Covariance – Population vs. Sample Data

Population

A population includes **all** individual data points within a dataset.

Covariance formula (population):

$$\mathrm{Cov}\ (X, Y) = \frac{\sum(X_i - \bar{X})(Y_j - \bar{Y})}{n}$$

Sample

A sample includes one or multiple data points **from the population.**

Covariance formula (sample):

$$\mathrm{Cov}\ (X, Y) = \frac{\sum(X_i - \bar{X})(Y_j - \bar{Y})}{n - 1}$$

Legend

X_i – X-variable values	**Y_j** – Y-variable values	**n** – number of observations (data points)
$\bar{X}$ – X-variable mean (average)	**$\bar{Y}$** – Y-variable mean (average)	

Why is "n-1" in the sample formula?

The number of observations is subtracted by one since the sample data is less certain than the population data (since there are less data points available). This effectively "debiases" the data by slightly increasing the sample covariance.

The sample data, containing fewer data points, usually contains less variation than the population data. Since the sample data covariance would be understated, reducing "n" by 1 increases the covariance slightly to account for this.

A larger sample dataset (which results in a higher "n") will be less effected by the "n-1" denominator.

Example: What is the average age of all individuals in the United States of America?

Population: this would be used if you were using a dataset that included **every** individual in the United States of America.

Sample: this would be used if you were using a dataset which included **a portion** of every individual in the United States of America (for example, the ages of 500 individuals across the U.S. who were selected randomly).