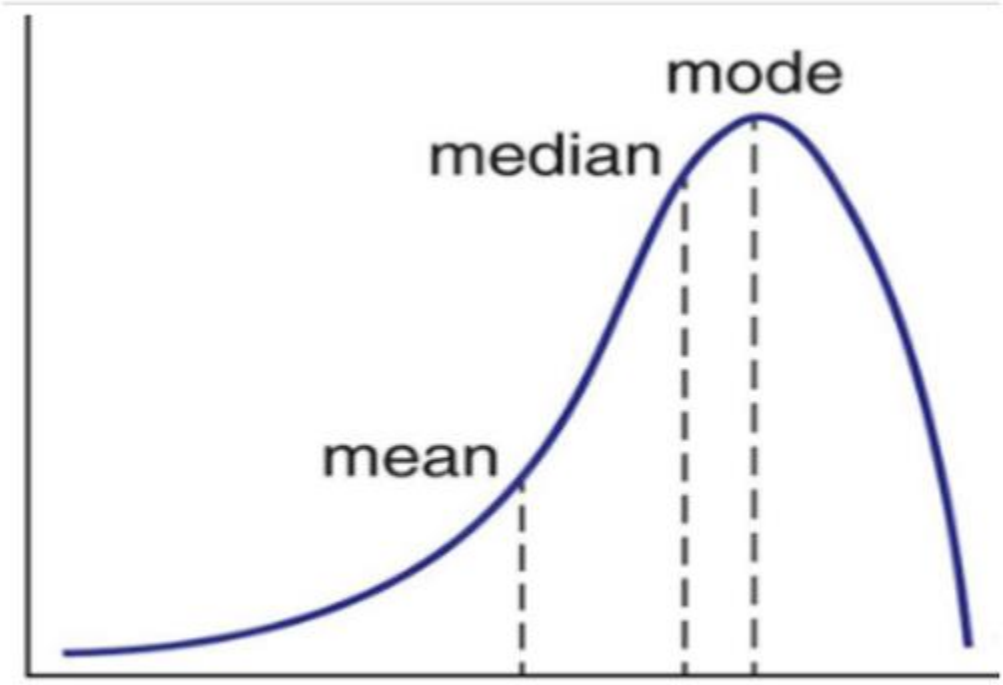
Extreme positive skewness is not desirable for a distribution, as a high level of skewness can cause misleading results. The data transformation tools are helping to make the skewed data closer to a normal distribution. For positively skewed distributions, the famous transformation is the log transformation. The log transformation proposes the calculations of the natural logarithm for each value in the dataset.

Negative Skewed or Left-Skewed (Negative Skewness)

A distribution with a long left tail, known as negatively skewed or left-skewed, stands in complete contrast to a positively skewed distribution. In statistics, negatively skewed distribution refers to

the distribution model where more values are plots on the right side of the graph, and the tail of the distribution is spreading on the left side.

In negatively skewed, the mean of the data is less than the median (a large number of data-pushed on the left-hand side). Negatively Skewed Distribution is a type of distribution where the mean, median, and mode of the distribution are negative rather than positive or zero.



Mode > Median > mean

Median is the middle value, and mode is the most frequent value. Due to an unbalanced distribution, the median will be higher than the mean.

How to Calculate the Skewness Coefficient?

Various methods can calculate skewness, with Pearson's coefficient being the most commonly used method.

Pearson's first coefficient of skewness
To calculate skewness values, subtract the mode from the mean, and then divide the difference by standard deviation.

$$\text{Pearson's first coefficient} = \frac{\text{Mean} - \text{Mode}}{\text{Standard Deviation}}$$

As Pearson's correlation coefficient differs from -1 (perfect negative linear relationship) to +1 (perfect positive linear relationship), including a value of 0 indicating no linear relationship, When we divide the covariance values by the standard deviation, it truly scales the value down to a limited range of -1 to +1. That accurately shows the range of the correlation values.

Pearson's first coefficient of skewness is helping if the data present high mode. However, if the data exhibits low mode or multiple modes, it is preferable not to use Pearson's first coefficient, and instead, Pearson's second coefficient may be superior, as it does not depend on the mode.

Pearson's second coefficient of skewness
subtract the median from the mean, multiply the difference by 3, and divide the product by the standard deviation.

$$\text{Pearson's second coefficient} = \frac{3\,(\text{Mean} - \text{Median})}{\text{Standard Deviation}}$$

$$\text{Mean} - \text{Mode} \approx 3\,(\text{Mean} - \text{Median})$$

Rule of thumb:

- For skewness values between -0.5 and 0.5, the data exhibit approximate symmetry.
- Skewness values within the range of -1 and -0.5 (negative skewed) or 0.5 and 1(positive skewed) indicate slightly skewed data distributions.
- Data with skewness values less than -1 (negative skewed) or greater than 1 (positive skewed) are considered highly skewed.

What Is Kurtosis?

Kurtosis is a statistical measure that quantifies the shape of a probability distribution. It provides information about the tails and peakedness of the distribution compared to a normal distribution.

Positive kurtosis indicates heavier tails and a more peaked distribution, while negative kurtosis suggests lighter tails and a flatter distribution. Kurtosis helps in analyzing the characteristics and outliers of a dataset.

The measure of Kurtosis refers to the tailedness of a distribution. Tailedness refers to how often the outliers occur.

Peakedness in a data distribution is the degree to which data values are concentrated around the mean. Datasets with high kurtosis tend to have a distinct peak near the mean, decline rapidly, and have heavy tails. Datasets with low kurtosis tend to have a flat top near the mean rather than a sharp peak.

In finance, kurtosis is used as a measure of financial risk. A large kurtosis is associated with a high level of risk for an investment because it indicates that there are high probabilities of extremely

large and extremely small returns. On the other hand, a small kurtosis signals a moderate level of risk because the probabilities of extreme returns are relatively low.

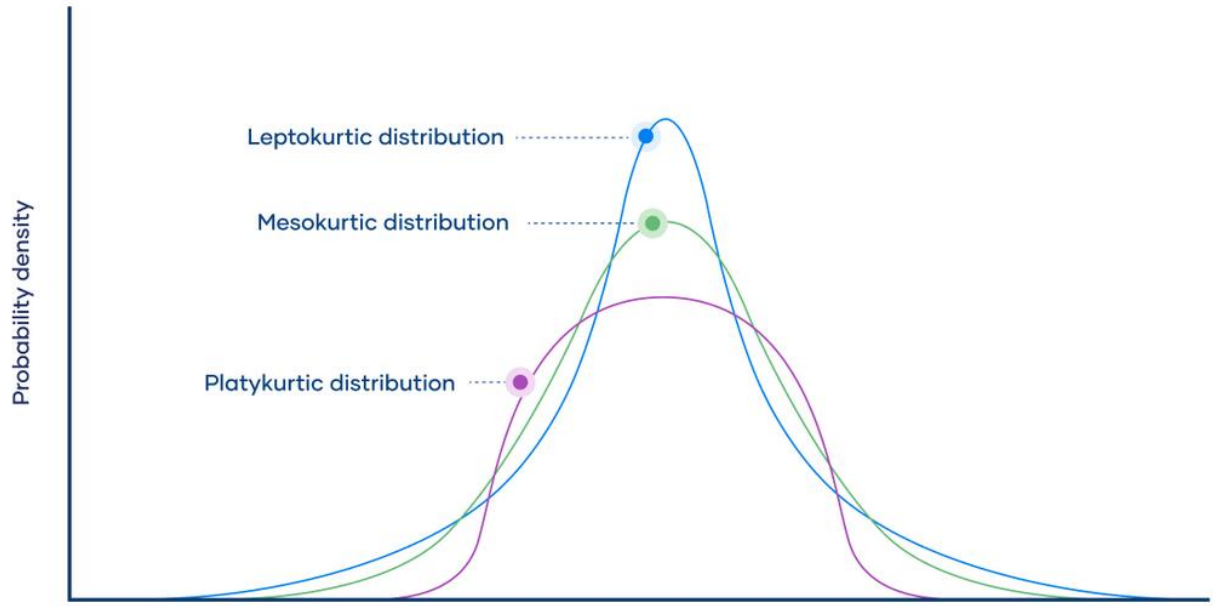
What Is Excess Kurtosis?

In statistics and probability theory, researchers use excess kurtosis to compare the kurtosis coefficient with that of a normal distribution. Excess kurtosis can be positive (Leptokurtic distribution), negative (Platykurtic distribution), or near zero (Mesokurtic distribution). Since normal distributions have a kurtosis of 3, excess kurtosis is calculated by subtracting kurtosis by 3.

Excess kurtosis = Kurt – 3

Types of Excess Kurtosis

1. Leptokurtic or heavy-tailed distribution (kurtosis more than normal distribution).
2. Mesokurtic (kurtosis same as the normal distribution).
3. Platykurtic or short-tailed distribution (kurtosis less than normal distribution).



Leptokurtic (Kurtosis > 3)

Leptokurtic has very long and thick tails, which means there are more chances of outliers. Positive values of kurtosis indicate that distribution is peaked and possesses thick tails. Extremely positive kurtosis indicates a distribution where more numbers are located in the tails of the distribution instead of around the mean.

Platykurtic (Kurtosis < 3)

Platykurtic having a thin tail and stretched around the center means most data points are present in high proximity to the mean. A platykurtic distribution is flatter (less peaked) when compared with the normal distribution.

Mesokurtic (Kurtosis = 3)

Mesokurtic is the same as the normal distribution, which means kurtosis is near 0. In Mesokurtic, distributions are moderate in breadth, and curves are a medium peaked height.

$$\text{Mesokurtic} = 3 - 3 = 0$$

Conclusion

Skewness and Kurtosis naturally complement each other in analyzing data distributions. Skewness, which measures the symmetry or asymmetry of data distribution, helps us understand if the data is pushed towards one side or the other. For instance, positive skewness indicates a distribution pushed towards the right side, while negative skewness implies a distribution pushed towards the left side. On the other hand, Kurtosis helps determine whether the data exhibits a heavy-tailed or light-tailed distribution. By incorporating both Skewness and Kurtosis into our analysis, we gain a more comprehensive understanding of the shape and characteristics of the data.

Skewed data may cause the tail region to act as an outlier for the statistical model, and such outliers can adversely impact the performance of the model, particularly in regression-based models. Some statistical models are robust to outliers like Tree-based models, but it will limit the possibility of trying other models. So, there is a necessity to transform the skewed data to be close enough to a Normal distribution.

Range

The range is defined as the difference between the highest and lowest value in a distribution of numbers. In order to compute the range, we must first locate the highest and lowest values. With a small number of values, one is able to inspect the set of numbers in order to identify these values.

When the set of numbers is large, however, a simple way to locate these values is to sort them in ascending order and then choose the first and last values, as we did in Chapter 3. Here is an example: Let us denote the lowest or first value with the symbol X1 and the highest value with Xn. Then the range (d) is:

$d = X_n - X_1$

with indices 1 and n defined after sorting the values.

Calculation is as follows:

Data set: 100, 95, 125, 45, 70

Sorted values: 45, 70, 95, 100, 125

Range = 125 – 45

Range = 80

Mean Absolute Deviation

A second method we use to describe variability is called the mean absolute deviation. This measure involves first calculating the mean of a set of observations or values and then determining the deviation of each observation from the mean of those values. Then we take the absolute value of each deviation, sum all of the deviations, and calculate their mean. The mean absolute deviation for a sample is:

$$\text{mean absolute deviation} = \frac{\sum_{i=1}^{n} |X_i - \bar{X}|}{n}$$

where n = number of observations in the data set.

The analogous formula for a finite population is:

$$\text{mean absolute deviation} = \frac{\sum_{i=1}^{N} |X_i - \mu|}{N}$$

where N = number of observations in the population.

Sample Variance and Standard Deviation

The symbols S^2 and S shall be used to denote sample variance and standard deviation, respectively, and are calculated by using Formulas:

$$S^2 = \frac{\sum_{i=1}^{n} (X_i - \bar{X})^2}{n-1}$$

$$S = \sqrt{\frac{\sum_{i=1}^{n} (X_i - \bar{X})^2}{n-1}}$$

where n is the sample size and $\bar{X}$ is the sample mean.

Example:

Blood Cholesterol Measurements for a Sample of 10 Persons:

(276, 304, 316, 188, 214, 252, 333, 271, 245, 198)

Person	X	$X-\overline{X}$	$(X-\overline{X})^2$	X_2
1	276	16.3	265.69	76,176
2	304	44.3	1,962.49	92,416
3	316	56.3	3,169.69	99,856
4	188	–71.7	5,140.89	35,344
5	214	–45.7	2,088.49	45,796
6	252	–7.7	59.29	63,504
7	333	73.3	5,372.89	110,889
8	271	11.3	127.69	73,441
9	245	–14.7	216.09	60,025
10	198	–61.7	3,806.89	39,204
Sum	2,597	0	22,210.10	696,651
		Mean = $\overline{X} = \Sigma X/n = 2{,}597/10 = 259.7$		
Variance	2467.788			
Std. Dev.	49.677			