Understanding Covariance in Excel: A Comparison of COVARIANCE.P and COVARIANCE.S

Authored by

Mohammed looti

October 27, 2025

Understanding Covariance: Quantifying the Relationship Between Variables

In the expansive field of statistics, **covariance** stands as a foundational measurement tool. It is specifically designed to quantify the degree and direction in which two distinct variables move together. By providing insight into this directional relationship, covariance helps analysts determine whether variables tend to increase or decrease in unison, or if their movements are inversely related.

The resulting value of the calculation offers immediate insight into the relationship's nature. A **positive covariance** indicates a direct relationship: as the magnitude of one variable increases, the other variable tends to increase as well. Conversely, a **negative covariance** signals an inverse relationship, where an upward movement in one variable is typically associated with a decline in the other. Crucially, a covariance value close to zero suggests a weak or non-existent linear relationship, though it is important to remember that this metric does not rule out potential non-linear dependencies between the data sets.

While covariance is essential for understanding the direction of association, its raw magnitude can be difficult to interpret practically because it is inherently dependent upon the measurement units of the variables involved. To overcome this limitation and achieve a standardized measure of relationship strength, analysts frequently rely on correlation. Correlation is essentially a normalized version of covariance, scaled to fall between -1 and +1, making the strength of the linear association universally comparable regardless of the variables' original units.

Population vs. Sample: The Crucial Statistical Distinction

Before executing any covariance calculation in Excel, a clear understanding of the difference between a population and a sample is paramount. A **population** is defined as the entire comprehensive group of individuals, objects, or data points that are the subject of the statistical investigation. For instance, if a study aims to analyze the performance of every single server operated by a large tech company, the entire fleet of servers constitutes the population.

In most real-world analytical scenarios, however, collecting data from an entire population is either logistically impossible or economically prohibitive. Consequently, researchers typically resort to gathering data from a **sample**, which is merely a representative subset of the larger population. The underlying objective when using a sample is to leverage the statistical findings derived from this smaller group to make robust inferences and draw accurate conclusions about the characteristics of the entire population from which it was selected.

The choice between calculating population covariance or sample covariance is determined solely by the scope of your available data. If your dataset encompasses every element of interest, you have a population; if it includes only a portion, you have a sample. This distinction is far more than

a conceptual detail; it directly influences the statistical formulas used, leading to differences in the denominator that adjust for the certainty (or uncertainty) associated with the data source.

COVARIANCE.P: Calculating Covariance for the Entire Population

Microsoft Excel provides the dedicated function **COVARIANCE.P** specifically for scenarios where the input data represents the **entire population** being studied. This function assumes that data collection has been comprehensive and exhaustive, meaning there is no missing information or need to estimate parameters for a larger, unobserved group. It is the appropriate choice when absolute certainty exists regarding the completeness of the dataset.

The underlying calculation in **COVARIANCE.P** follows the standard population covariance formula. This approach involves summing the products of the deviations of each variable from its respective mean and then dividing this sum by the total number of observations, denoted as 'n'. Since the entire population is known, no statistical correction is required to estimate population parameters, simplifying the denominator.

The mathematical representation used by the **COVARIANCE.P** function is provided below:

Population covariance = $\Sigma(x_i-x)(y_i-y) / n$

Where the components of the formula signify the following:

Σ: The standard symbol for "summation," directing the addition of all calculated values across the data range.
xi: Represents the ith individual observation for the variable x.
x: Represents the mean (average) value for variable x, calculated across the entire population dataset.
yi: Represents the ith individual observation for the variable y.
y: Represents the mean (average) value for variable y, calculated across the entire population dataset.
n: Represents the total count of paired observations or data points constituting the complete population.

Analysts must exercise caution and only employ **COVARIANCE.P** when they are unequivocally certain that their data set constitutes the full population. Applying this function incorrectly to a sample dataset will result in a biased calculation that consistently underestimates the true population covariance.

COVARIANCE.S: The Standard Function for Sample Estimation

In contrast to the population function, Excel's **COVARIANCE.S** is the function tailored for situations

where the data represents a **sample** extracted from a larger, unobserved population. As working with samples is the standard practice in most applied statistical analysis, **COVARIANCE.S** is generally the more frequently used function for covariance calculations.

When working with a sample, the primary statistical objective is to produce an accurate estimate of the population's characteristics. To ensure that the sample covariance serves as an unbiased estimator of the true population covariance, a necessary adjustment is implemented in the formula's denominator. Instead of simply dividing by 'n' (the sample size), the division is performed by 'n-1'. This critical correction is known in statistical theory as Bessel's correction.

The mathematical structure utilized by the **COVARIANCE.S** function is detailed below:

Sample covariance = $\Sigma(xi-x)(yi-y) / (n-1)$

The terms within the formula are defined as follows:

Σ: Denotes the summation of all products within the specified data range.
xi: Represents the ith individual observation for variable x within the sample.
x: Represents the mean (average) value for variable x, calculated only across the sample data.
yi: Represents the ith individual observation for variable y within the sample.
y: Represents the mean (average) value for variable y within the sample.
n: Represents the total count of paired observations or data points in the sample.

Dividing by 'n-1' effectively accounts for the statistical degrees of freedom lost when population parameters (like the mean) are estimated using sample data. This adjustment systematically inflates the resulting covariance slightly, ensuring that the calculation is a statistically sound and unbiased estimate of the underlying population parameter.

The Impact of Denominators: N vs. N-1

The fundamental and most impactful difference between **COVARIANCE.P** and **COVARIANCE.S** resides exclusively in their denominators: **COVARIANCE.P** uses the total number of observations (**n**), while **COVARIANCE.S** uses the sample size minus one (**n-1**). This minor mathematical tweak is the source of significant statistical consequences in estimation and inference.

The introduction of **n-1** in **COVARIANCE.S** is the implementation of Bessel's correction, a vital step when dealing with partial data. When population parameters are unknown and estimated from a subset (a sample), the sample variance and covariance tend to systematically underestimate the true population values if divided by 'n'. By dividing by 'n-1', the calculation increases slightly, providing an unbiased estimator--meaning that if we took infinite samples, the average of their calculated covariances would accurately equal the true population covariance.

A direct, observable consequence of this methodology is that **COVARIANCE.S** will invariably produce a larger numerical result than **COVARIANCE.P** when applied to the identical dataset (assuming n is greater than 1). While this difference diminishes for extremely large data sets (where n and n-1 are nearly identical), the distinction is highly relevant and statistically critical when analyzing small or moderately sized samples. Misapplying **COVARIANCE.P** to a sample guarantees a downwardly biased result, potentially leading to inaccurate statistical modeling or flawed conclusions.

Illustrative Example: Comparing Excel Outputs

To crystallize the theoretical difference, let us examine a practical application using Excel. Consider a scenario where we have gathered data representing 15 observations of two correlated variables--say, a set of basketball players' total points scored and their total assists over a season. We wish to calculate the covariance between these metrics.

The data is structured in an Excel sheet as follows, with the paired observations clearly visible:

	A	B	C	D	E	F
1	**Points**	**Assists**				
2	7	2				
3	7	4				
4	8	4				
5	10	3				
6	12	2				
7	14	3				
8	14	2				
9	15	4				
10	17	6				
11	18	5				
12	22	4				
13	22	5				
14	24	7				
15	29	11				
16	32	8				
17						
18						
19						
20						
21						
22						

Applying both functions to this data set reveals the effect of the denominator adjustment. Assuming

'Points' are in cells A2:A16 and 'Assists' are in B2:B16, the formulas would be `=COVARIANCE.S(A2:A16, B2:B16)` and `=COVARIANCE.P(A2:A16, B2:B16)`. The resulting values are displayed in the screenshot below:

	A	B	C	D	E	F
1	**Points**	**Assists**				Formula
2	7	2		**COVARIANCE.P**	14.64444	=COVARIANCE.P(A2:A16, B2:B16)
3	7	4		**COVARIANCE.S**	15.69048	=COVARIANCE.S(A2:A16, B2:B16)
4	8	4				
5	10	3				
6	12	2				
7	14	3				
8	14	2				
9	15	4				
10	17	6				
11	18	5				
12	22	4				
13	22	5				
14	24	7				
15	29	11				
16	32	8				
17						
18						
19						
20						
21						
22						
23						

As demonstrated, the sample covariance (**COVARIANCE.S**) is calculated as approximately **15.69**, while the corresponding population covariance (**COVARIANCE.P**) yields approximately **14.64**. The sample covariance is indeed larger, which is the expected outcome due to the division by n-1. Since both results are positive, we can conclude there is a positive linear relationship, indicating that players who score more points also tend to record more assists.

Selecting the Correct Function for Your Statistical Analysis

The decision of whether to use **COVARIANCE.P** or **COVARIANCE.S** is entirely predicated on the methodology of your data collection. In the vast majority of practical data analysis scenarios, researchers and business analysts are operating with samples rather than complete populations. Limitations in time, budget, and access make it generally impractical or impossible to observe every single element of the group of interest.

For this reason, **COVARIANCE.S** is the standard, default function for calculating covariance in Excel. It is mathematically engineered to provide an unbiased estimate of the true population covariance when you are working with a partial, representative subset of the data. When the goal is to generalize findings from your observed data to the larger universe it represents, **COVARIANCE.S** is the statistically correct choice.

Conversely, **COVARIANCE.P** is reserved for highly specific and rare circumstances where the analyst has absolute confirmation that the dataset includes every possible member of the group under investigation. An example might include calculating the covariance between sales and marketing spend for all five regional branches of a specific small business. If you are certain you possess the entire population data, then **COVARIANCE.P** is correct; otherwise, defaulting to **COVARIANCE.S** is the soundest practice in applied statistics.

Additional Resources for Statistical Analysis

To further enhance your understanding of statistical concepts and Excel functions, consider exploring the following tutorials that delve into the nuances of other commonly used functions:

QUARTILE.EXC vs. QUARTILE.INC in Excel: What's the Difference?