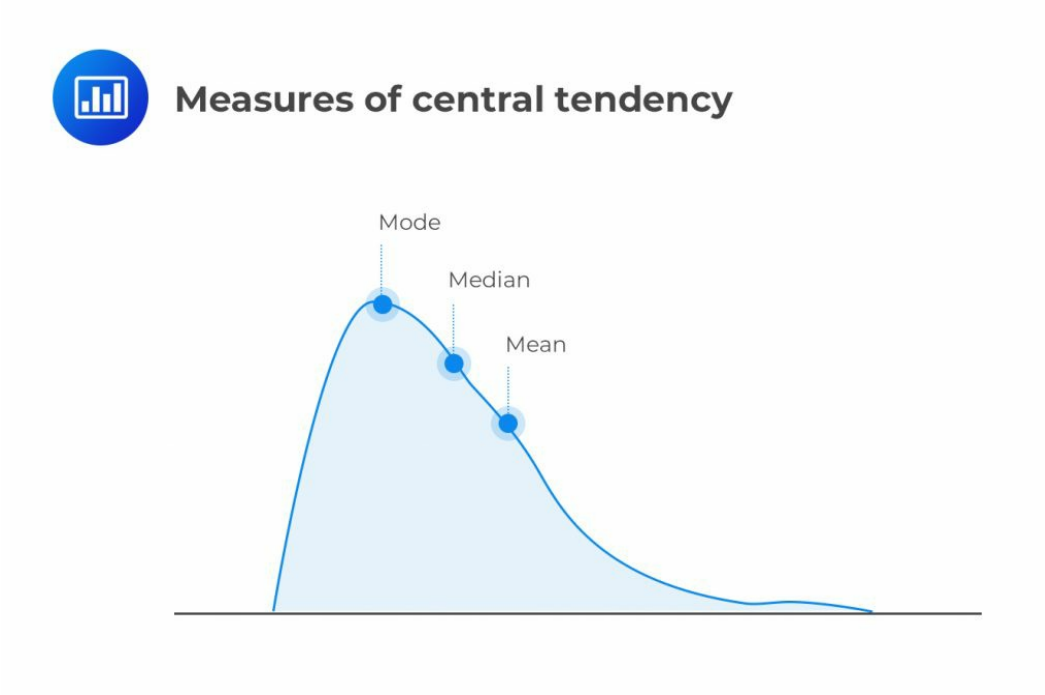
Learning Module 3: Statistical Measures of Asset Returns

LOS 3a: calculate, interpret, and evaluate measures of central tendency and location to address an investment problem

The center of any data is identified via a measure of central tendency. A measure of central tendency for a series of returns reveals the center of the empirical distribution of returns. They include mean, mode, and median.



Measures of location help us understand where data points tend to cluster. These measures include central tendency measures such as mean, median, and mode. There are also other measures that provide different insights into how the data is spread out or located within a distribution.

Measures of Central Tendency

Arithmetic Mean

The arithmetic mean is the sum of the values of the observations in a dataset divided by the

number of observations.

Recall the formula: denoted by $\bar{R}_i$ arithmetic mean for an asset *i* is a simple process of finding the average holding period returns. It is given by:

$$\bar{R}_i = \frac{R_{i,1} + R_{i,1} + \ldots + R_{i,T-1} + R_{iT}}{T} = \frac{1}{T}\sum_{t=1}^{T} R_{it}$$

Where:

R_{it} = Return of asset *i* in period *t*.

T = Total number of periods.

For example, if a share has returned 15%, 10%, 12%, and 3% over the last four years, then the arithmetic mean is computed as follows:

$$\bar{R}_i = \frac{1}{T}\sum_{t=1}^{T} R_{it} = \frac{1}{4}(15\% + 10\% + 12\% + 3\%) = 10\%$$

Population Mean and Sample Mean

The population mean is the summation of all the observed values in the population, $\sum X_i$ divided by the total number of observations, N. The population mean differs from the sample mean, which is based on a few observed values chosen from the population. Thus:

$$\text{Population mean} = \frac{\sum X_i}{N}$$

$$\text{Sample mean} = \frac{\sum X_i}{n}$$

Analysts use the sample mean to *estimate* the actual population mean.

The population mean and the sample mean are both arithmetic means. The arithmetic mean for any data set is unique and is computed using all the data values. Among all the measures of central tendency, it is the only measure for which the sum of the deviations from the mean is

zero.

Example: Calculating the Arithmetic Mean

The following are the annual returns realized from a given asset between 2005 and 2015.

{12% 13% 11.5% 14% 9.8% 17% 16.1% 13% 11% 14%}

Calculate the population mean.

Solution

$$\text{Population mean} = \frac{0.12 + 0.13 + 0.115 + 0.14 + 0.098 + 0.17 + 0.161 + 0.13 + 0.11 + 0.14}{10}$$
$$= 0.1314 = 13.14\%$$

Median

The median is the statistical value located at the center of a data set organized in ascending or descending order.

Consider a sample of n observations. For an odd-numbered of observations, the median is the observation that is located in $\frac{(n+1)}{2}$ position. On the other hand, if the number of observations is even, the media is the mean value of the observations located in $\frac{n}{2}$ and $\frac{n+2}{2}$ position.

Unlike the arithmetic mean, the median resists the effects of extreme observations. However, it only gives the relative position of the ranked observations without considering all observations relating to the size of the observations.

Example: Calculating the Median

The following are the annual returns on a given asset realized between 2005 and 2015.

{12% 13% 11.5% 14% 9.8% 17% 16.1% 13% 11% 14%}

The median is *closest* to:

Solution

First, we arrange the returns in ascending order:

{9.8% 11% 11.5% 12% 13% 13% 14% 14% 16.1% 17%}

Since the number of observations is even, the median return will be the middle point of the two middle values in the positions $\frac{n}{2}$ and $\frac{(n+2)}{2}$.

The value occupying $\frac{n}{2} = \frac{10}{5} = 5$th position is 13, and the value located in $\frac{(n+2)}{2} = \frac{12}{2} = 6$th position is 13, so that the mode is:

$$\frac{13\% + 13\%}{2} = 13\%$$

Mode

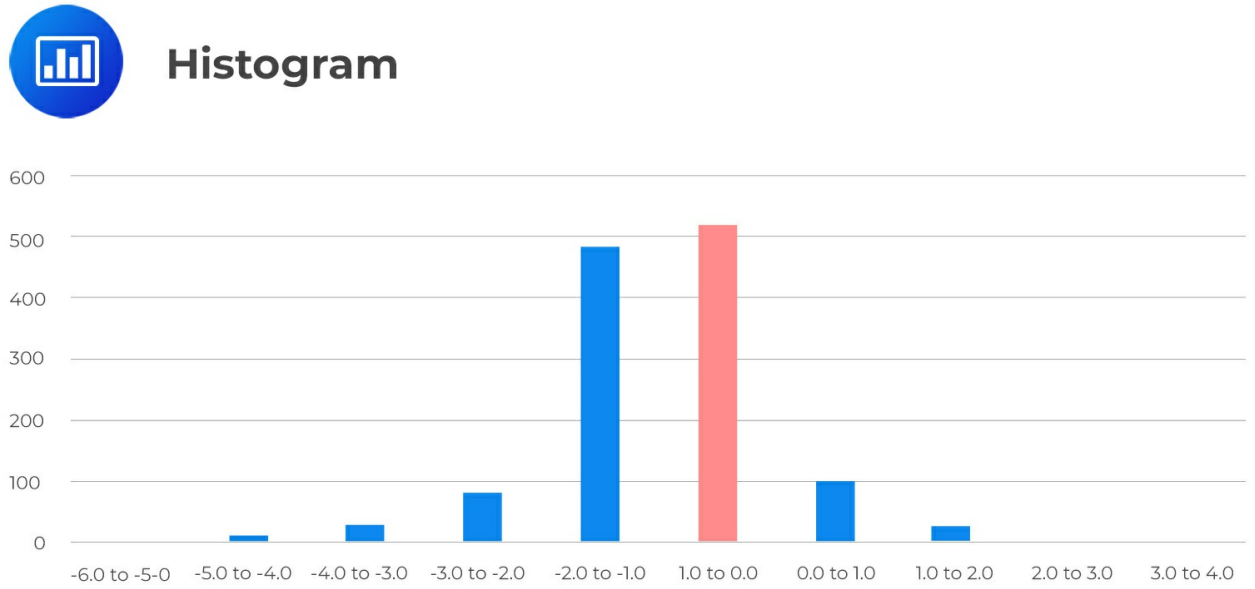
The mode is the value that appears most often in a dataset. Sometimes, a dataset has a mode; sometimes, it doesn't. If all the observations in a dataset are different and no value repeats more than others, then the dataset has no mode.

A dataset with one mode is called unimodal. When there are two modes, it's called bimodal. If the distribution has three frequently occurring values, it's termed trimodal.

An interval with the highest frequency is called the modal interval (or intervals) in a frequency distribution. For instance, in the frequency distribution below, the modal interval is -1.0 to 0.0 with an absolute frequency of 520.

Return Bin (%)	Absolute Frequency	Relative Frequency(%)	Cumulative Absolute Frequency	Cumulative] Relative Frequency (%)
-6.0 to -5.0	2	0.16	2	0.16
-5.0 to -4.0	8	0.64	10	0.80
-4.0 to -3.0	27	2.16	37	2.96
-3.0 to -2.0	80	6.40	117	9.36
-2.0 to -1.0	485	38.80	602	48.16
-1.0 to 0.0	520	41.60	1,122	89.76
0.0 to 1.0	100	8.00	1,222	97.76
1.0 to 2.0	24	1.92	1,246	99.68
2.0 to 3.0	3	0.24	1,249	99.92
3.0 to 4.0	1	0.08	1,250	100.00

In a histogram, the modal interval always has the highest bar.



The mode is the only measure of central tendency that can be used with nominal data. Nominal data refers to a type of data that is categorized into distinct categories or groups without any inherent order or numerical value. Examples of nominal data include gender (male, female), eye color (blue, brown, green), and marital status (single, married, divorced).

Example: Calculating the Mode

Determine the mode from the following data set:

{20% 23% 20% 16% 21% 20% 16% 23% 25% 27% 20%}

Solution

The mode is 20%. It occurs four times, a frequency higher than any other value in the data set. Clearly, this dataset is unimodal.

Dealing With Outliers

An outlier may represent a distinct value in a population. In addition, it may show that there was an error in recording the value, or it was generated from a different population.

When working with a sample with outliers, we can potentially transform the variable or choose another variable that achieves the same objective. If these observations prove to be impossible, there are three options:

Option 1: Take no action and use the data as it is. If these observations are accurate, then this is appropriate.

Option 2: Remove outliers using the trimmed mean. For example, when calculating the central tendency, a 4 percent trimmed mean excludes the lowest 2% and highest 2% of values.

Option 3: Substitute a different value for the outliers. The winsorized mean is an illustration of a central tendency that does this. For instance, when computing a 96% winsorized mean, the value at or above the lowest and highest 2% is assigned the lowest and highest 2% values.

Measures of Location

Quartiles, quintiles, deciles, and percentiles are values or cut points that partition a finite number of observations into nearly equal-sized subsets. The number of partitions depends on the type of cut point involved.

Quartiles

They divide data into **four** parts. The first quartile, Q_1, is referred to as the lower quartile, and the last quartile, $Q_{4,}$ is the upper quartile. Q_1 splits the data into the lower 25% and upper 75% values. Similarly, the upper quartile subdivides the data into the lower 75% of the values and the upper 25%. The difference between the upper and lower quartiles is known as the **interquartile range**, which indicates the spread of the middle 50% of the data.

Quintiles

Though rarely used in practice, quintiles split a set of data into **five** equal parts, i.e., fifths. Therefore, the second quintile splits data into the lower 40% of the values and the upper 60%.

Deciles

The deciles subdivide data into **ten** equal parts. There are 10 deciles in any data set. For example, the fourth decile splits data into the lower 40% of the values and the upper 60%.

Percentiles

Percentiles split data into **100** equal parts, i.e., hundredths. So, for instance, the 77^{th} percentile splits the data into the lower 77% of the values and the upper 23%.

Financial analysts commonly use the four types of subdivisions to rank investment performance. You should note that quartiles, quintiles, and deciles can all be expressed as percentiles. For instance, the first quartile is just the 25^{th} percentile. Similarly, the fourth decile is simply the 40^{th} percentile. This enables the application of the formula below.

$$\text{Position of percentile} = \frac{(n+1)\,y}{100}$$

Example: Calculating Quartiles

Given the following distribution of returns, calculate the lower quartile.

{10% 23% 12% 21% 14% 17% 16% 11% 15% 19%}

Solution

First, we have to arrange the values in ascending order:

{10% 11% 12% 14% 15% 16% 17% 19% 21% 23%}

Next, we establish the position of the first quartile. This is simply the 25^{th} percentile. Therefore:

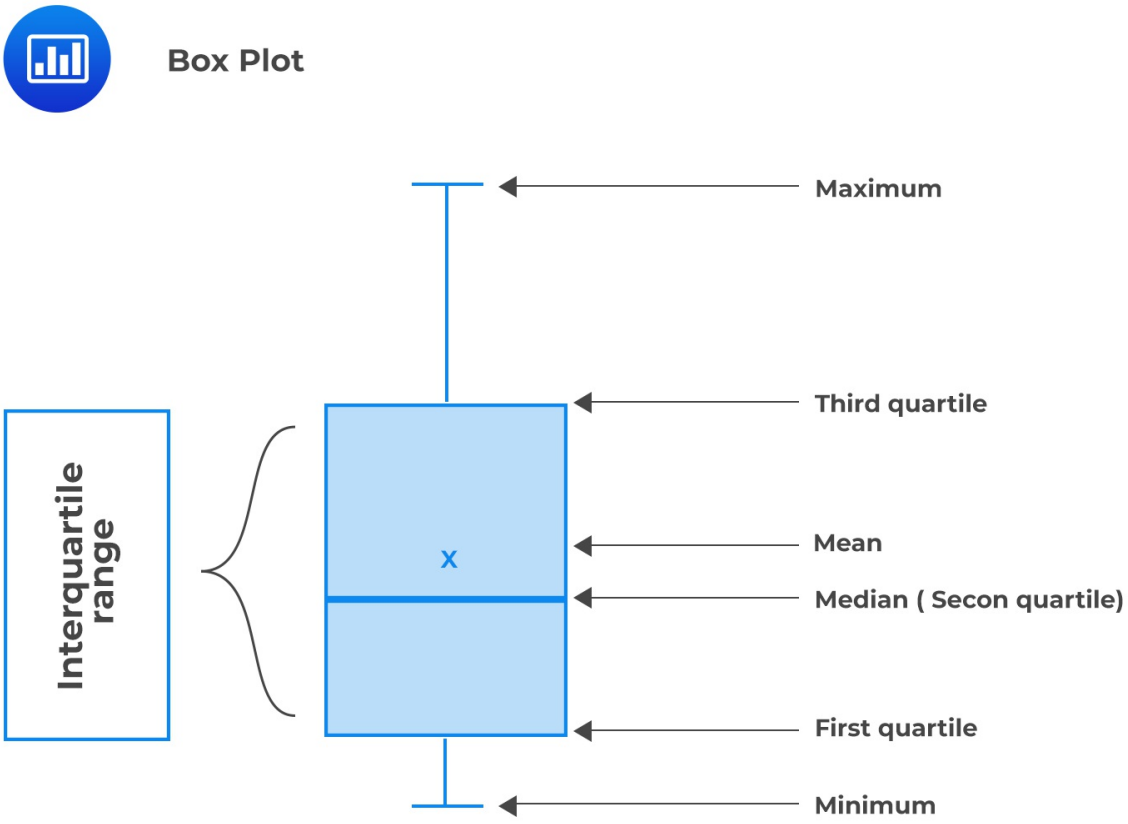
$$P_{25} = \frac{(10 + 1)\,25}{100} = 2.75^{th} \text{ value}$$

Since the value is not straightforward, we have to extrapolate between the 2^{nd} and the 3^{rd} data points. The 25^{th} percentile is three-fourths (0.75) of the way from the 2^{nd} data point (11%) to the 3^{rd} data point (12%):

$$11\% + 0.75 \times (12 - 11) = 11.75\%$$

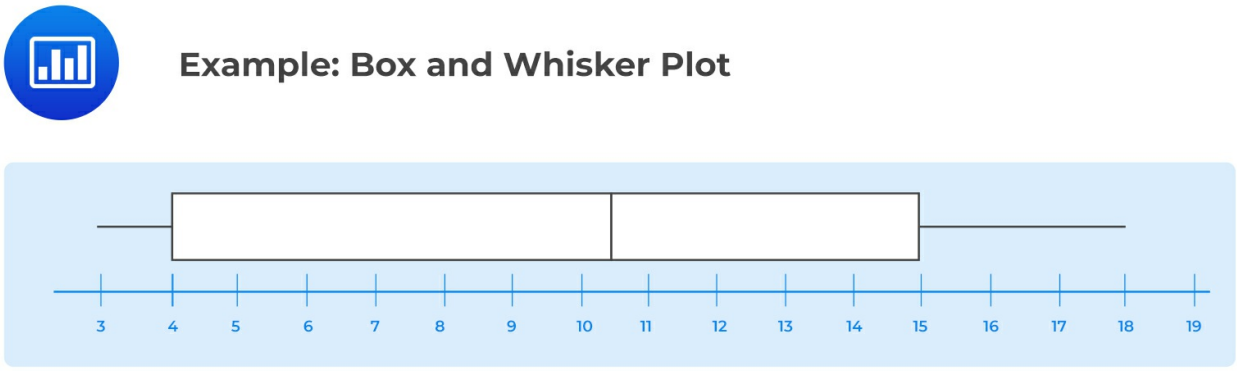
Box and Whisker Plot

Box and whisker plot is used to display the dispersion of data across quartiles. A box and whisker plot consists of a "box" with "whiskers" connected to the box. It shows the following five-number summary of a set of data.



Example: Box and Whisker Plot

Consider the following box and whisker plot:



1. Which of the following is *most likely* the median?

A. 10.

B. 10.5.

C. 11.

Solution

The correct answer is B.

$$\text{Median or Quartile 2 (Q2)} = \frac{(10 + 11)}{2} = 10.5$$

2. Which of the following is most likely the interquartile range?

A. 4.5.

B. 6.5.

C. 11.

Solution

The correct answer is C.

$$\text{Interquartile range is } Q_3 - Q_1 = 15 - 4 = 11$$

3. Which of the following is *most likely* the 3rd quartile?

A. 4.

B. 15.

C. 18.

Solution

$$\text{Quartile 3 (Q3)} = \frac{3 \times (n + 1)}{4} = \frac{3 \times (16 + 1)}{4} = 12.75\text{th term, which is } 14.75$$

Quantiles in Investment Practice

Quantiles have two main purposes in investment practice. First, quantiles are used to rank performance. Secondly, quantiles can be used in investment research for comparison purposes. For example, companies can be clustered into deciles to compare the performance of small companies with the large ones. In this case, the first decile will contain the portfolio of companies with the smallest market values, while the tenth decile will contain the companies with the largest market values.

Question

A mutual fund achieved the following rates of growth over an 11-month period:

{3% 2% 7% 8% 2% 4% 3% 7.5% 7.2% 2.7% 2.09%}

The 5^{th} decile from the data is *closest* to:

A. 2%.

B. 3%.

C. 4%.

Solution

The correct answer is B.

First, you should re-arrange the data in ascending order:

{2% 2% 2.09% 2.7% 3% 3% 4% 7% 7.2% 7.5% 8%}

Secondly, you should establish the 5^{th} decile. This is simply the 50^{th} percentile and is actually the median:

$$\begin{aligned} P_{50} &= \frac{(1 + 11)\,50}{100} \\ &= 12 \times 0.5 \\ &= 6, \text{ i.e., the 6th data point} \end{aligned}$$

Therefore, the 5th decile = 50th percentile = Median = 3%

LOS 3b: calculate, interpret, and evaluate measures of dispersion to address an investment problem

Measures of dispersion are used to describe the variability or spread in a sample or population. They are usually used in conjunction with measures of central tendency, such as the mean and the median. Specifically, measures of dispersion are the range, variance, absolute deviation, and standard deviation.

Measures of dispersion are essential because they give us an idea of how well the measures of central tendency represent the data. For example, if the standard deviation is large, then there are large differences between individual data points. Consequently, the mean may not be representative of the data.

Range

The range is the difference between the highest and the lowest values in a dataset, i.e.,

$$\text{Range} = \text{Maximum value} - \text{Minimum value}$$

Example: Calculating the Range

Consider the following scores of 10 level I candidates:

{78 56 67 51 43 89 57 67 78 50}

$$\text{Range} = 89 - 43 = 46$$

Advantage of the Range

- The range is easy to compute.

Disadvantages of the Range

- The range is not a reliable dispersion measure. It provides limited information about the distribution because it uses only two data points.

- The range is sensitive to outliers.

Mean Absolute Deviation (MAD)

MAD is a measure of dispersion representing the **average of the absolute values** of the deviations of individual observations from the arithmetic mean. Therefore,

$$\text{MAD} = \frac{\sum |X_i - \bar{X}|}{n}$$

Remember that the sum of deviations from the arithmetic means is always zero, which is why we use **absolute values**.

Example: Calculating Mean Absolute Deviation

Six financial analysts have reported the following returns on six different large-cap stocks over 2021:

{6% 7% 12% 2% 3% 11%}

Calculate the mean absolute deviation and interpret it.

Solution

First, we have to calculate the arithmetic mean:

$$\bar{X} = \frac{(6\% + 7\% + 12\% + 2\% + 3\% + 11\%)}{6} = 6.83\%$$

Next, we can now compute the MAD:

$$\begin{aligned}\text{MAD} &= \frac{\{|6\% - 6.83\%| + |7\% - 6.83\%| + |12\% - 6.83\%| + |2\% - 6.83\%| + |3\% - 6.83\%| + |11\% - 6.8}{6} \\ &= \frac{0.83 + 0.17 + 5.17 + 4.83 + 3.83 + 4.17}{6} \\ &= 3.17\%\end{aligned}$$

Interpretation: On average, an individual return deviates by 3.17% from the mean return of 6.83%.

Sample Variance and Sample Standard Deviation

The sample variance,s^2, is the measure of dispersion that applies when working with a sample instead of a population.

$$s^2 = \frac{\sum (X_i - \bar{X})^2}{n - 1}$$

Where:

$\bar{X}$ = Sample mean.

n = Number of observations.

Note that we are dividing by n–1. This is necessary to remove **bias**.

The sample standard deviation, s, is simply the square root of the sample variance.

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

Example: Calculating Sample Mean and Variance

Assume that the returns realized in the previous example were sampled from a population comprising 100 returns. The sample mean and the corresponding sample variance are *closest* to:

Solution

The sample mean will still be 6.83%.

Hence,

$$s^2 = \frac{\{(6\% - 6.83\%)^2 + (7\% - 6.83\%)^2 + (12\% - 6.83\%)^2 + (2\% - 6.83\%)^2 + (3\% - 6.83\%)^2 + (11\%}{5}$$
$$= 0.001656$$

Therefore,

$$s = 0.001656^{\frac{1}{2}}$$
$$= 0.0407$$

Downside Deviation and Coefficient of Variation

When trying to estimate downside risk (i.e., returns below the mean), we can use the following measures:

- **Semi-variance**: The average squared deviation below the mean.
- **Semi-deviation** (also known as semi-standard deviation): The positive square root of semi-variance.
- **Target semi-variance**: The sum of the squared deviations from a specific target return.
- **Target semi-deviation**: The square root of target semi-variance.

Sample Target Semi-Deviation

The target semi deviation, s_{Target} , is calculated as follows:

$$s_{Target} = \sqrt{\sum_{\text{for all } X_i \leq B}^{n} \frac{(X_i - B)^2}{n - 1}}$$

Where B is the target and n is the total number of sample observations.

Yearly returns of an equity mutual fund are provided as follows.

Month	Return %
2010	36%
2011	29%
2012	10%
2013	52%
2014	41%
2015	16%
2016	10%
2017	23%
2018	−10%
2019	−19%
2020	2%

What is the target downside deviation if the target return is 20%?

Solution

Month	Return %	Deviation from the 20% target	Deviation below the target	Squared deviations below the target
2010	36.00	16.00	−	−
2011	29.00	9.00	−	−
2012	10.00	(10.00)	(10.00)	100
2013	52.00	32.00	−	
2014	41.00	21.00	−	
2015	16.00	(4.00)	(4.00)	16
2016	10.00	(10.00)	(10.00)	100
2017	23.00	3.00	−	
2018	(10.00)	(30.00)	(30.00)	900
2019	(19.00)	(39.00)	(39.00)	1,521
2020	2.00	(18.00)	(18.00)	324
Sum				2,961

Here n = 11–1 = 10 so that:

$$\text{Target semi-deviation} = \left(\frac{2961}{10}\right)^{0.5} = 17.21\%$$

Coefficient of Variation

The coefficient of variation, CV, is a measure of spread that describes the amount of variability of data relative to its mean. It has **no units**, so we can use it as an alternative to the standard deviation to compare the variability of data sets that have different means. The coefficient of variation is given by:

$$CV = \frac{s}{\bar{X}}$$

Where:

s = Standard deviation of a sample.

$\bar{X}$ = Mean of the sample.

***Note*:** The formula can be replaced with $\frac{?}{?}$ when dealing with a population.

Procedure to Follow While Calculating the Coefficient of Variation:

1. Compute the mean of the data.
2. Calculate the sample standard deviation of the data set, s.
3. Find the ratio of s to the mean, x?.

Example: Coefficient of Variation

What is the relative variability for the samples 40, 46, 34, 35, and 45 of a population?

Solution

Step 1: Calculate the mean.

$$\text{Mean} = \frac{(40 + 46 + 34 + 35 + 45)}{5} = \frac{200}{5} = 40$$

Step 2: Calculate the sample standard deviation. (Start with the variance, s^2.)

$$s^2 = \frac{(40-40)^2+\ldots+(45-40)^2}{4}$$
$$= \frac{122}{4}$$
$$= 30.5$$

Note: Since it is the sample standard deviation (not the population standard deviation), we use $n-1$ as the denominator.

Therefore,

$$s = \sqrt{30.5} = 5.52268$$

Step 3: Calculate the ratio.

$$\frac{\text{Mean}}{s} = \frac{5.52268}{40} = 0.13806 \text{ or } 13.81\%$$

Interpreting the Coefficient of Variation

In finance, the coefficient of variation is used to measure the **risk per unit of return**. For example, imagine that the mean monthly return on a T-Bill is 0.5% with a standard deviation of 0.58%. Suppose we have another investment, say, Y, with a 1.5% mean monthly return and standard deviation of 6%; then,

$$CV_{T-Bill} = \frac{0.58}{0.5} = 1.16$$

$$CV_Y = \frac{6}{1.5} = 4$$

Interpretation: The dispersion per unit monthly return of T-Bills is less than that of Y. Therefore, investment Y is riskier than an investment in T-Bills.

Question 1

If a security has a mean expected return of 10% and a standard deviation of 5%, its coefficient of variation is *closest* to:

A. 0.005.

B. 0.500.

C. 2.000.

Solution

The correct answer is **B**.

$$\text{CV} = \frac{\text{S}}{\text{x?}} = \frac{0.05}{0.10} = 0.5$$

Where:

s = The standard deviation of the sample.

x? = The mean of the sample.

A is incorrect. It assumes the following calculation.

$$\text{CV} = \frac{0.05}{10} = 0.005$$

C is incorrect. It assumes the following calculation.

$$\text{CV} = \frac{10}{5} = 2$$

Question 2

You have been given the following data:

{12 13 54 56 25}

Assuming that this is a sample from a certain population, the sample standard deviation is *closest* to:

A. 21.62.

B. 374.00.

C. 1,870.00.

The correct answer is **A**.

$$\bar{X} = \frac{(12 + 13 + \cdots + 25)}{5} = \frac{160}{5} = 32$$

Hence,

$$s^2 = \frac{\{(12 - 32)^2 + (13 - 32)^2 + (54 - 32)^2 + (56 - 32)^2 + (25 - 32)^2\}}{4}$$
$$= \frac{1870}{4} = 468$$

Therefore,

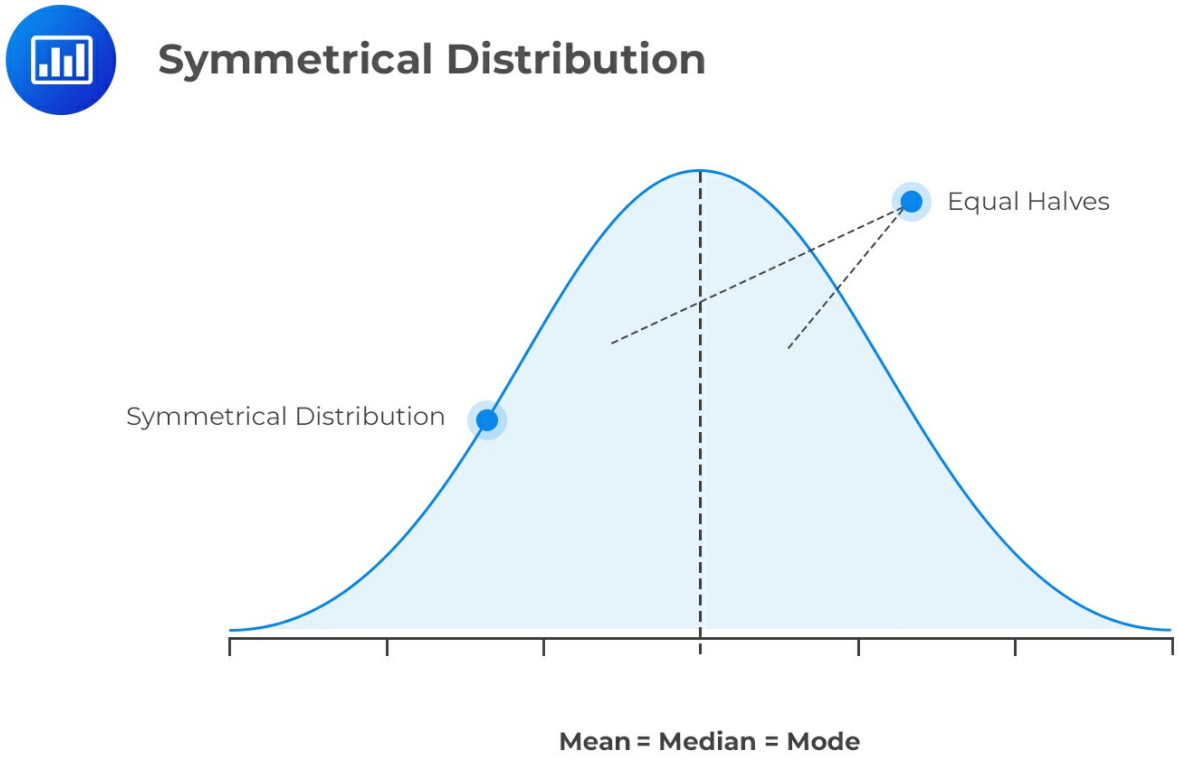
$$s = \sqrt{468} = 21.62$$

LOS 3c: interpret and evaluate measures of skewness and kurtosis to address an investment problem

Since the deviations from the mean are squared when calculating variance, we cannot determine whether significant deviations are more likely to be positive or negative. In order to identify other crucial distributional traits, we must look beyond measures of central tendency, location, and dispersion.

Skewness

Skewness refers to the degree of deviation from a symmetrical distribution, such as the normal distribution. A symmetrical distribution has identical shapes on either side of the mean.



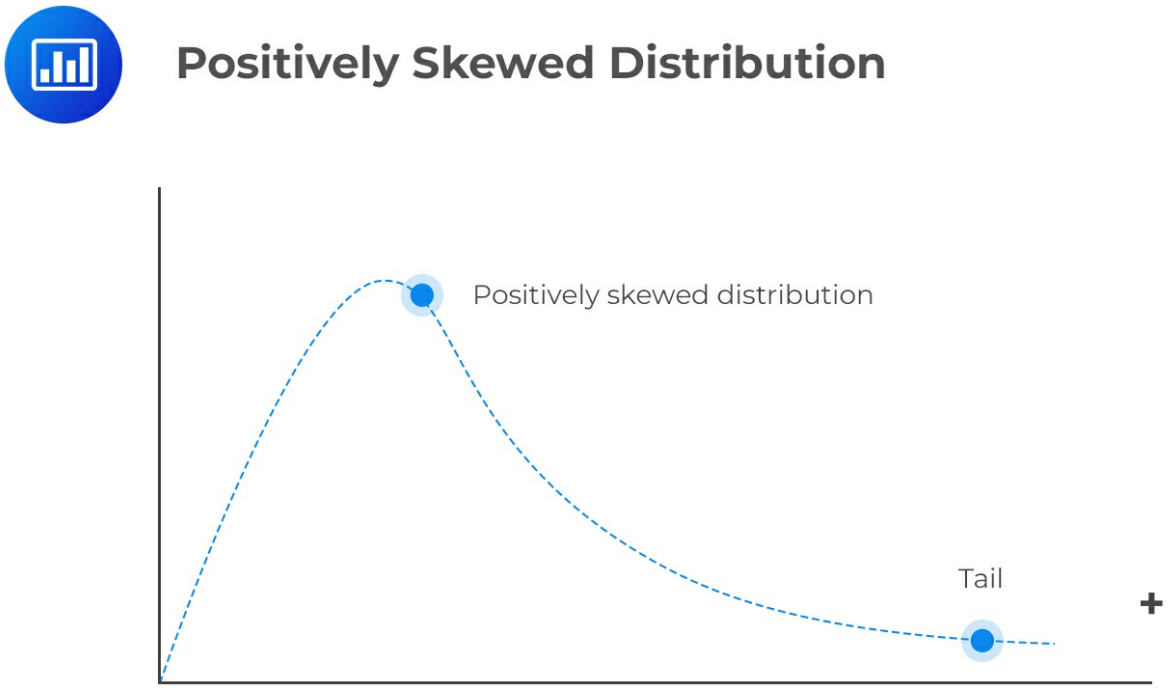
Distributions that are nonsymmetrical have unequal shapes on either side of the mean, leading to skewness. This is because nonsymmetrical distributions depart from the usual bell shape of the

normal distribution.

Skewness can be positive, negative, or, in some cases, undefined. The shape of a skewed distribution depends on outliers, which are extremely negative and positive observations.

Positive Skewness

A positively skewed distribution has a **long right tail** because of many outliers or extreme values on the right side. Perhaps the best way to remember its shape is to consider its points in a positive direction. Most data points are concentrated on the left side.

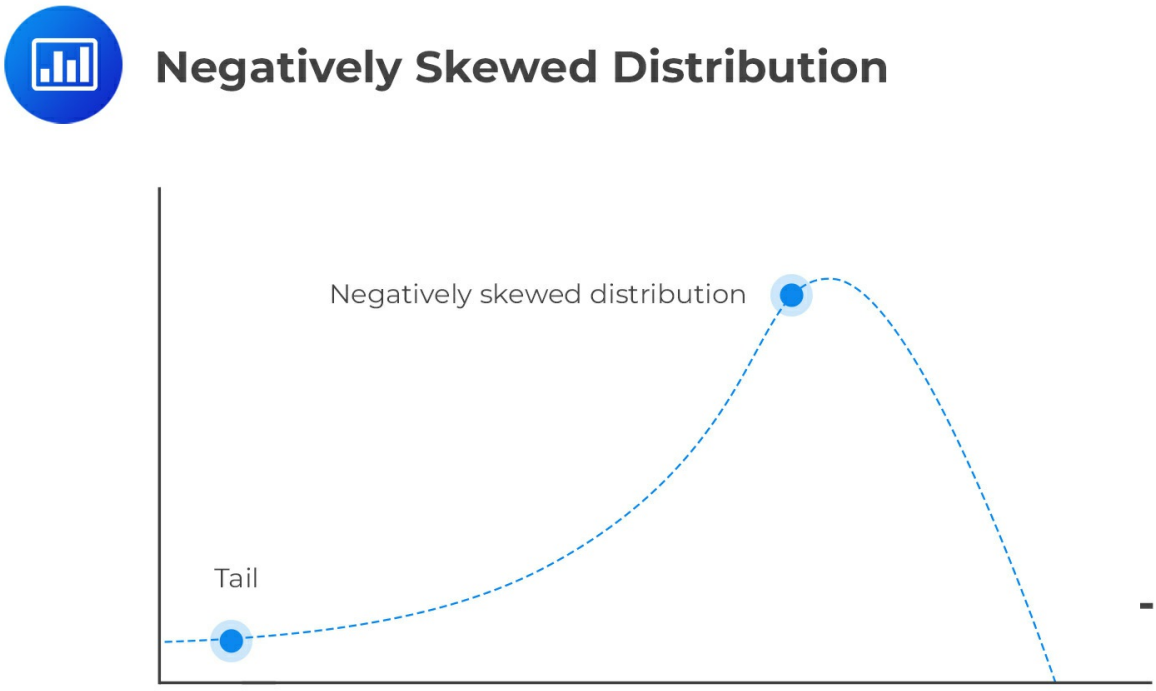


An example of a positively skewed distribution would be the income of individuals living in a specific country.

Negative Skewness

A negatively skewed distribution has a long left **tail** resulting from many outliers on the left side of the distribution. Therefore, we could say that it points in the negative direction. This is

because the right side harbors most of the data points.



Application of Skewness

Skewness matters in finance. Market data often show positive or negative skewness, like stock prices or mortgage costs. Investors can predict if future prices will be above or below the mean based on the skewness of the market segment.

Calculating Sample Skewness

The approximate sample skewness when sample is large ($n \geq 100$) is given by:

$$\text{Skewness} = \left(\frac{1}{n}\right) \frac{\sum_{i=1}^{n} (X_i - \bar{X})^3}{s^3}$$

Where:

$\bar{X}$ = Sample mean.

s = Sample standard deviation.

n = Number of observations.

A positive value indicates positive skewness. A 'zero' value indicates that the data is not skewed. Lastly, a negative value indicates negative skewness or a negatively skewed distribution.

Example: Calculating Skewness

Suppose we have the following observations:

{12 13 54 56 25}

What is the skewness of the data?

Solution

First, we must determine the sample mean and the sample standard deviation:

$$\bar{X} = \frac{(12 + 13 + 54 + 56 + 25)}{5} = \frac{160}{5} = 32$$

$$s^2 = \frac{(12 - 32)^2 + (13 - 32)^2 + \ldots (25 - 32)^2}{4}$$

$$= 467.5$$

Therefore,

$$s = \sqrt{467.5} = 21.62$$

Now we can work out the skewness:

$$\text{Skewness} = \left(\frac{1}{n}\right) \frac{\sum_{i=1}^{n} (X_i - \bar{X})^3}{s^3}$$

$$\text{Skewness} = \left(\frac{1}{5}\right) \frac{(-20)^3 + (-19)^3 + 22^3 + 24^3 + (-7)^3}{21.62^3}$$

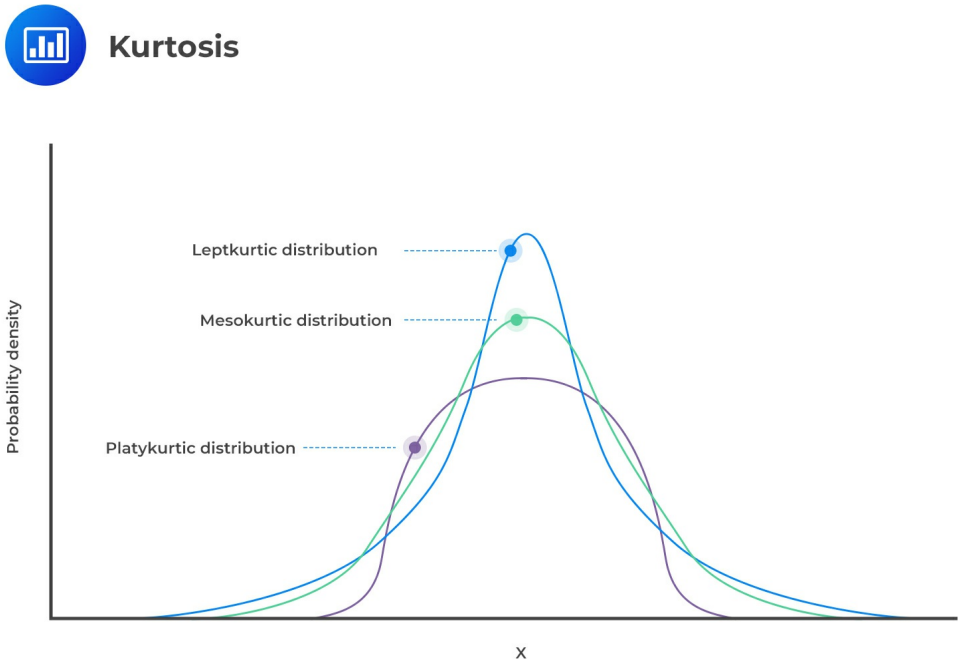
$$\text{Skewness} = 0.1835$$

Skewness is positive. Hence, the data has a positively skewed distribution.

Kurtosis

Kurtosis refers to the measurement of the degree to which a given distribution is more or less 'peaked' relative to the normal distribution. The concept of kurtosis is instrumental in decision-making. In this regard, we have three categories of distributions:

- Leptokurtic.
- Mesokurtic.
- Platykurtic.



Leptokurtic

A leptokurtic distribution is more peaked than the normal distribution. The higher peak results from the clustering of data points along the *x*-axis. The tails are also fatter than those of a normal distribution. The coefficient of kurtosis is usually more than 3.

The term "lepto" means thin or skinny. When analyzing historical returns, a leptokurtic distribution means that small changes are less frequent since historical values are clustered

around the mean. However, there are also large fluctuations represented by the fat tails.

Platykurtic

A platykurtic distribution has extremely dispersed points along the x-axis, resulting in a lower peak when compared to a normal distribution. "Platy" means broad. Hence, the prefix fits the distribution's shape, which is wide and flat. The points are less clustered around the mean compared to a leptokurtic distribution. The coefficient of kurtosis is usually less than 3.

Returns that follow this type of distribution have fewer major fluctuations compared to leptokurtic returns. However, you should note that fluctuations represent the riskiness of an asset. More fluctuations represent more risk and vice versa. Therefore, platykurtic returns are less risky than leptokurtic returns.

Mesokurtic

Lastly, mesokurtic distributions have a curve that is similar to that of a normal distribution. In other words, the distribution is mainly normal.

The majority of equity return series are found to have fat tails. Suppose a return distribution has fat tails, and we apply statistical models that do not consider distribution. In that case, we will overestimate the probability of either extremely poor or very favorable outcomes.

Investors often study a stock's daily trading volume distribution to assess its trading liquidity. It helps them see if the market can handle a large trade in that stock. This is useful for investors who want to make big investments or exit their positions in a particular stock.

Calculating Sample Kurtosis

Sample kurtosis is always measured relative to the kurtosis of a normal distribution, which is 3. Therefore, we are always interested in the "excess "kurtosis, i.e.,

$$\text{Excess kurtosis} = \text{Sample kurtosis} - 3$$

Where:

$$\text{Sample Excess Kurtosis} = \left(\frac{1}{n}\right)\frac{\sum_{i=1}^{n}\left(X_i - \bar{X}\right)^4}{s^4}$$

Positive excess kurtosis indicates a leptokurtic distribution. A zero value indicates a mesokurtic distribution. Lastly, a negative excess kurtosis represents a platykurtic distribution.

Example: Calculating Kurtosis

Using the data from the example above (12, 13, 54, 56, and 25), determine the type of kurtosis present.

$$\bar{X} = \frac{(12 + 13 + 54 + 56 + 25)}{5} = \frac{160}{5} = 32$$
$$s^2 = \frac{(12 - 32)^2 + (13 - 32)^2 + \ldots (25 - 32)^2}{4} = 467.5$$
$$s = \sqrt{467.5} = 21.62$$

Therefore,

$$\text{Excess Kurtosis} = \left(\frac{1}{n}\right)\frac{\sum_{i=1}^{n}\left(X_i - \bar{X}\right)^4}{s^4} - 3$$
$$\text{Excess Kurtosis} = \left(\frac{1}{5}\right)\frac{(-20)^4 + (-19)^4 + 22^4 + 24^4 + (-7)^4}{21.62^4} - 3$$
$$\text{Excess Kurtosis} = 2.2139$$

Since the excess kurtosis is negative, we have a platykurtic distribution.

Question 1

The skewness of the normal distribution is *most likely:*

A. Zero.
B. Positive.
C. Negative.

Solution

The correct answer is **A**.

Since the normal curve is symmetric about its mean, its skewness is zero.

B is incorrect because a positively skewed distribution has positive skewness.

C is incorrect because a negatively skewed distribution has negative skewness.

Question 2

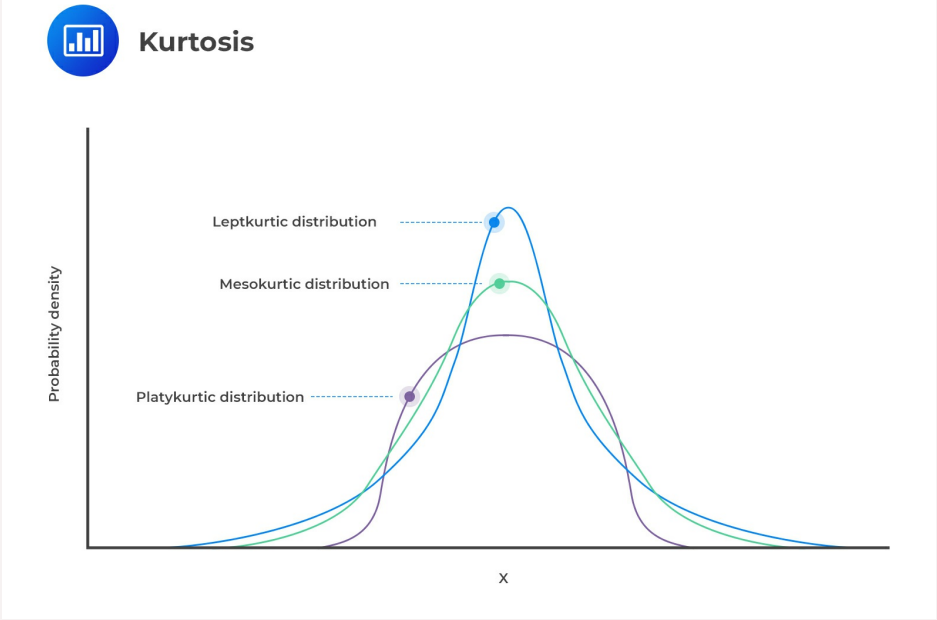
A frequency distribution in which there are too few scores at the extremes of the distribution is *most likely* called:

A. Platykurtic.
B. Leptokurtic.
C. Mesokurtic.

Solution

The correct answer is **A**.

A platykurtic distribution has "thin" tails and is flatter compared to a normal distribution. It implies that there are fewer scores at the extremes of the distribution, which aligns with the question's description.



Question 3

When most of the data are concentrated on the left of the distribution, it is *most likely* called:

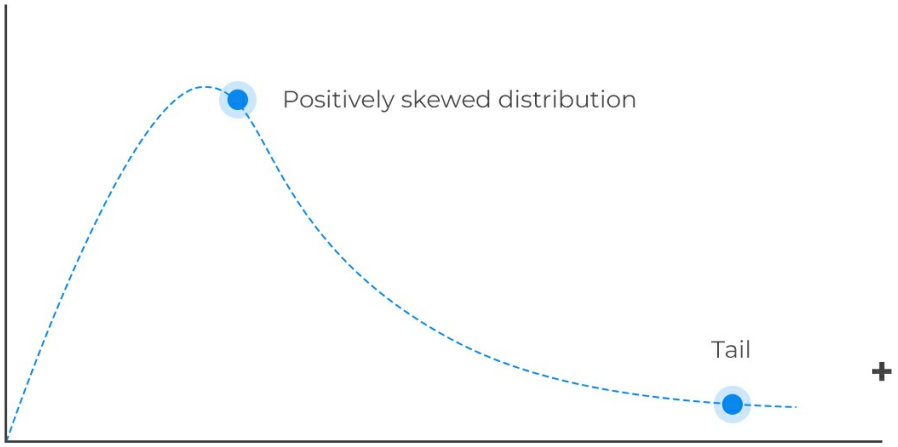
A. Symmetric distribution.
B. Positively skewed distribution.
C. Negatively skewed distribution.

Solution

The correct answer is **B**.

A distribution is said to be skewed to the right or positively skewed when most of the data are concentrated on the left of the distribution. A distribution is said to be skewed to the left or negatively skewed if most of the data are concentrated on the right of the distribution. The left tail clearly extends farther from the distribution's center than the right tail.

Positively Skewed Distribution



A is incorrect. A symmetric distribution is one in which the left and right sides mirror each other.

C is incorrect. A distribution is said to be skewed to the left or negatively skewed if most of the data are concentrated on the right of the distribution. The left tail extends farther away from the mean than the right tail.

LOS 3d: Interpret the correlation between two variables to address an investment problem

Covariance

Covariance is a measure of how two variables move together. The sample covariance of X and Y is calculated as follows:

$$s_{XY} = \frac{\sum_{i=1}^{N} (X_i - \bar{X})(Y_i - \bar{Y})}{n - 1}$$

The formula above implies that the sample covariance is the mean of the product of the deviations in the two random variables and from their sample means.

If the covariance between two random variables is positive, it means they move in the same direction. When one is below its mean, the other is also below its mean, and vice versa.

A major drawback of covariance is that it is difficult to interpret since its value can vary from negative infinity to positive infinity.

Correlation

Correlation is a standardized measure of the linear relationship between two variables. It takes the covariance and divides it by the product of the standard deviations of both variables. As a result, its value ranges between -1 and +1 and is easier to interpret.

The sample correlation coefficient is calculated as follows:

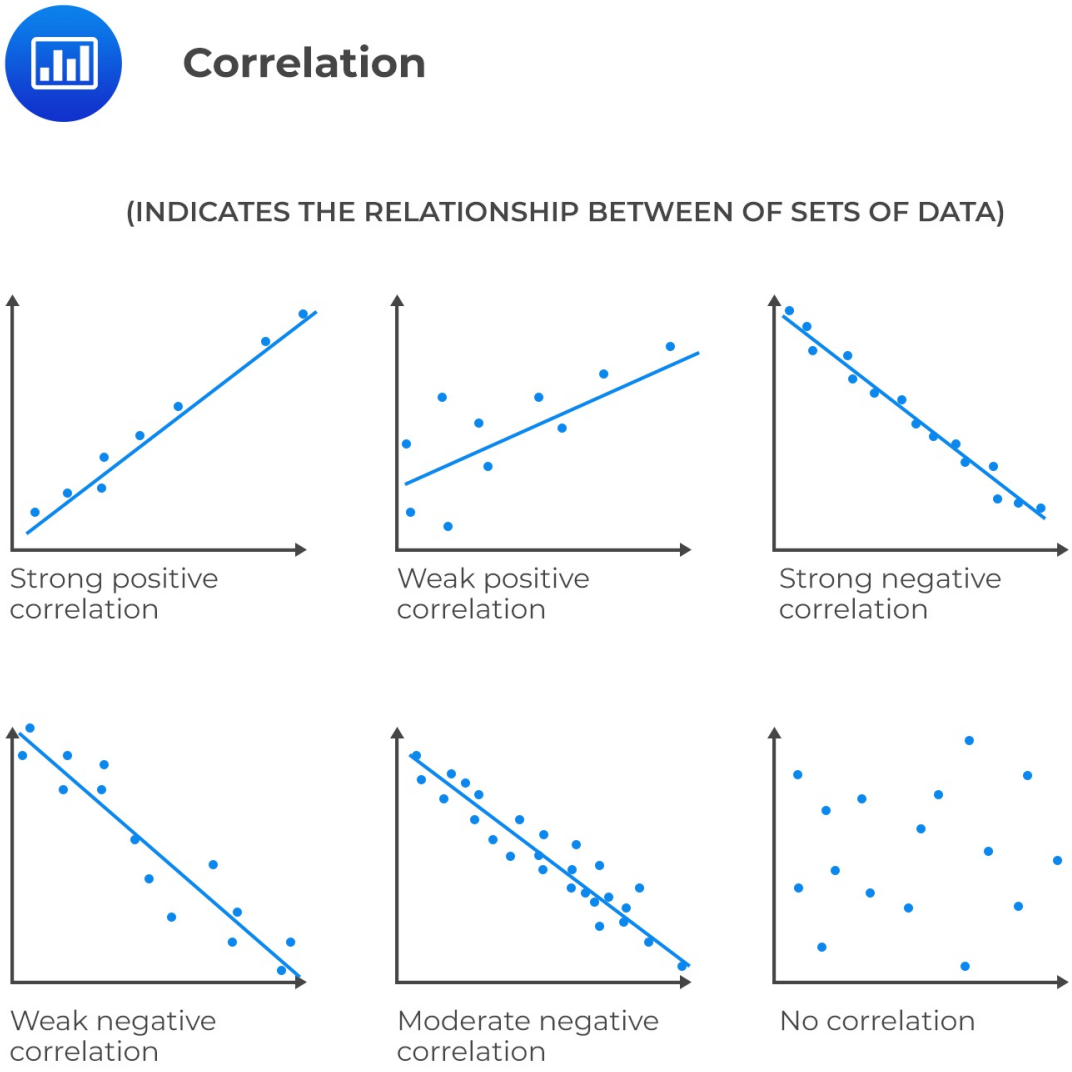
$$r_{xy} = \frac{S_{xy}}{S_x \times S_y}$$

Where:

s_{XY} = Covariance between variable X and Y.

s_X = Standard deviation of variable X.

s_Y = Standard deviation of variable Y.



Properties of Correlation

- Correlation ranges between -1 to +1 for two random variables, X and Y.
- A correlation of 0 (uncorrelated variables) indicates no linear (straight line) relationship exists between the variables.

- A positive correlation close to +1 indicates a strong positive linear relationship.
    - A correlation of 1 indicates a perfect linear relationship.
- A negative correlation close to -1 indicates a strong negative linear relationship.
    - A correlation of -1 indicates a perfect inverse linear relationship.

Limitations of Correlation Analysis

- Two variables can have a very low correlation despite having a strong ***nonlinear*** relationship.
- Correlation can be an unreliable measure when outliers are present in the data.
- Correlation does not imply causation. This implies that the correlation may be spurious. A spurious correlation refers to:
    - Correlation between two variables due to chance relationships in a particular dataset.
    - Correlation arising between variables when they are divided by a third variable.
    - Correlation between two variables arising from their relation to a third variable.

Question

The correlation coefficient between X and Y is 0.7, and the covariance is 29. If the variance of Y is 25, the variance of X is *closest* to:

A. 8.29.

B. 29.

C. 68.65.

Solution

The correct answer is **C**.

$$r_{XY} = \frac{s_{XY}}{s_X \times S_Y}$$
$$\Rightarrow 0.7 = \frac{29}{X \times 5}$$
$$\therefore X = 8.2857$$

$$\text{Variance} = 8.2857^2 = 68.65$$