Page no.75

**Skewness and Kurtosis:**

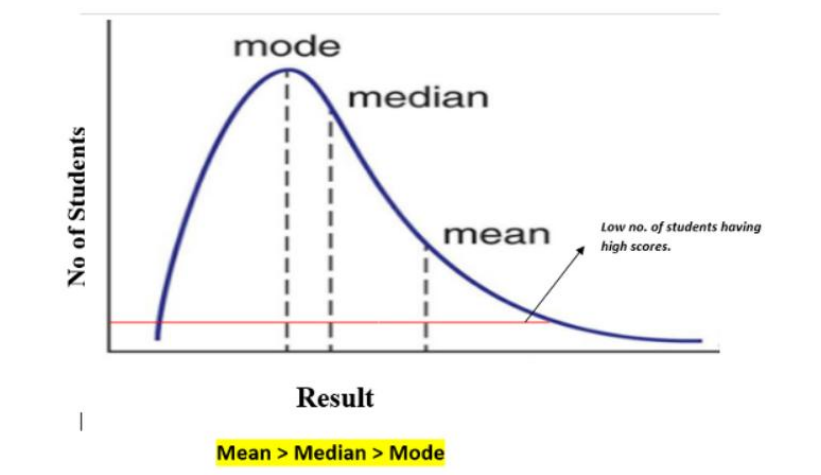
Skewness checks / describes (distribution) symmetry of data.

How symmetric is data and how asymmetric is data

**Mean = Median = Mode** then data is **symmetric**

When data is **symmetric** **skewness = 0**

**Positive Skewed or Right-Skewed (Positive Skewness) :**



Low values have high frequency

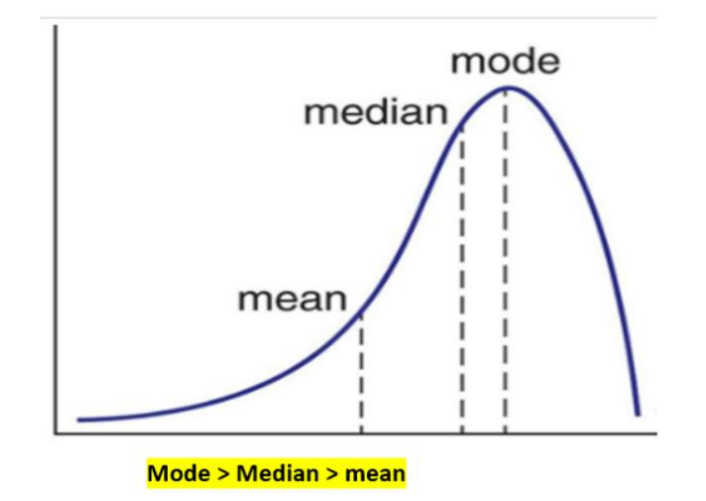
High values have low frequency

Ex. More number of employees have low salary, less employees have more salary

Ex. Result of difficult exam in class

Tail at right side

**Negative Skewed or Left-Skewed (Negative Skewness) :**

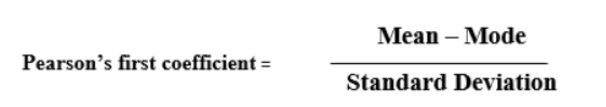
****

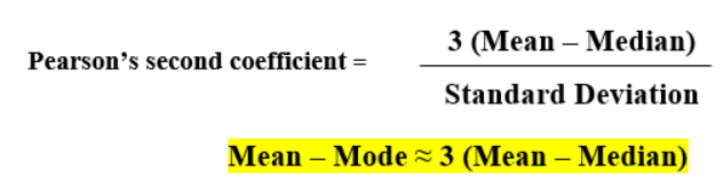
Low values have low frequency

High values have high frequency

Tail at left side

Ex. Result of very easy exam in class

To find skewness we use this formula:  




**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.

**Kurtosis :**

A statistical measure that quantifies the shape of a probability distribution.

It provides information about the tails and “peak”ness 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.

Types of Kurtosis

1. Mesokurtic: A distribution with mesokurtic kurtosis has a similar peak and tail

shape as the normal distribution. It has a kurtosis value of around 0, indicating

that its tails are neither too heavy nor too light compared to a normal distribution.

2. Leptokurtic: A distribution with leptokurtic kurtosis has heavier tails and a

sharper peak than the normal distribution. It has a positive kurtosis value,

indicating that it has more extreme outliers than a normal distribution. This type

of distribution is often associated with higher peakedness and a greater

probability of extreme values.

3. Platykurtic: A distribution with platykurtic kurtosis has lighter tails and a flatter

peak than the normal distribution. It has a negative kurtosis value, indicating that

it has fewer extreme outliers than a normal distribution. This type of distribution

is often associated with less peakedness and a lower probability of extreme

values.

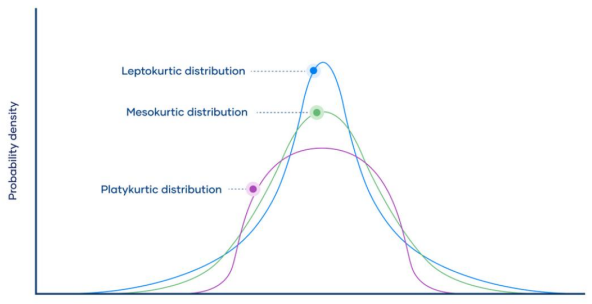
Types of Excess Kurtosis

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).

High frequency of values is at the peak.



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.

Key Differences of Skewness and Kurtosis

1. Skewness evaluates how much a distribution deviates from symmetry, while

Kurtosis gauges the degree of its peakiness or flatness.

2. Skewness is a measure derived from the third moment, whereas Kurtosis stems

from the fourth moment.

3. The range of values for both Skewness and Kurtosis spans from negative infinity

to positive infinity.

4. Perfect symmetry and normality are indicated by both zero skewness and zero

kurtosis.

5. Skewness can impact the central tendency of a distribution, whereas kurtosis can

influence its tail behavior.

6. Both Skewness and Kurtosis provide insights into the shape characteristics of

Distributions

One liner 🡪 Skewness = symmetry | kurtosis = peakedness

Page no.18

**Correlation :**

* statistical relationship between the two entities.
* height and weight of a person are related, and taller people tend to be heavier than shorter people.

positive correlation 🡪 One value increases ALONG WITH IT another value is increasing

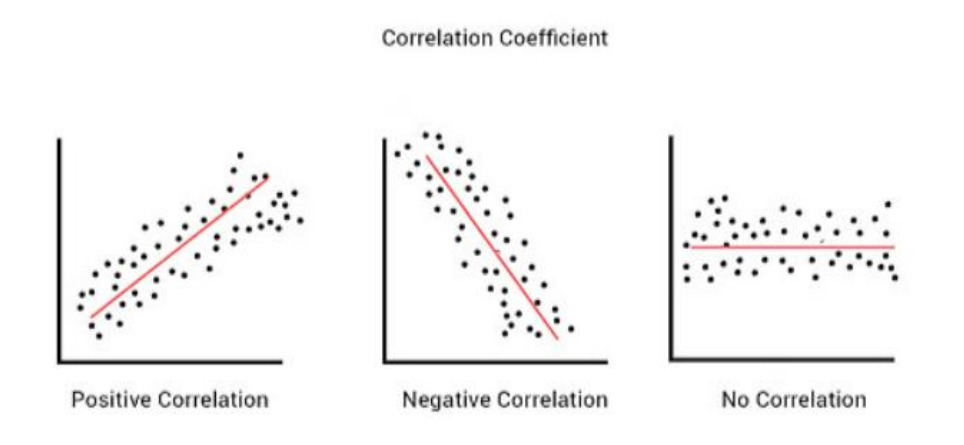
negative correlation 🡪 One value decreasing ALONG WITH IT another value is decreasing

(Spurious) No Correlation: Appear to be related but are not. Has no underlying casual connection. Relation occurred by third value (a confounder / confounding variable)

**Causation :**

positive correlation 🡪 One value increases DUE TO IT another value is increasing

negative correlation 🡪 One value decreasing DUE TO IT another value is decreasing



**Correlation Coefficient :**

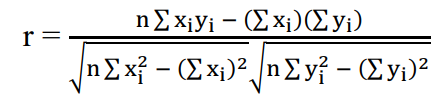
The correlation coefficient, r, is a summary measure that describes the extent of the statistical

relationship between two interval or ratio level variables. The correlation coefficient is scaled

so that it is always between -1 and +1. When r is close to 0 this means that there is little

relationship between the variables and the farther away from 0 r is, in either the positive or

negative direction, the greater the relationship between the two variables.

****

Where,

* n = no.of observations
*  🡪 n\* sum ( xi2 )
*  🡪 ( sum( xi ) )2
* same for y.
* to solve , have to write columns of x, y , x2 , y2, xi\*yi

Note:

1) r lies between −1 & 1 i.e. −1  r  1

2) If r = 1, there is perfect positive correlation

3) If 0 < r < 1, there is positive correlation

4) If r = −1, there is perfect negative correlation

5) If −1 < r < 0, there is negative correlation

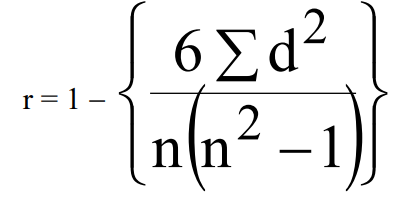
6) If r = 0, there is no correlation

7) Correlation Coefficient is independent of change of origin & change of scale.

**Spearman’s Rank Correlation Coefficient:**

In this method, ranks are assigned to the data. The ranks are given to the x-series & yseries separately. The highest observation is given rank ‘1’, the next highest observation is given rank ‘2’ and so on. Suppose, R1 & R2 are the ranks of the x & y respectively and

**d = R1 − R2 then**

****

where n = number of pairs of observations

**Ex: Calculate the Spearman’s rank correlation coefficient for the following data.**

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **X** | 15 | 12 | 16 | 13 | 17 | 14 | 18 | 11 |
| **y** | 17 | 14 | 20 | 25 | 23 | 24 | 22 | 21 |

**Ans :**

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **X** | 15 | 12 | 16 | 13 | 17 | 14 | 18 | 11 |
| **y** | 17 | 14 | 20 | 25 | 23 | 24 | 22 | 21 |
|  |  |  |  |  |  |  |  |  |
| **R1** ( rank x ) | 4 | 7 | 3 | 6 | 2 | 5 | 1 | 3 |
| **R2** ( rank y ) | 7 | 8 | 6 | 1 | 3 | 2 | 4 | 5 |
|  |  |  |  |  |  |  |  |  |
| **d ( R2 – R1 )** | 3 | 1 | 3 | -5 | 1 | 3 | 3 | 2 |
| **d2** | 9 | 1 | 9 | 25 | 1 | 9 | 9 | 4 |

Now Put values into formula.

**Causation is not correlation**

**Covariance :**

Covariance is a measure of the relationship between two random variables and to what

extent, they change together. Or we can say, in other words, it defines the changes between

the two variables, such that change in one variable is equal to change in another variable.

This is the property of a function of maintaining its form when the variables are linearly

transformed. Covariance is measured in units, which are calculated by multiplying the units

of the two variables.

Two types:

1. Positive Covariance

2. Negative Covariance

**Positive Covariance**

If the covariance for any two variables is positive, that means, both the variables move in the

same direction. Here, the variables show similar behaviour. That means, if the values (greater

or lesser) of one variable corresponds to the values of another variable, then they are said to

be in positive covariance.

**Negative Covariance**

If the covariance for any two variables is negative, that means, both the variables move in the

opposite direction. It is the opposite case of positive covariance, where greater values of

one variable correspond to lesser values of another variable and vice-versa.

**Covariance Formula**

Covariance formula is a statistical formula, used to evaluate the relationship between two

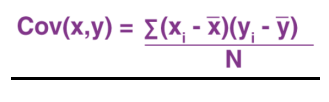
variables. It is one of the statistical measurements to know the relationship between the

variance between the two variables. Let us say X and Y are any two variables, whose

relationship has to be calculated. Thus the covariance of these two variables is denoted by

Cov(X,Y). The formula is given below for both population covariance and sample

covariance.

****

If cov(X, Y) is greater than zero, then we can say that the covariance for any two variables is

positive and both the variables move in the same direction.

If cov(X, Y) is less than zero, then we can say that the covariance for any two variables is

negative and both the variables move in the opposite direction.

If cov(X, Y) is zero, then we can say that there is no relation between two variables.

**Regression :**

Correlation just checks whether variables are related or not but regression predicts their relationship.