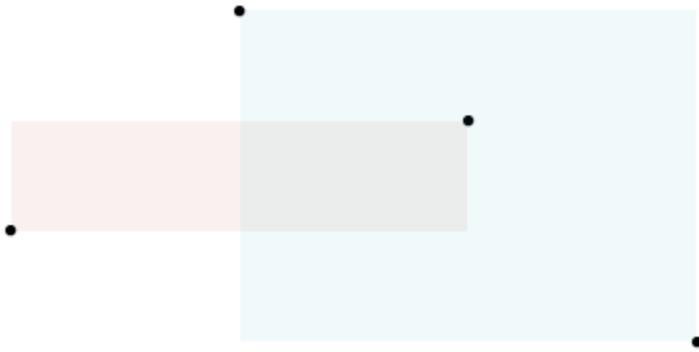


## COVARIANCE EXPLAINED GRAPHICALLY

Sometimes we can "augment knowledge" with an unusual or different approach. I would like this reply to be accessible to kindergartners and also have some fun, so **everybody get out your crayons!**

Given paired  $(x,y)$  data, draw their scatterplot. (The younger students may need a teacher to produce this for them. :-) Each pair of points  $(x_i, y_i)$ ,  $(x_j, y_j)$  in that plot determines a rectangle: it's the smallest rectangle, whose sides are parallel to the axes, containing those points. Thus the points are either at the upper right and lower left corners (a "positive" relationship) or they are at the upper left and lower right corners (a "negative" relationship).

**Draw all possible such rectangles.** Color them transparently, making the positive rectangles red (say) and the negative rectangles "anti-red" (blue). In this fashion, wherever rectangles overlap, their colors are either enhanced when they are the same (blue and blue or red and red) or cancel out when they are different.

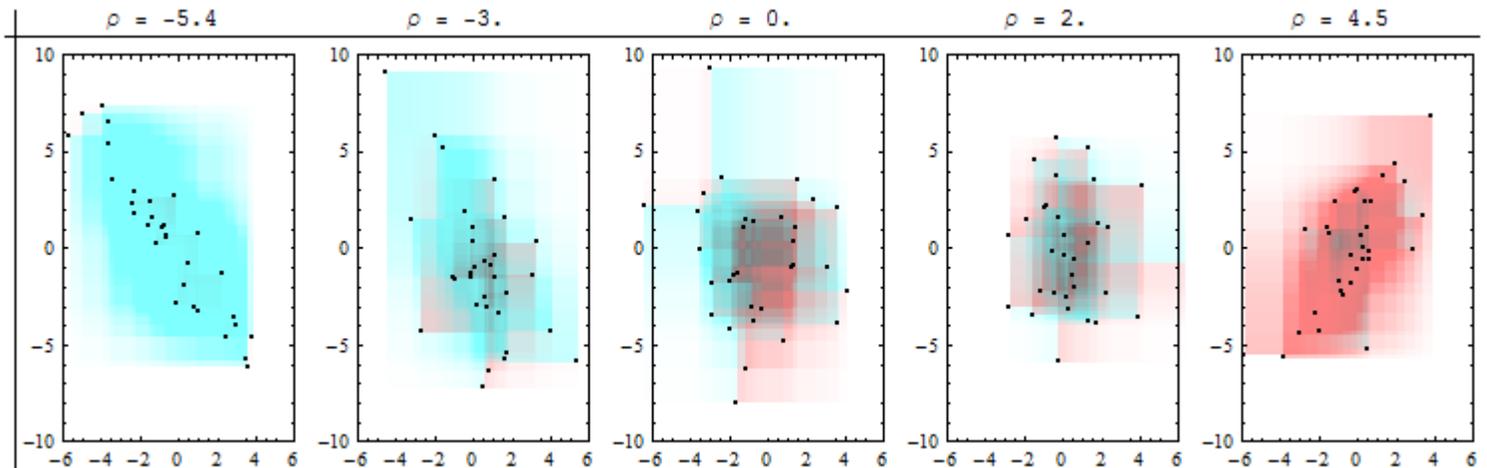


*(In this illustration of a positive (red) and negative (blue) rectangle, the overlap ought to be white; unfortunately, this software does not have a true "anti-red" color. The overlap is gray, so it will darken the plot, but on the whole the net amount of red is correct.)*

Now we're ready for the explanation of covariance.

**The covariance is the net amount of red in the plot** (treating blue as negative values).

Here are some examples with 32 binormal points drawn from distributions with the given covariances, ordered from most negative (bluest) to most positive (reddest).



**Let's deduce some properties of covariance.** Understanding of these properties will be accessible to anyone who has actually drawn a few of the rectangles. :-)

- **Bilinearity.** Because the amount of red depends on the size of the plot, covariance is directly proportional to the scale on the x-axis and to the scale on the y-axis.

- **Correlation.** Covariance increases as the points approximate an upward sloping line and decreases as the points approximate a downward sloping line. This is because in the former case most of the rectangles are positive and in the latter case, most are negative.
- **Relationship to linear associations.** Because non-linear associations can create mixtures of positive and negative rectangles, they lead to unpredictable (and not very useful) covariances. Linear associations can be fully interpreted by means of the preceding two characterizations.
- **Sensitivity to outliers.** A geometric outlier (one point standing away from the mass) will create many large rectangles in association with all the other points. It alone can create a net positive or negative amount of red in the overall picture.

Incidentally, this definition of covariance differs from the usual one only by a universal constant of proportionality (independent of the data set size). The mathematically inclined will have no trouble performing the algebraic demonstration of the equivalence.

---

Covariance is a measure of how much two variables change together. Compare this to Variance, which is just the range over which one measure (or variable) varies.

In studying social patterns, you might hypothesize that wealthier people are likely to be more educated, so you'd try to see how closely measures of wealth and education stay together. You would use a measure of covariance to determine this.

...

I'm not sure what you mean when you ask how does it apply to statistics. It is one measure taught in many stats classes. Did you mean, when should you use it?

You use it when you want to see how much two or more variables change in relation to each other.

Think of people on a team. Look at how they vary in geographic location compared to each other. When the team is playing or practicing, the distance between individual members is very small and we would say they are in the same location. And when their location changes, it changes for all individuals together (say, travelling on a bus to a game). In this situation, we would say they have a high level of covariance. But when they aren't playing, then the covariance rate is likely to be pretty low, because they are all going to different places at different rates of speed.

So you can predict one team member's location, based on another team member's location when they are practicing or playing a game with a high degree of accuracy. The covariance measurement would be close to 1, I believe. But when they are not practicing or playing, you would have a much smaller chance

of predicting one person's location, based on a team member's location. It would be close to zero, probably, although not zero, since sometimes team members will be friends, and might go places together on their own time.

However, if you randomly selected individuals in the United States, and tried to use one of them to predict the other's locations, you'd probably find the covariance was zero. In other words, there is absolutely no relation between one randomly selected person's location in the US, and another's.

Adding another one (by 'CatofGrey') that helps augment the intuition:

In probability theory and statistics, covariance is the measure of how much two random variables vary together (as distinct from variance, which measures how much a single variable varies).

If two variables tend to vary together (that is, when one of them is above its expected value, then the other variable tends to be above its expected value too), then the covariance between the two variables will be positive. On the other hand, if one of them is above its expected value and the other variable tends to be below its expected value, then the covariance between the two variables will be negative.

---

## DIFFERENCE BETWEEN COVARIANCE AND CORRELATION

The requirements of these types of questions strike me as a bit bizarre. Here is a *mathematical* concept/formula, yet I want to talk about it in some context completely devoid of mathematical symbols. I also think it should be stated that the actual algebra necessary to understand the formulas, I would think, should be taught to most individuals before higher education (no understanding of matrix algebra is needed, just simple algebra will suffice). So, at first instead of completely ignoring the formula and speaking of it in some magical and heuristic types of analogies, lets just look at the formula and try to explain the individual components in small steps. The difference in terms of covariance and correlation, when looking at the formulas, should become clear. Whereas speaking in terms of analogies and heuristics I suspect would obfuscate two relatively simple concepts and their differences in many situations.

So lets start out with a formula for the **sample covariance** (these I have just taken and adopted from wikipedia);

$$\frac{1}{N-1} \sum_{i=1}^N (x_i - \bar{x})(y_i - \bar{y})$$

To get everyone up to speed, lets explicitly define all of the elements and operations in the formula.

- $x_i$  and  $y_i$  are each measurements of two separate attributes of the same observation
- $\bar{x}$  and  $\bar{y}$  are the means (or average) of each attribute
- For  $\frac{1}{N-1}$ , lets just say this means we divide the final result

by  $\sum_{i=1}^n (x_i - \bar{x})^2$ .

- $\sum_{i=1}^n x_i$  may be a foreign symbol to some, so it would likely be useful to explain this operation. It is simply the sum of all  $x_i$  separate observations, and  $n$  represents the total number of observations.

At this point, I might introduce a simple example, to put a face on the elements and operations so to speak. So for example, let's just make up a table, where each row corresponds to an observation (and  $x$  and  $y$  are labeled appropriately). One would likely make these examples more specific (e.g. say  $x$  represents age and  $y$  represents weight), but for our discussion here it should not matter.

```
x y
---
2 5
4 8
9 3
5 6
0 8
```

At this point if you feel the sum operation in the formula may not have been fully comprehended, you can introduce it again in a much simpler context. Say just present that  $\sum_{i=1}^n (x_i)$  is the same as saying in this example;

```
x
--
2
4
9
5
+ 0
--
20
```

Now that mess should be cleared up, and we can work our way into the second part of the formula,  $(x_i - \bar{x})(y_i - \bar{y})(x_i - \bar{x})(y_i - \bar{y})$ . Now, assuming people already know what the mean,  $\bar{x}$  and  $\bar{y}$  stand for, and I would say, being hypocritical of my own comments earlier in the post, one can just refer to the mean in terms of simple heuristics (e.g. the middle of the distribution). One can then just take this process one operation at a time. The statement  $(x_i - \bar{x})(x_i - \bar{x})$  is just examining the deviations/distance between each observation, and the mean of all observations for that particular attribute. Hence when an observation is further from the mean, this operation will be given a higher value. One can then refer back to the example table given, and simply demonstrate the operation on the  $x$  vector of observations.

```
x x_bar (x - x_bar)
2 4      -2
4 4       0
9 4       5
5 4       1
0 4      -4
```

The operation is the same for  $y$  vector, but just for reinforcement you can present that operation as well.

```
y y_bar (y - y_bar)
5 6      -1
```

8	6	2
3	6	-3
6	6	0
8	6	2

Now, the terms  $(x_i - \bar{x})(x_i - \bar{x})$  and  $(y_i - \bar{y})(y_i - \bar{y})$  should not be ambiguous, and we can go onto the next operation, multiplying these results together,  $(x_i - \bar{x}) \cdot (y_i - \bar{y})(x_i - \bar{x}) \cdot (y_i - \bar{y})$ . As gung points out in the comments, this is frequently called the cross product (perhaps a useful example to bring back up if one were introducing basic matrix algebra for statistics).

Take note of what happens when multiplying, if two observations are both a large distance above the mean, the resulting observation will have an even larger positive value (the same is true if both observations are a large distance below the mean, as multiplying two negatives equals a positive). Also note that if one observation is high above the mean and the other is well below the mean, the resulting value will be large (in absolute terms) and negative (as a positive times a negative equals a negative number). Finally note that when a value is very near the mean for either observation, multiplying the two values will result in a small number. Again we can just present this operation in a table.

$(x - \bar{x})$	$(y - \bar{y})$	$(x - \bar{x}) * (y - \bar{y})$
-2	-1	2
0	2	0
5	-3	-15
1	0	0
-4	2	-8

Now if there are any statisticians in the room they should be boiling with anticipation at this point. We can see all the separate elements of what a covariance is, and how it is calculated come into play. Now all we have to do is sum up the final result in the preceding table, divide by  $N-1$  and *voila*, the covariance should no longer be mystical (all with only defining one greek symbol).

$(x - \bar{x}) * (y - \bar{y})$

```

-----
  2
  0
-15
  0
+ -8
-----
-21

```

$$-21 / (5-1) = -5.25$$

At this point you may want to reinforce where the 5 is coming from, but that should be as simple as referring back to the table and counting the number of observations (lets again leave the difference between sample and population to another time).

Now, the covariance in and of itself does not tell us much (it can, but it is needless at this point to go into any interesting examples without resorting to magically, undefined references to the audience). In a good case scenario, you won't really need to sell why we should care what the covariance is, in other circumstances, you may just have to hope your audience is captive and will take your word for it. But,

continuing on to develop the difference between what the covariance is and what the correlation is, we can just refer back to the formula for correlation. To prevent greek symbol phobia maybe just say  $\rho$  is the common symbol used to represent correlation.

$$\rho = \frac{\text{Cov}(x,y)}{\sqrt{\text{Var}(x)\text{Var}(y)}} \quad \rho = \frac{\text{Cov}(x,y)}{\sqrt{\text{Var}(x)\text{Var}(y)}}$$

Again, to reiterate, the numerator in the preceding formula is simply the covariance as we have just defined, and the denominator is the square root of the product of the *variance* of each individual series. If you need to define the variance itself, you could just say that the variance is the same thing as the covariance of a series with itself (i.e.  $\text{Cov}(x,x) = \text{Var}(x)$ ). And all the same concepts that you introduced with the covariance apply (i.e. if a series has many values a far ways from its mean, it will have a high variance). Maybe note here that a series can not have a negative variance as well (which should logically follow from the math previously presented).

So the only new components we have introduced are in the denominator,  $\sqrt{\text{Var}(x)\text{Var}(y)}$ . So we are dividing the covariance we just calculated by the product of the variances of each series. One could go into the treatment about why dividing by  $\sqrt{\text{Var}(x)\text{Var}(y)}$  will always result in a value between -1 and 1, but I suspect the Cauchy-Schwarz inequality should be left off of the agenda for this discussion. So again, I'm a hypocrite and resort to some, *take my word for it*, but at this point we can introduce all the reasons why we use the correlation coefficient. One can then relate these math lessons back to the heuristics that have been given in the other statements, such as [Peter Flom's response](#) to one of the other questions. While this was criticized for introducing the concept in terms of causal statements, that lesson should be on the agenda at some point as well.

I understand in some circumstances this level of treatment would not be appropriate. *The senate needs the executive summary*. In that case, well you can refer back to the simple heuristics that people have been using in other examples, but Rome wasn't built in a day. And to the senate whom asks for the executive summary, if you have so little time perhaps you should just take my word for it, and dispense with the formalities of analogies and bullet-points.

<http://stats.stackexchange.com/questions/18082/how-would-you-explain-the-difference-between-correlation-and-covariance>