# Statistical Measures of Asset Returns

## 2. MEASURES OF CENTRAL TENDENCY AND LOCATION

---

### MEASURES OF CENTRAL TENDENCY

#### The Arithmetic Mean

Analysts and portfolio managers often want one number that describes a representative possible outcome of an investment decision. The arithmetic mean is one of the most frequently used 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.

---

#### The Sample Mean

The sample mean is the arithmetic mean, or arithmetic average, computed for a sample.

**Sample Mean Formula.** The **sample mean** or average, $\bar{X}$ (read "X-bar"), is the arithmetic mean value of a sample:

$$\bar{X} = \frac{\sum_{i=1}^{n} X_i}{n}, \tag{1}$$

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

A property and potential drawback of the arithmetic mean is its sensitivity to extreme values, or outliers. Because all observations are used to compute the mean and are given equal weight (i.e., importance), the arithmetic mean can be pulled sharply upward or downward by extremely

large or small observations, respectively. The most common approach in this situation is to report the median, or middle value, in place of or in addition to the mean.

---

### The Median

A second important measure of central tendency is the median.

**Definition of Median.** The median is the value of the middle item of a dataset that has been sorted in ascending or descending order. In an odd-numbered sample of $n$ observations, the median is the value of the observation that occupies the $(n+1)/2$ position. In an even-numbered sample, we define the median as the mean of the values of the observations occupying the $n/2$ and $(n+2)/2$ positions (the two middle observations).

Whether we use the calculation for an even- or odd-numbered sample, an equal number of observations lie above and below the median. A distribution has only one median. A potential advantage of the median is that, unlike the mean, outliers do not affect it.

The median, however, does not use all the information about the size of the observations; it focuses only on the relative position of the ranked observations. Calculating the median may also be more complex. Mathematicians express this disadvantage by saying that the median is less mathematically tractable than the mean.

---

### The Mode

A third important measure of central tendency is the mode.

**Definition of Mode.** The mode is the most frequently occurring value in a dataset. A dataset can have more than one mode, or even no mode. When a dataset has a single value that is observed most frequently, its distribution is said to be **unimodal**. If a dataset has two most frequently occurring values, then it has two modes and its distribution is referred to as **bimodal**. When all the values in a dataset are different, the distribution has no mode because no value occurs more frequently than any other value.

Stock return data and other data from continuous distributions may not have a modal outcome. Exhibit 2 presents the frequency distribution of the daily returns for the EAA Equity Index over the past five years.

---

## Dealing with Outliers

In practice, although an extreme value or outlier in a financial dataset may represent a rare value in the population, it may also reflect an error in recording the value of an observation or an observation generated from a different population. After having checked and eliminated errors, we can address what to do with extreme values in the sample.

When dealing with a sample that has extreme values, there may be a possibility of transforming the variable (e.g., a log transformation) or of selecting another variable that achieves the same purpose. If, however, alternative model specifications or variable transformations are not possible, then three options exist for dealing with 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.

The first option is appropriate if the values are legitimate, correct observations, and it is important to reflect the whole of the sample distribution. Because outliers may contain meaningful information, excluding or altering these values may reduce valuable information. Further, because identifying a data point as extreme leaves it up to the judgment of the analyst, leaving in all observations eliminates the need to judge a value as extreme.

The second option excludes the extreme observations. One measure of central tendency in this case is the **trimmed mean**, which is computing an arithmetic mean after excluding a stated small percentage of the lowest and highest values. For example, a 5 percent trimmed mean discards the lowest 2.5 percent and the highest 2.5 percent of values and computes the mean of the remaining 95 percent of values. A trimmed mean is used in sports competitions when judges' lowest and highest scores are discarded in computing a contestant's score.

The third option involves substituting values for the extreme values. A measure of central tendency in this case is the **winsorized mean**. It is calculated after assigning one specified low value to a stated percentage of the lowest values in the dataset and one specified high value to a stated percentage of the highest values in the dataset. For example, a 95 percent winsorized mean sets the bottom 2.5 percent of values in the dataset equal to the value at or below which 2.5 percent of all the values lie (as will be seen shortly, this is called the "2.5th percentile" value) and the top 2.5 percent of values in the dataset equal to the value at or below which 97.5 percent of all the values lie (the "97.5th percentile" value).

Often comparing the statistical measures of datasets with outliers included and with outliers excluded can reveal important insights about the dataset. Such comparison can be particularly helpful when investors analyze the behavior of asset returns and rate, price, spread and volume changes.

In Example 1, we show the differences among these options for handling outliers using daily returns for the fictitious Euro-Asia-Africa (EAA) Equity Index in Exhibit 2.

### Measures of Location

Having discussed measures of central tendency, we now examine an approach to describing the location of data that involves identifying values at or below which specified proportions of the data lie. For example, establishing that 25 percent, 50 percent, and 75 percent of the annual returns on a portfolio provides concise information about the distribution of portfolio returns. Statisticians use the word **quantile** as the most general term for a value at or below which a stated fraction of the data lies. In the following section, we describe the most commonly used quantiles—quartiles, quintiles, deciles, and percentiles—and their application in investments.

#### Quartiles, Quintiles, Deciles, and Percentiles

We know that the median divides a distribution of data in half. We can define other dividing lines that split the distribution into smaller sizes. **Quartiles** divide the distribution into quarters, **quintiles** into fifths, **deciles** into tenths, and **percentiles** into hundredths. The **interquartile range (IQR)** is the difference between the third quartile and the first quartile, or $IQR = Q_3 - Q_1$.

Example 2 illustrates the calculation of various quantiles for the daily return on the EAA Equity Index.

## 3. MEASURES OF DISPERSION

Few would disagree with the importance of expected return or mean return in invest ments: To understand an investment more completely, however, we also need to know how returns are dispersed around the mean. Dispersion is the variability around the central tendency. If mean return addresses reward, then dispersion addresses risk and uncertainty.

In this lesson, we examine the most common measures of dispersion: range, mean absolute deviation, variance, and standard deviation. These are all measures of absolute dispersion. **Absolute dispersion** is the amount of variability present without comparison to any reference point or benchmark

## The Range

We encountered range earlier when we discussed the construction of frequency distributions. It is the simplest of all the measures of dispersion.

In this lesson, we examine the most common measures of dispersion: range, mean absolute deviation, variance, and standard deviation. These are all measures of absolute dispersion. Absolute dispersion is the amount of variability present without comparison to any reference point or benchmark.

**Definition of Range.** The **range** is the difference between the maximum and minimum values in a dataset:

$$\text{Range} = \text{Maximum value} - \text{Minimum value} \tag{2}$$

An alternative way to report the range is to specify both the maximum and minimum values. This alternative definition provides more information as the range is reported as "from Maximum Value to Minimum Value."

One advantage of the range is ease of computation. A disadvantage is that the range uses only two pieces of information from the distribution. It cannot tell us how the data are distributed (i.e., the shape of the distribution). Because the range is the difference between the maximum and minimum values in the dataset, it is also sensitive to extremely large or small observations ("outliers") that may not be representative of the distribution.

---

## Mean Absolute Deviations

Measures of dispersion can be computed using all the observations in the distribution rather than just the highest and lowest. We could compute measures of dispersion as the arithmetic average of the deviations around the mean, but the problem is that deviations around the mean always sum to 0. Therefore, we need to find a way to address the problem of negative deviations canceling out positive deviations.

One solution is to examine the absolute deviations around the mean as in the **mean absolute deviation** (MAD).

**MAD Formula.** The MAD for a sample is:

$$\text{MAD} = \frac{\sum_{i=1}^{n} |X_i - \bar{X}|}{n} \tag{3}$$

where $\bar{X}$ is the sample mean, $n$ is the number of observations in the sample, and the $||$ indicate the absolute value of what is contained within these bars.

The MAD uses all of the observations in the sample and is thus superior to the range as a measure of dispersion. One technical drawback of MAD is that it is difficult to manipulate mathematically compared with the next measure we will introduce, sample variance.

---

## Sample Variance and Sample Standard Deviation

A second approach to the problem of positive and negative deviations canceling out is to square them. Variance and standard deviation, which are based on squared deviations, are the two most widely used measures of dispersion. **Variance** is defined as the average of the squared deviations around the mean. **Standard deviation** is the square root of the variance.

---

### Sample Variance

In investments, we often do not know the mean of a population of interest, so we estimate it using the mean from a sample drawn from the population. The corresponding measure of dispersion is the sample variance or standard deviation.

**Sample Variance Formula.** The **sample variance**, $s^2$, is:

$$s^2 = \frac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{n-1}, \tag{4}$$

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

The variance calculation takes care of the problem of negative deviations from the mean canceling out positive deviations by squaring those deviations.

For the sample variance, by dividing by the sample size minus 1 (or $n-1$) rather than $n$, we improve the statistical properties of the sample variance. The quantity $n-1$ is also known as the number of degrees of freedom in estimating the population variance. To estimate the population variance with $s^2$, we must first calculate the sample mean, which itself is an estimated parameter. Therefore, once we have computed the sample mean, there are only $n-1$ independent pieces of information from the sample; that is, if you know the sample mean and $n-1$ of the observations, you could calculate the missing sample observation.

### Sample Standard Deviation

Variance is measured in squared units associated with the mean, and we need a way to return to those original units. Standard deviation, the square root of the variance, solves this problem and is more easily interpreted than the variance.

A useful property of the sample standard deviation is that, unlike sample variance, it is expressed in the same unit as the data itself. If the dataset is percentage of daily returns for an index, then both the average and the standard deviation of the dataset is in percentage terms, while the variance is in squared percentage of daily returns.

**Sample Standard Deviation Formula.** The **sample standard deviation**, $s$, is:

$$s = \sqrt{\frac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{n-1}}, \tag{5}$$

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

Because the standard deviation is a measure of dispersion about the arithmetic mean, we usually present the arithmetic mean and standard deviation together when summarizing data. When we are dealing with data that represent a time series of percentage changes, presenting the geometric mean (i.e., representing the compound rate of growth) is also very helpful.

## Downside Deviation and Coefficient of Variation

An asset's variance or standard deviation of returns is often interpreted as a measure of the asset's risk. Variance and standard deviation of returns take account of returns above and below the mean, or upside and downside risks, respectively. However, investors are typically concerned only with **downside risk**—for example, returns below the mean or below some specified minimum target return. As a result, analysts have developed measures of downside risk.

---

### Downside Deviation

In practice, we may be concerned with values of return (or another variable) below some level other than the mean. For example, if our return objective is 6.0 percent annually (our minimum acceptable return), then we may be concerned particularly with returns below 6.0 percent a year. The target downside deviation, also referred to as the **target semideviation**, is a measure of dispersion of the observations (here, returns) below a target, for example 6.0 percent. To calculate a sample target semideviation, we first specify the target. After identifying observations below the target, we find the sum of those squared negative deviations from the target, divide that sum by the total number of observations in the sample minus 1, and finally, take the square root.

**Sample Target Semideviation Formula.** The target semideviation, $S_{Target}$, is:

$$S_{\text{target}} = \sqrt{\sum_{\text{for all} X_i \leq B}^{n} \frac{(X_i - B)^2}{n-1}} \tag{6}$$

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

We illustrate this in Example 3.

---

**Coefficient of Variation**

**Coefficient of Variation Formula.** The **coefficient of variation** (CV) is the ratio of the standard deviation of a set of observations to their mean value:

$$CV = \frac{s}{\bar{X}}, \tag{7}$$

where $s$ is the sample standard deviation and $\bar{X}$ is the sample mean.

---

## 4. MEASURES OF SHAPE OF A DISTRIBUTION

Mean and variance may not adequately describe an investment's distribution of returns. In calculations of variance, for example, the deviations around the mean are squared, so we do not know whether large deviations are likely to be positive or negative. We need to go beyond measures of central tendency, location, and dispersion to reveal other important characteristics of the distribution. One important characteristic of interest to analysts is the degree of symmetry in return distributions.

---

**sample skewness**

The approximation for computing **sample skewness** when $n$ is large (100 or more) is:

$$\text{Skewness} \approx \left(\frac{1}{n}\right) \frac{\sum_{i=1}^{n}(X_i - \bar{X})^3}{s^3}. \tag{8}$$

---

**Excess kurtosis**

The calculation for kurtosis involves finding the average of deviations from the mean raised to the fourth power and then standardizing that average by dividing by the standard deviation raised to the fourth power. A normal distribution has kurtosis of 3.0, so a fat-tailed distribution has a kurtosis above 3.0 and a thin-tailed distribution has a kurtosis below 3.0.

**Excess kurtosis** is the kurtosis relative to the normal distribution. For a large sample size ($n = 100$ or more), **sample excess kurtosis** ($K_E$) is approximately as follows:

$$K_E \approx \left(\frac{1}{n}\right) \frac{\sum_{i=1}^{n}(X_i - \bar{X})^4}{s^4} - 3. \tag{9}$$

As with skewness, this measure is free of scale. Many statistical packages report estimates of sample excess kurtosis, labeling this simply "kurtosis."

Excess kurtosis thus characterizes kurtosis relative to the normal distribution. A normal distribution has excess kurtosis equal to 0. A fat-tailed distribution has excess kurtosis greater than 0, and a thin-tailed distribution has excess kurtosis less than 0. A return distribution with positive excess kurtosis, a fat-tailed return distribution, has more frequent extremely large deviations from the mean than a normal distribution.

---

# 5. CORRELATION BETWEEN TWO VARIABLES

## Scatter Plot

A scatter plot is a useful tool for displaying and understanding potential relationships between two variables. Suppose an analyst is interested in the relative performance of two sectors, information technology (IT) and utilities, compared to the market index over a specific five-year period. The analyst has obtained the sector and market index returns for each month over the five years under investigation. Exhibit 24 presents a scatterplot of returns for the IT sector index versus the S&P 500, and Exhibit 25 presents a scatterplot of returns for the utilities sector index versus the S&P 500.

---

## Covariance and Correlation

---

### Correlation

**Correlation** is a measure of the linear relationship between two random variables. The first step in considering how two variables vary together, however, is constructing their covariance.

---

**Definition of Sample Covariance.**

The **sample covariance** ($s_{XY}$) is a measure of how two variables in a sample move together:

$$s_{XY} = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{n-1}. \tag{10}$$

Equation 10 indicates that the sample covariance is the average value of the product of the deviations of observations on two random variables ($X_i$ and $Y_i$) from their sample means. If the random variables are returns, the units would be returns squared. Also, note the use of $n-1$ in the denominator, which ensures that the sample covariance is an unbiased estimate of population covariance.

Stated simply, covariance is a measure of the joint variability of two random variables. If the random variables vary in the same direction, for example, $X$ tends to be above its mean when $Y$ is above its mean, and $X$ tends to be below its mean when $Y$ is below its mean, then their covariance is positive. If the variables vary in the opposite direction relative to their respective means, then their covariance is negative.

The size of the covariance measure alone is difficult to interpret as it involves squared units of measure and so depends on the magnitude of the variables. This brings us to the normalized version of covariance, which is the correlation coefficient.

---

**Definition of Sample Correlation Coefficient.**

The **sample correlation coefficient** is a standardized measure of how two variables in a sample move together. The sample correlation coefficient ($r_{XY}$) is the ratio of the sample covariance to the product of the two variables' standard deviations:

$$r_{XY} = \frac{s_{XY}}{s_X s_Y}. \tag{11}$$

Importantly, the correlation coefficient expresses the strength of the linear relationship between the two random variables.

## Properties of Correlation

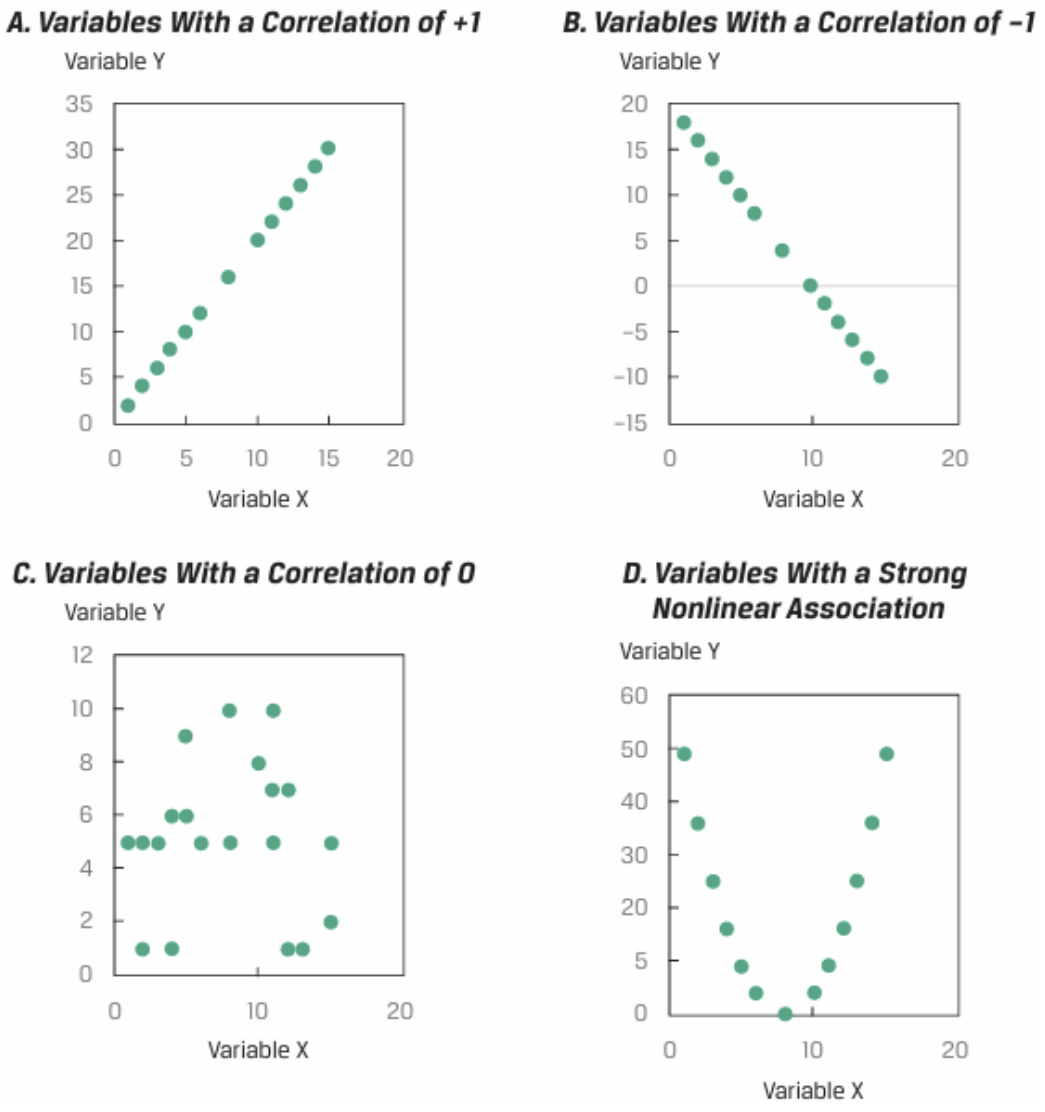
We now discuss the correlation coefficient, or simply correlation, and its properties in more detail:

1. Correlation ranges from $-1$ and $+1$ for two random variables, $X$ and $Y$:

$$-1 \leq r_{XY} \leq +1.$$

2. A correlation of 0, termed uncorrelated, indicates an absence of any linear relationship between the variables.
3. A positive correlation close to $+1$ indicates a strong positive linear relationship. A correlation of 1 indicates a perfect linear relationship.
4. A negative correlation close to $-1$ indicates a strong negative (i.e., inverse) linear relationship. A correlation of $-1$ indicates a perfect inverse linear relationship.

**Exhibit 26: Scatter Plots Showing Various Degrees of Correlation**




Panel C shows a scatter plot of two variables with a correlation of 0; they have no linear relation. This graph shows that the value of variable X tells us nothing about the value of variable Y. Panel D shows a scatter plot of two variables that have a non-linear relationship. Because the correlation coefficient is a measure of the linear association between two variables, it would not be appropriate to use the correlation coefficient in this case.

---

## Limitations of Correlation Analysis

Exhibit 26 illustrates that correlation measures the linear association between two variables, but it may not always be reliable. Two variables can have a strong nonlinear relation and still have a very low correlation. A nonlinear relation between variables X and Y is shown in Panel D. Even though these two variables are perfectly associated, there is no linear association between them and hence no meaningful correlation.

Moreover, with visualizations too, including scatter plots, we must be on guard against unconsciously making judgments about causal relationships that may or may not be supported by the data.

---

**spurious correlation**

The term **spurious correlation** has been used to refer to:

- correlation between two variables that reflects chance relationships in a particular dataset;
- correlation induced by a calculation that mixes each of two variables with a third variable; and
- correlation between two variables arising not from a direct relation between them but from their relation to a third variable.

---