14.6: Correlation Formula- Covariance Divided by Variability

The best way to learn the formula for correlations is to learn about two ideas and what they look like mathematically. We'll start with co-variance, which will become the numerator (top). Then we'll talk about standard deviations AGAIN for the denominator (bottom). To preview where we're going, here's the formula when we already have the standard deviation computed:

$$r = \frac{\left(\frac{\sum((x_{Each} - \bar{X}_x) \times (y_{Each} - \bar{X}_y))}{(N-1)}\right)}{(s_x \times s_y)}$$

We will go through each part below!

Numerator: Co-Variation

Because we have two quantitative variables, we will have two characteristics or score on which people will vary. What we want to know is do people vary on the scores together? That is, as one score changes, does the other score also change in a predictable or consistent way? This notion of variables differing together is called covariance (the prefix “co” meaning “together”).

Standard Deviation Refresher

We'll talk about standard deviations again in the denominator, but for now, let's look at the formula for standard deviation:

$$s = \sqrt{\frac{\sum(X - \overline{X})^2}{N-1}}$$

We use X to represent each score on the variable at hand, and $\overline{X}$ to represent the mean of that variable. The numerator of this formula is the Sum of Squares, which we have seen several times for various uses. Recall that squaring a value is just multiplying that value by itself. Thus, we can write the same equation as:

$$s = \sqrt{\frac{\sum((X - \overline{X})(X - \overline{X}))}{N-1}}$$

This is the same formula and works the same way as before, where we multiply the deviation score by itself (we square it) and then sum across squared deviations. You can trust me, or you can try it yourself on any data that has all scores and the calculated standard deviation provided!

Now, let’s look at the formula for covariance. In this formula, we will still use X to represent the score on one variable (but called x_{Each} to make it clear that you *subtract the mean from each number*), and we will now use y_{Each} to represent the score on the second variable. Some statisticians use $\bar{Y}$ to represent the mean of the second variable, but we've used subscripts for this whole book so we're going to keep that up. Either option will confuse a portion of you, so Dr. MO is sorry if you're the portion that is confused by $\bar{X}_y$ the mean of the "y" variable. When we start having variables, we'll use those as the subscripts again.

$$\text{COV} = \frac{\sum((x_{Each} - \bar{X}_x) \times (y_{Each} - \bar{X}_y))}{(N-1)}$$

As we can see, this is the exact same structure as the second formula presented (showing how squaring numbers is just multiplying them by themselves). Now, instead of multiplying the deviation score on one variable by itself , we take the two deviation scores from a single person on *each variable* and multiply them together. We do this for each person (exactly the same as we did for standard deviations) and then sum them to get our numerator. You can use Table 14.6.1 to help you calculate both the standard deviations and the correlation. If the standard deviations are provided, then you can skip the columns for $(x - \overline{X_x})^2$ and $(y - \overline{X_y})^2$.

Table 14.6.1: Sum of Products table

Participant	X	$(X - \overline{X_x})$	$(X - \overline{X_x})^2$ (skip this column if have SD)	Y	$(Y - \overline{X - y})$	$(Y - \overline{X_y})^2$ (skip this column if have SD)	$(X - \overline{X - x})(Y$
A							
B							
C							
and so on...							

Participant	X	$(X-\overline{X_x})$	$(X-\overline{X_x})^2$ (skip this column if have SD)	Y	$(Y-\overline{X-y})$	$(Y-\overline{X_y})^2$ (skip this column if have SD)	$(X-\overline{X-x})(Y$
Sum each column:	$\Sigma =$	$\Sigma =$	$\Sigma =$	$\Sigma =$	$\Sigma =$	$\Sigma =$	$\Sigma =$

The column headers tell you exactly what to do in that column. We list our raw data for the X and Y variables in the X and Y columns, respectively, then add them up so we can calculate the mean of each variable. We then take those means and subtract them from the appropriate raw score to get our deviation scores for each person on each variable, and the columns of deviation scores will both add up to zero. We will square our deviation scores for each variable to get the sum of squares for X and Y so that we can compute the variance and standard deviation of each (we will use the standard deviation in our equation below). Finally, we take the deviation score from each variable and multiply them together to get our product score. Summing this column will give us our sum of products. It is very important that you multiply the raw deviation scores from each variable, NOT the squared deviation scores.

Let's get back to the formula to see where we're at:

$$r = \frac{\left(\frac{\sum((x_{Each} - \bar{X}_x) \times (y_{Each} - \bar{X}_y))}{(N-1)}\right)}{(s_x \times s_y)}$$

We took each score from the first variable ("x") and subtracted that variables mean from it. Then subtracted the mean of the second variable ("y") from each score of that variable. Then we multiplied them all together, and finally added up all of those products.

When we add up all of the answers from the the last column in Table 14.6.1 to calculate find the numerator of the numerator, also known as the numerator of the covariation formula ($\text{COV} = \frac{\sum((x_{Each} - \bar{X}_x) \times (y_{Each} - \bar{X}_y))}{(N-1)}$) from the table, and then we only have to divide by $N-1$ to get our covariance (the numerator of the correlation formula). Note that N is the number of people, so the number the pairs. It is not the number of scores. For example, if we measured 10 participants' happiness scores and their chocolate supply, we would have 10 scores for happiness and 10 scores for chocolate, and N would also be 10 (n = 10) because we had 10 people.

Unlike the sum of squares, both our sum of products and our covariance can be positive, negative, or zero, and they will always match (e.g. if our sum of products is positive, our covariance will always be positive). A positive sum of products and covariance indicates that the two variables are related and move in the same direction. That is, as one variable goes up, the other will also go up, and vice versa. A negative sum of products and covariance means that the variables are related but move in opposite directions when they change, which is called an inverse relation. In an inverse relation, as one variable goes up, the other variable goes down. If the sum of products and covariance are zero, then that means that the variables are not related. As one variable goes up or down, the other variable does not change in a consistent or predictable way.

The previous paragraph brings us to an important definition about relations between variables. What we are looking for in a relation is a consistent or predictable pattern. That is, the variables change together, either in the same direction or opposite directions, in the same way each time. It doesn't matter if this relation is positive or negative, only that it is not zero. If there is no consistency in how the variables change within a person, then the relation is zero and does not exist. We will revisit this notion of direction vs zero relation later on.

Denominator: Standard Deviations

As you can see in the formula for Pearson's correlation:

$$r = \frac{\left(\frac{\sum((x_{Each} - \bar{X}_x) \times (y_{Each} - \bar{X}_y))}{(N-1)}\right)}{(s_x \times s_y)}$$

that the denominator is just multiplying the two standard deviations together. It looks like this:

$$r_{denominator} = \left(\sqrt{\frac{\sum(x-\overline{X_x})^2}{N-1}}\right) \times \left(\sqrt{\frac{\sum(y-\overline{X_y})^2}{N-1}}\right)$$

Easy-peasy!

Full Formula to Calculate Standard Deviations

Okay, I don't want to scare you, but I do want you to be prepared. Although, firstly, you will probably never calculate a standard deviation or a correlation by hand outside of this class. Secondly, even if your professor asks you to calculate a correlation, it is sorta unlikely that they wouldn't just give you the standard deviations. That's a lot of calculations to do by hand! But just in case, here is the very fullest of formulas for Pearson's r:

$$\frac{\left(\frac{\sum((x-\bar{X}_x)*(y-\bar{X}_y))}{(N-1)}\right)}{\left(\sqrt{\frac{\sum(x-\overline{\bar{X}_x})^2}{N-1}}\right)\times\left(\sqrt{\frac{\sum(y-\overline{\bar{X}_y})^2}{N-1}}\right)}$$

Next up, let's practice using these formulas!
