LM03 Statistical Measures of Asset Returns

Ver 1.0

1. Introduction

In this learning module we will learn how to summarize and analyze important aspects of financial returns. This learning module covers:

- Measures of central tendency and location
- Measures of dispersion
- Measures of the shape of return distributions
- Covariance and correlation between two variables

2. Measures of Central Tendency and Location

A 'population' is defined as all members of a specified group. A 'parameter' describes the characteristics of a population.

A 'sample' is a subset drawn from a population. A 'sample statistic' describes the characteristic of a sample.

For example, all stocks listed on a country's exchange refers to a population. If 30 stocks are selected from the listed stocks, then this refers to a sample.

Sample statistics—such as measures of central tendency, measures of dispersion, skewness, and kurtosis—help make probabilistic statements about investment returns.

'Measures of central tendency' specify where data are centered.

'Measures of location' include not only measures of central tendency but other measures that explain the location or distribution of data.

Measures of Central Tendency

The Arithmetic Mean

The arithmetic mean is the sum of the observations divided by the number of observations. It is the most frequently used measure of the middle or center of data.

The Sample Mean

The sample mean is the arithmetic mean calculated for a sample. It is expressed as:

$$\overline{X} = \frac{\sum_{i=1}^{n} X_i}{n}$$

where: n is the number of observations in the sample.

If the sample data is: 2, 4, 4, 6, 10, 10, 12, 12, and 12 the sample mean can be calculated as:

$$\overline{X} = \frac{2 + 4 + 4 + 6 + 10 + 10 + 12 + 12 + 12}{9} = 8$$

A drawback of the arithmetic mean is that it is sensitive to extreme values (outliers). It can be pulled sharply upward or downward by extremely large or small observations,

respectively.

The Median

The median is the midpoint of a data set that has been sorted into ascending or descending order.

For odd number of observations: 2,5,7,11,14 → Median = 7

For even number of observations: 3, 9, 10, 20 → Median = (9 + 10)/2 = 9.5

As compared to a mean, a median is less affected by extreme values (outliers).

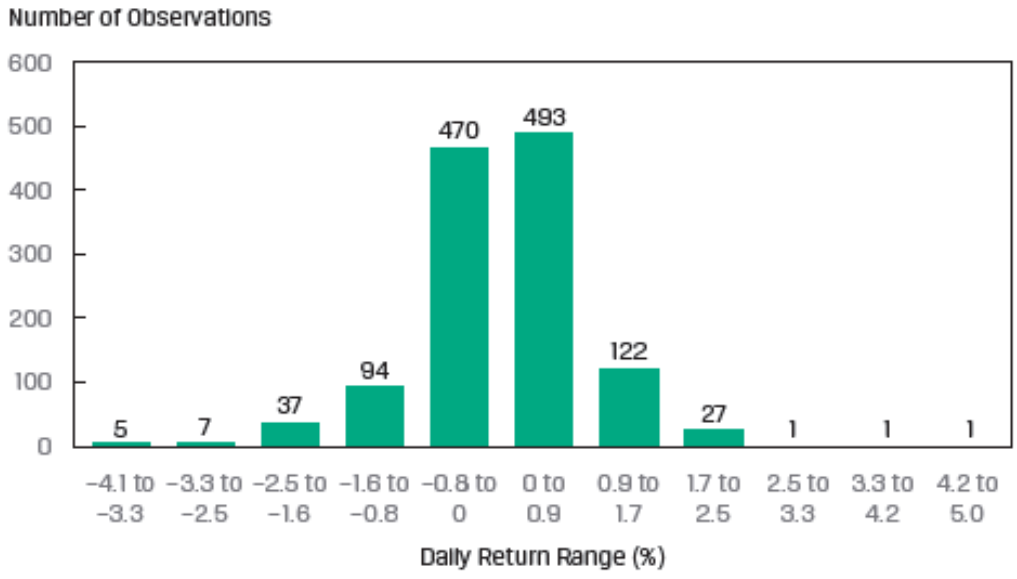
The Mode

The mode is the most frequently occurring value in a distribution.

For the following data set: 2, 4, 5, 5, 7, 8, 8, 8, 10, 12 → Mode = 8

A distribution can have more than one mode, or even no mode. When a distribution has one mode it is said to be unimodal. If a distribution has two or three modes, it is called bimodal or trimodal respectively.

When working with continuous data such as stock returns, 'modal interval' is often used instead of a mode. The data is divided into bins and the bin with the highest frequency is considered the modal interval. The following exhibit demonstrates this concept by plotting a histogram of the daily returns on an index. The highest bar in the histogram '0.0 to 0.9%' is the modal interval.



Dealing with Outliers

When data contains outliers, there are three options to deal with the extreme values:

Option 1: Do nothing; use the data without any adjustment.

Option 2: Delete all the outliers.

Option 3: Replace the outliers with another value.

Option 1 is appropriate in cases when the extreme values are genuine.

Option 2 excludes extreme observations. A **trimmed mean** excludes a stated percentage of the lowest and highest values and then calculates the arithmetic mean of the remaining values. For example, a 5% trimmed mean discards the lowest 2.5% and the highest 2.5% of values and computes the mean of the remaining 95% of values.

Option 3 replaces extreme observations with observations closest to them. A **winsorized mean** assigns a stated percentage of the lowest values equal to one specified low value and a stated percentage of the highest values equal to one specified high value, and then computes a mean from the restated data. For example, a 95% winsorized mean sets the bottom 2.5% of values equal to the value at or below which 2.5% of all the values lie (the "2.5th percentile" value) and the top 2.5% of values equal to the value at or below which 97.5% of all the values lie (the "97.5th percentile" value).

Measures of Location

Quartiles, Quintiles, Deciles, and Percentiles

A quantile is a value at or below which a stated fraction of the data lies. Some examples of quantiles include:

- Quartiles: The distribution is divided into quarters.
- Quintiles: The distribution is divided into fifths.
- Deciles: The distribution is divided into tenths.
- Percentile: The distribution is divided into hundredths.

The formula for the position of a percentile in a data set with n observations sorted in ascending order is:

$$L_y = \frac{(n + 1)y}{100}$$

where:
y is the percentage point at which we are dividing the distribution.
n is the number of observations.
L_y is the location (L) of the percentile (P_y) in an array sorted in ascending order.

Some important points to remember are:

- When the location, L_y, is a whole number, the location corresponds to an actual observation.
- When L_y is not a whole number or integer, L_y lies between the two closest integer numbers (one above and one below) and we use linear interpolation between those two places to determine P_y.
- Interquartile range is the difference between the third and the first quartiles.

Example
Consider the data set:

47 35 37 32 40 39 36 34 35 31 44

1. Find the 75th percentile point
2. Find the 1st quartile and 3rd quartile
3. Calculate the interquartile range
4. Find the 5th decile point
5. Find the 6th decile point.

Solution to 1:
First arrange the data in ascending order:

31, 32, 34, 35, 35, 36, 37, 39, 40, 44, 47

Location of the 75th percentile is the:
L_{75} = (11 + 1) (75/100) = 9th value. i.e. P_{75} = 40

With a small data set, such as this one, the location and the value is approximate. As the data set becomes larger, the location and percentile value estimates become more precise.

Solution to 2:
Location of the 1st quartile is:

L_{25} = (11 + 1) (25/100) = 3rd value. i.e. P_{25} = 34

Location of the 3rd quartile is:

L_{75} = (11 + 1) (75/100) = 9th value. i.e. P_{75} = 40

Solution to 3:
The interquartile range is the difference between the third and first quartiles, 40 – 34 = 6

Solution to 4:
Location of the 5th decile is:

L_{50} = (11 + 1) (50/100) = 6th value. i.e. P_{50} = 36

Solution to 5:
L_{60} = (11 + 1) (60/100) = 7.2

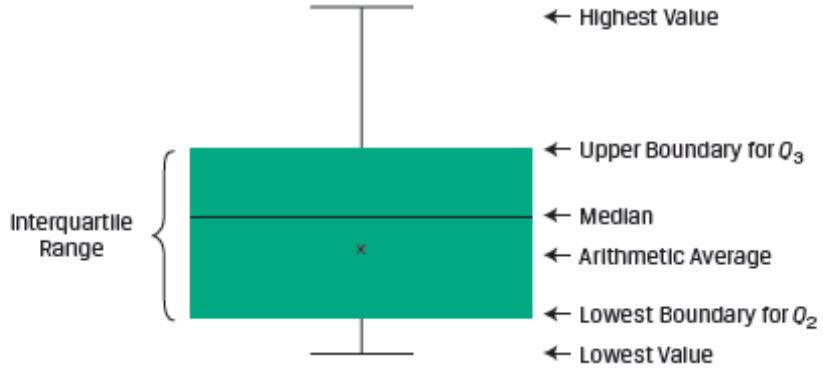
Use linear interpolation, which estimates an unknown value on the basis of two known values that surround it.

In this case, the 7th value is 37 and the 8th value is 39. The 6th decile is: P_{60} = 37+ 0.4 (0.2 times the linear distance between 37 and 39). P_{60} = 37.4

Box and Whiskers Plot

A box and whiskers plot is used to visualize the dispersion of data across quartiles. The box represents the interquartile range. The whiskers represent the highest and lowest values of

the distribution. Exhibit 5 shows a sample box and whisker plot.



There are several variations of the box and whiskers plot. Sometimes the whiskers may be a function of the interquartile range instead of the highest and lowest values.

Quantiles in Investment Practice

Quantiles are used in:

- Portfolio performance evaluation: The performance of investment managers is often evaluated in terms of the percentile or quartile in which they fall relative to the performance of their peers.
- Investment research: For example, companies can be ranked based on their market capitalization and sorted into deciles. The first decile contains companies with smallest market values and the tenth decile contains companies with the largest market values. Such a classification allows analysts to compare the performance of small companies with large ones.

3. Measures of Dispersion

Measures of central tendency tell us where the investment results (expected returns) are centered. However, to evaluate an investment we also need to know how returns are dispersed around the mean. Measures of dispersion describe the variability of outcomes around the mean.

The Range

The range is the difference between the maximum and minimum values in a data set. It is expressed as:

Range = Max value – Min Value

If the annual returns data is: 10%, -5%, 10%, 25%. What is the range?

Here the maximum return is 25% and the minimum return is -5%. The range is 25% – (-5%) = 30%.

Another way to specify the range is to mention the actual minimum and maximum values. For example, for the above data the range is “from -5% to 25%”.

The range is easy to compute; however, it does not tell us much about how the data is distributed.

Mean Absolute Deviations

It is the average of the absolute values of deviations from the mean. It is expressed as:

$$\text{MAD} = \left[\sum_{i=1}^{n} |X_i - \overline{X}|\right] / n$$

where: $\overline{X}$ is the sample mean and n is the number of observations in the sample.

Example

Consider the following data set: 8, 12, 10, 8 and 5. Calculate the mean absolute deviation.

Solution:

$\overline{X}$= (8 + 12 + 10 + 8 + 5) / 5 = 8.6

$$\text{MAD} = \frac{|8 - 8.6| + |12 - 8.6| + |10 - 8.6| + |8 - 8.6| + |5 - 8.6|}{5}$$

$$\text{MAD} = \frac{0.6 + 3.4 + 1.4 + 0.6 + 3.6}{5} = 1.92$$

Sample Variance and Sample Standard Deviation

'Variance' is defined as the average of the squared deviations around the mean. 'Standard deviation' is the positive square root of the variance.
'Sample variance' applies when we are dealing with a subset, or sample, of the total population. It is expressed as:

$$s^2 = \sum_{i=0}^{n} (X_i - \overline{X})^2 / (n - 1)$$

where: $\overline{X}$ is the sample mean and n is the number of observations in the sample.

'Sample standard deviation' is defined as the positive square root of the sample variance.

Example

Calculate the sample variance for the following data set: 8, 12, 10, 8 and 5.

Solution:

$$s^2 = \frac{[(8 - 8.6)^2 + (12 - 8.6)^2 + (10 - 8.6)^2 + (8 - 8.6)^2 + (5 - 8.6)^2]}{5 - 1}$$

$s^2 = 6.80\%$

The sample standard deviation is the positive square root of the sample variance. For the sample data given above, $s = \sqrt{6.80}$ = 2.61%

Using a financial calculator to calculate variance and standard deviations

The sample standard deviation can easily be computed using a financial calculator. Assume the following data set: 10%, -5%, 10%, 25%, the calculator key strokes are shown below:

Keystrokes	*Description*	*Display*
[2nd] [DATA]	Enters data entry mode	
[2nd] [CLR WRK]	Clears data register	X01
10 [ENTER]		X01 = 10
[↓] [↓] 5+/- [ENTER]		X02 = -5
[↓] [↓] 10 [ENTER]		X03 = 10
[↓] [↓] 25 [ENTER]		X04 = 25
[2nd] [STAT] [ENTER]	Puts calculator into stats mode	
[2nd] [SET]	Press repeatedly till you see →	1-V
[↓]	Number of data points	N = 4
[↓]	Mean	X = 10
[↓]	Sample standard deviation	Sx = 12.25
[↓]	Population standard deviation	σx = 10.61

Notice that the calculator gives both the sample and the population standard deviation. On the exam we will have to determine whether we are dealing with population or sample data and choose the appropriate value.

Downside Deviation and Coefficient of Variation

Variance and standard deviation of returns take account of returns above and below the mean, but often investors are concerned only with downside risk, for example returns below the mean.

The target downside deviation, or target semideviation, is a measure of the risk of being below a given target. It is calculated as the square root of the average squared deviations from the target, but it includes only those observations below the target (B).
The sample target semideivation can be calculated as:

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

Example:
Suppose the monthly returns on a portfolio are as shown:

Month	**Return (%)**
Jan	6
Feb	4
Mar	-2
Apr	-5

May	5
Jun	2
Jul	1
Aug	0
Sep	4
Oct	3
Nov	0
Dec	2

Calculate the target downside deviation when the target return is 4%.

Solution:

Month	**Observation**	**Deviation from the 4% target**	**Deviation below the target**	**Squared deviations below the target**
Jan	6	2	-	-
Feb	4	0	-	-
Mar	-2	-6	-6	36
Apr	-5	-9	-9	81
May	5	1	-	-
Jun	2	-2	-2	4
Jul	1	-3	-3	9
Aug	0	-4	-4	16
Sep	4	0	-	-
Oct	3	-1	-1	1
Nov	0	-4	-4	16
Dec	2	-2	-2	4
Sum				**167**

$$\text{Target semideviation} = \sqrt{\frac{167}{11}} = 3.8964\%$$

The target downside deviation will be less than the standard deviation, because deviations above the target are ignored. As the target is increased, the target downside deviation will increase.

Coefficient of Variation

Coefficient of variation expresses how much dispersion exists relative to the mean of a distribution and allows for direct comparison of dispersion across different data sets, even if the means are drastically different from one another. It is used in investment analysis to compare relative risks. When evaluating investments, a lower value is better. Coefficient of variation is expressed as:

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

where: s = sample standard deviation of a set of observations and $\overline{X}$ = sample mean

Example
Investment A has a mean return of 7% and a standard deviation of 5%. Investment B has a mean return of 12% and a standard deviation of 7%. Calculate the coefficients of variation.

Solution
The coefficients of variation can be calculated as follows:

$$CV_A = \frac{5\%}{7\%} = 0.71$$

$$CV_B = \frac{7\%}{12\%} = 0.58$$

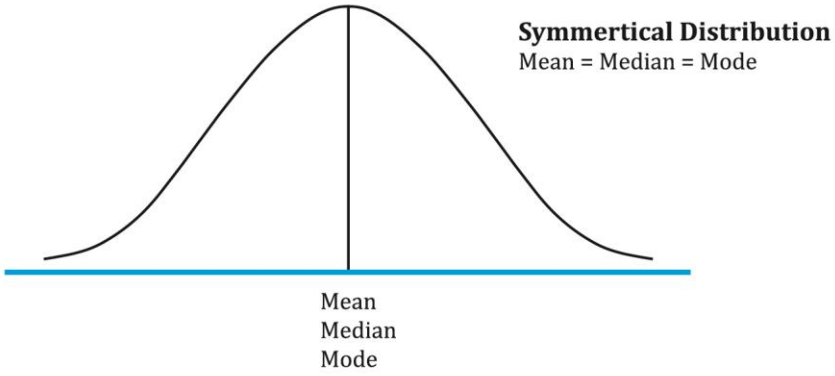
This metric shows that Investment A is riskier than Investment B.

4. Measures of Shape of a Distribution

Mean and variance may not adequately describe an investment's distribution of returns. To reveal other important characteristics of the distribution, we must look beyond measures of central tendency, location, and dispersion. One such characteristic is the degree of symmetry in return distributions.

Symmetrical distribution
A distribution is said to be symmetrical when the distribution on either side of the mean is a mirror image of the other.

In a normal distribution, mean = median = mode.

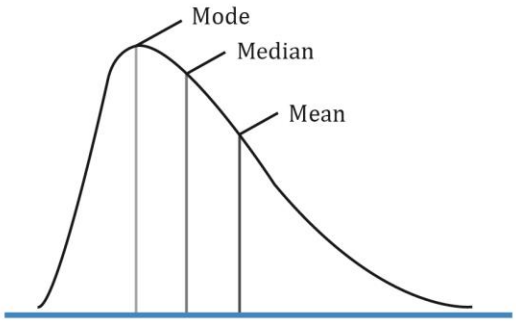


If a distribution is non-symmetrical, it is said to be skewed. Skewness is a measure of the asymmetry of the probability distribution. Skewness can be negative or positive.

Positively skewed distribution
A positively skewed distribution has a long tail on the right side, which means that there will be limited but frequent downside returns and unlimited but less frequent upside returns.

Here the mean > median > mode. The extreme values affect the mean the most which is

pulled to the right. They affect the mode the least.

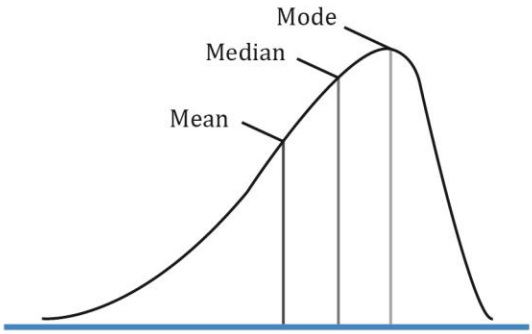


Right - Skewed (Positive Skewness)

Negatively skewed distribution

A negatively skewed distribution has a long tail on the left side, which means that there will be limited but frequent upside returns and unlimited but less frequent downside returns.

Here the mean < median< mode. The extreme values affect the mean the most which is pulled to the left. They affect the mode the least.



Left - Skewed (Negative Skewness)

Instructor's Note: Investors prefer positive skewness because it has a higher chance of very large returns and also because it has a higher mean return.

Example:

Which of the following distribution is most likely characterized by frequent small losses and a few extreme gains?

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

Solution:

C is correct. A positively skewed distribution is characterized by frequent small losses and a

few extreme gains.

Example:

Which of the following is most likely to be true for a negatively skewed distribution?

A. Mean < Median < Mode
B. Mode < Median < Mean
C. Median < Mean < Mode

Solution:

A is correct. In a negatively skewed distribution, the mean < median < mode.

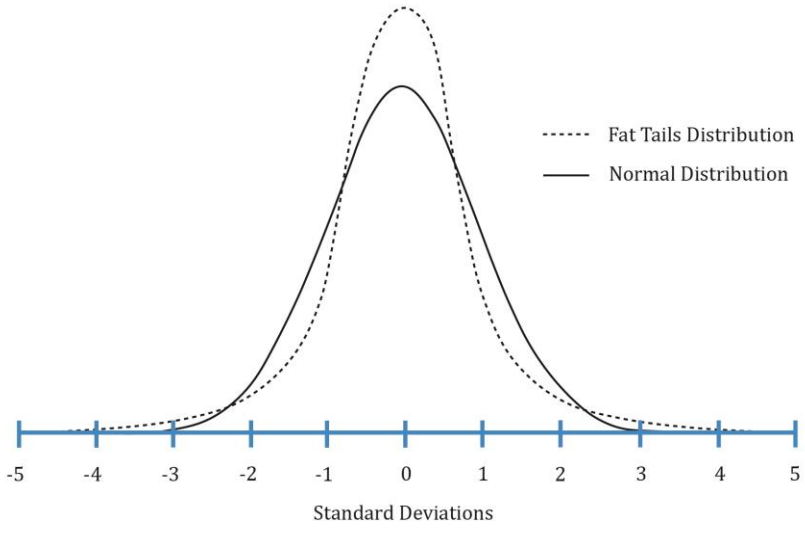
Kurtosis

Kurtosis is a measure of the combined weight of the tails of a distribution relative to the rest of the distribution.

Excess kurtosis = kurtosis - 3. An excess kurtosis with an absolute value greater than one is considered significant.

- A 'leptokurtic' distribution has fatter tails than a normal distribution. It has an excess kurtosis greater than 0.
- A 'platykurtic' distribution has thinner tails than a normal distribution. It has an excess kurtosis less than 0.
- A 'mesokurtic' distribution is identical to a normal distribution. It has an excess kurtosis equal to 0.

The following figure shows a leptokurtic distribution. As compared to a normal distribution, a leptokurtic distribution is more likely to generate observations in the tail region. It is also more likely to generate observations near the mean. However, to have the total probabilities sum to 1, it will generate fewer observations in the remaining regions (i.e. regions between the central and the two tail regions)

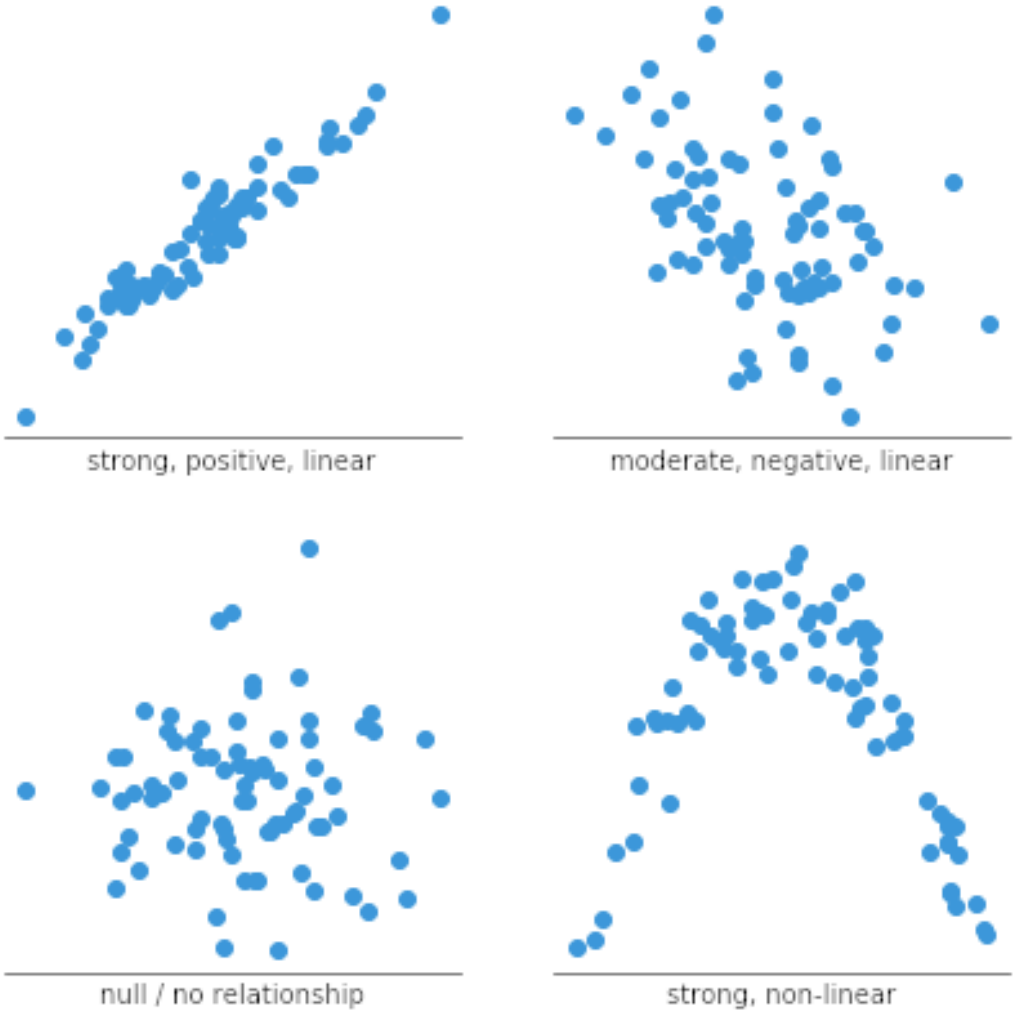


5. Correlation Between Two Variables

Scatter Plot

A scatter plot is a type of graph used to visualize the joint variation in two numerical variables. It is constructed with the x-axis representing one variable and the y-axis representing the other variable. Dots are drawn to indicate the values of the two variables at different points in time.

The pattern of a scatter plot may indicate no relationship, linear relationship or a non-linear relationship between the two variables. In case of a linear relationship, a positive slope indicates that the variables move in the same direction; whereas a negative slope indicates that the variables move in opposite directions.



Covariance and Correlation

Covariance

Covariance is a measure of how two variables move together. The formula for computing the

'sample covariance' of X and Y is:

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

The problem with covariance is that it can vary from negative infinity to positive infinity which makes it difficult to interpret. To address this problem, we use another measure called correlation.

Correlation

Correlation is a standardized measure of the linear relationship between two variables with values ranging between -1 and +1.

The 'sample correlation coefficient' can be calculated as:

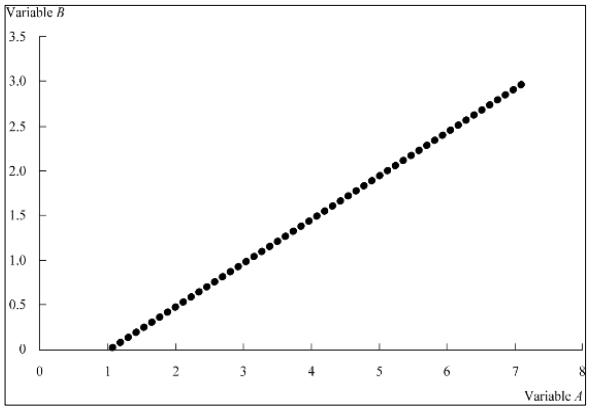
$$r_{XY} = \frac{s_{XY}}{s_x * s_y}$$

Properties of Correlation

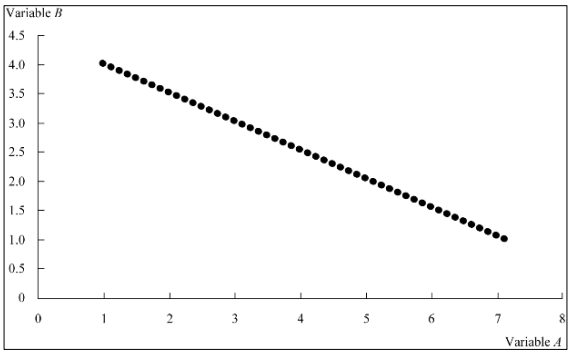
- Correlation ranges from -1 and +1.
- A correlation of 0 (uncorrelated variables) indicates an absence of any linear (straight-line) relationship between the variables.
- A correlation of +1 indicates a perfect positive relationship.
- A correlation of -1 indicates a perfect negative relationship.

The three scatter plots below show a positive linear, negative linear, and no linear relation between two variables A and B. They have correlation coefficients of +1, -1 and 0 respectively.

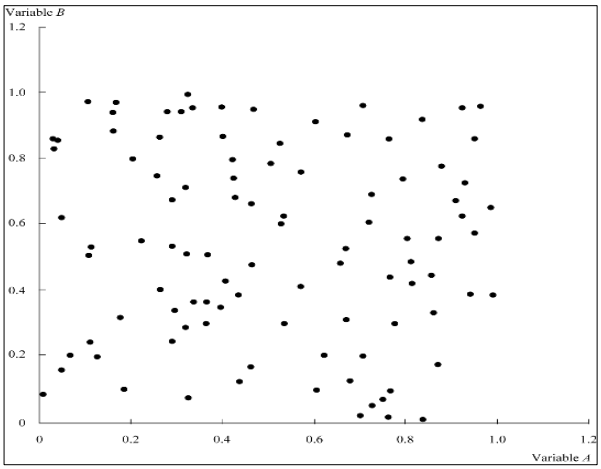
Variables with a correlation of 1.



Variables with a correlation of -1.



Variables with a correlation of 0.



Limitations of Correlation Analysis

The correlation analysis has certain limitations:

- Two variables can have a strong non-linear relation and still have a very low correlation.
- The correlation can be unreliable when outliers are present.
- The correlation may be spurious. 'Spurious correlation' refers to the following situations:
  - The correlation between two variables that reflects chance relationships in a particular data set.
  - The correlation induced by a calculation that mixes each of two variables with a third variable.
  - The correlation between two variables arising not from a direct relation between them, but from their relation to a third variable. Ex: shoe size and vocabulary of school children. The third variable is age here. Older shoe sizes simply imply that they belong to older children who have a better vocabulary.

Summary

LO: Calculate, interpret, and evaluate measures of central tendency and location to address an investment problem.

Measures of central tendency specify where data are centered.

The arithmetic mean is the sum of the observations divided by the number of observations. It is the most frequently used measure of the middle or center of data.

The median is the midpoint of a data set that has been sorted into ascending or descending order.

The mode is the most frequently occurring value in a distribution.

Measures of location

A quantile is a value at or below which a stated fraction of the data lies. Some examples of quantiles include:

- Quartiles: The distribution is divided into quarters.
- Quintiles: The distribution is divided into fifths.
- Deciles: The distribution is divided into tenths.
- Percentile: The distribution is divided into hundredths.

The formula for the position of a percentile in a data set with n observations sorted in ascending order is:

$$L_y = \frac{(n + 1)y}{100}$$

A box and whiskers plot is used to visualize the dispersion of data across quartiles. The box represents the interquartile range. The whiskers represent the highest and lowest values of the distribution. There are several variations of the box and whiskers plot. Sometimes the whiskers may be a function of the interquartile range instead of the highest and lowest values.

LO: Calculate, interpret, and evaluate measures of dispersion to address an investment problem.

Measures of dispersion describe the variability of outcomes around the mean.

The range is the difference between the maximum and minimum values in a data set.

MAD is the average of the absolute values of deviations from the mean.

Variance is defined as the average of the squared deviations around the mean. Standard deviation is the positive square root of the variance.

Coefficient of variation expresses how much dispersion exists relative to the mean of a distribution and allows for direct comparison of dispersion across different data sets, even if the means are drastically different from one another.

The target downside deviation, or target semideviation, is a measure of the risk of being below a given target. It is calculated as the square root of the average squared deviations from the target, but it includes only those observations below the target (B).

LO: Interpret and evaluate measures of skewness and kurtosis to address an investment problem.

Skewness is a measure of the asymmetry of the probability distribution. If a distribution is non-symmetrical, it is said to be skewed. Skewness can be negative or positive.

A positively skewed distribution has a long tail on the right side, which means that there will be frequent small losses and few large gains.

A negatively skewed distribution has a long tail on the left side, which means that there will be frequent small gains and few large losses.

Kurtosis is a measure of the combined weight of the tails of a distribution relative to the rest of the distribution.

Excess kurtosis = kurtosis - 3. An excess kurtosis with an absolute value greater than one is considered significant.

- A leptokurtic distribution has fatter tails than a normal distribution. It has an excess kurtosis greater than 0.
- A platykurtic distribution has thinner tails than a normal distribution. It has an excess kurtosis less than 0.
- A mesokurtic distribution is identical to a normal distribution. It has an excess kurtosis equal to 0.

LO: Interpret correlation between two variables to address an investment problem.

Correlation is a statistic that measures the degree to which two variables move in relation to each other.

- Correlation ranges from -1 and +1.
- A correlation of 0 (uncorrelated variables) indicates an absence of any linear (straight-line) relationship between the variables.
- A correlation of +1 indicates a perfect positive relationship.
- A correlation of -1 indicates a perfect negative relationship.

