Covariance

So far we've been looking almost exclusively at measures of one variable. But oftentimes we're interested in the relationship between two variables. Specifically, in this lesson, we want to look at the **covariance** of variables, which we can define as a measurement of how much two random variables vary together.

We already know that the variance for one variable is a measure of how much that variable varies away from its own mean. So when we expand that idea to two variables, *co*variance tells us how two variables change together. We can also say that covariance reflects the directional relationship between two random variables, but not the magnitude of the relationship.

We can calculate the covariance of a population with N members or of a sample with n members, for the variables x and y, such that each data point in the sample is a paired observation in the form (x_i, y_i), as

Population covariance

$$\sigma_{xy} = \frac{\sum (x_i - \mu_x)(y_i - \mu_y)}{N}$$

Sample covariance

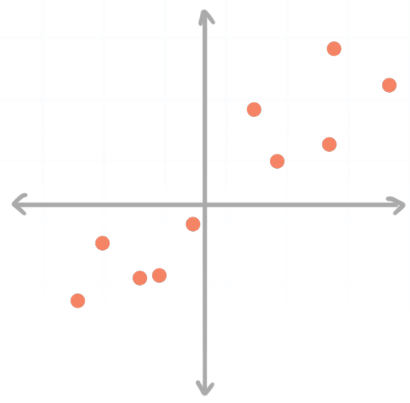
$$s_{xy} = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{n - 1}$$

Note that μ_x and μ_y are population means, and $\bar{x}$ and $\bar{y}$ are sample means. In addition, for very large samples, the number of members of the sample, $n - 1$ (as an unbiased estimate), will be roughly equal to the number of members of a population, N.

Interpreting covariance

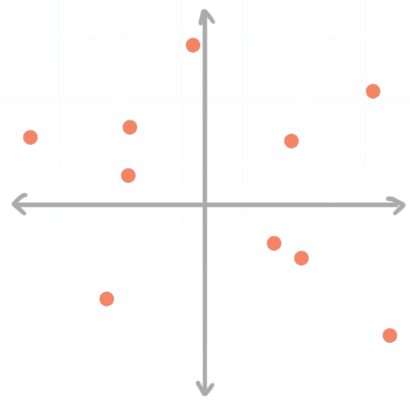
If a positive change in one variable causes a positive change in the other variable (or a negative change in one causes a negative change in the other), then the variables have a positive relationship and we should expect positive covariance. But if a positive change in one variable causes a negative change in the other variable, then the variables have a negative relationship and we should expect negative covariance.

And if the variables have no real discernible relationship, we'll expect a value for covariance that's close to 0.



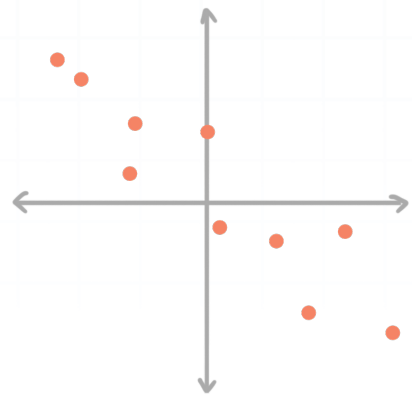
Positive covariance

Positive linear relationship between the variables



Approximately 0 covariance

No linear relationship between the variables



Negative covariance

Negative linear relationship between the variables

The limitation of covariance

Based on what we know now about the sign of covariance, we'd expect that a larger positive covariance means a stronger positive linear relationship, while a larger negative covariance means a stronger negative linear relationship.

However, the covariance calculation is unit-sensitive, meaning it changes with different units of measure.

For example, we can find a value for covariance using data given in hours per day, but we could also convert the hours to minutes, and do the same covariance calculation for the data in minutes per day. Because converting to minutes multiplies the hourly data by 60, using the minutes data will produce a much larger covariance figure than we'd find if we used the hourly data instead.

Let's do an example where we calculate covariance for the same set of data, but using two different units of measure.

Example

A coffee shop records sales and number of customers for a sample of hours throughout the week. Calculate the covariance of the data in dollars, then again in cents.

Customers	8	6	3	10	6	9
Revenue (dollars)	32.15	28.75	19.50	44.00	27.70	39.90
Revenue (cents)	3,215	2,875	1,950	4,400	2,760	3,990

We'll find the mean number of customers,

$$\bar{x} = \frac{8 + 6 + 3 + 10 + 6 + 9}{6}$$

$$\bar{x} = \frac{42}{6}$$

$$\bar{x} = 7$$

and the mean revenue in dollars.

$$\bar{y} = \frac{32.15 + 28.75 + 19.50 + 44.00 + 27.70 + 39.90}{6}$$

$$\bar{y} = \frac{192}{6}$$

$$\bar{y} = 32$$

Now we'll use the means to find the sample covariance.

$$s_{xy} = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{n - 1}$$

$$\sum (x_i - \bar{x})(y_i - \bar{y}) = (8 - 7)(32.15 - 32) + (6 - 7)(28.75 - 32)$$

$$+(3 - 7)(19.50 - 32) + (10 - 7)(44.00 - 32)$$

$$+(6 - 7)(27.70 - 32) + (9 - 7)(39.90 - 32)$$

$$\sum (x_i - \bar{x})(y_i - \bar{y}) = (0.15) - (-3.25) - 4(-12.5) + 3(12) - (-4.3) + 2(7.90)$$

$$\sum (x_i - \bar{x})(y_i - \bar{y}) = 0.15 + 3.25 + 50 + 36 + 4.3 + 15.80$$

$$\sum (x_i - \bar{x})(y_i - \bar{y}) = 109.5$$

$$s_{xy} = \frac{109.5}{6 - 1}$$

$$s_{xy} = 21.9$$

But now let's make the same covariance calculation, but this time with the data for cents, instead of dollars.

$$s_{xy} = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{n - 1}$$

$$\sum (x_i - \bar{x})(y_i - \bar{y}) = (8 - 7)(3{,}215 - 3{,}200) + (6 - 7)(2{,}875 - 3{,}200)$$

$$+(3 - 7)(1{,}950 - 3{,}200) + (10 - 7)(4{,}400 - 3{,}200)$$

$$+(6 - 7)(2{,}770 - 3{,}200) + (9 - 7)(3{,}990 - 3{,}200)$$

$$\sum (x_i - \bar{x})(y_i - \bar{y}) = (15) - (-325) - 4(-1{,}250) + 3(1{,}200) - (-430) + 2(790)$$

$$\sum (x_i - \bar{x})(y_i - \bar{y}) = 15 + 325 + 5{,}000 + 3{,}600 + 430 + 1{,}580$$

$$\sum (x_i - \bar{x})(y_i - \bar{y}) = 10{,}950$$

$$s_{xy} = \frac{10{,}950}{6 - 1}$$

$$s_{xy} = 2{,}190$$

So we see that we find a much larger covariance value for the cents data than the dollars data, even though both data sets are actually identical.

To correct for covariance's sensitivity to the units of measure of each variable, we prefer instead to calculate the correlation coefficient, which we'll look at in the next lesson.

In contrast to covariance, we'll see that the correlation coefficient is much more useful for determining the strength of the linear relationship between two variables, since its value is limited to the interval $[-1,1]$.