3.3: Turning the numbers into a measure of co-variance

"OK, so if you are saying that co-variance is just another word for correlation or relationship between two measures, I'm good with that. I suppose we would need some way to measure that." Correct, back to our table…notice anything new?

subject	chocolate	happiness	Chocolate_X_Happiness
1	1	1	1
2	2	2	4
3	2	3	6
4	3	3	9
5	3	3	9
6	5	5	25
7	4	6	24
8	5	5	25
9	9	5	45
10	6	9	54
Sums	40	42	202
Means	4	4.2	20.2

We've added a new column called "Chocolate_X_Happiness", which translates to Chocolate scores multiplied by Happiness scores. Each row in the new column, is the product, or multiplication of the chocolate and happiness score for that row. Yes, but why would we do this?

Last chapter we took you back to Elementary school and had you think about division. Now it's time to do the same thing with multiplication. We assume you know how that works. One number times another, means taking the first number, and adding it as many times as the second says to do,

\(2*2= 2+2=4 \nonumber \)

\(2*6= 2+2+2+2+2+2 = 12\), or \(6+6=12\), same thing.

Yes, you know all that. But, can you bend multiplication to your will, and make it do your bidding when need to solve a problem like summarizing co-variance? Multiplication is the droid you are looking for.

We know how to multiple numbers, and all we have to next is think about the consequences of multiplying sets of numbers together. For example, what happens when you multiply two small numbers together, compared to multiplying two big numbers together? The first product should be smaller than the second product right? How about things like multiplying a small number by a big number? Those products should be in between right?.

Then next step is to think about how the products of two measures sum together, depending on how they line up. Let's look at another table:

scores	X	Y	A	B	XY	AB
1	1	1	1	10	1	10
2	2	2	2	9	4	18
3	3	3	3	8	9	24
4	4	4	4	7	16	28
5	5	5	5	6	25	30
6	6	6	6	5	36	30
7	7	7	7	4	49	28

scores	X	Y	A	B	XY	AB
8	8	8	8	3	64	24
9	9	9	9	2	81	18
10	10	10	10	1	100	10
Sums	55	55	55	55	385	220
Means	5.5	5.5	5.5	5.5	38.5	22

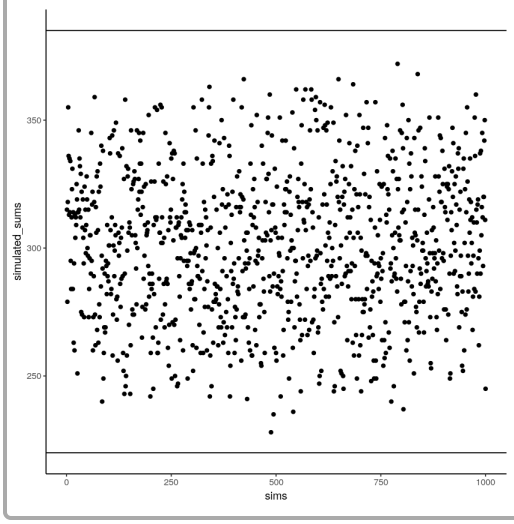
Look at the X and Y column. The scores for X and Y perfectly co-vary. When X is 1, Y is 1; when X is 2, Y is 2, etc. They are perfectly aligned. The scores for A and B also perfectly co-vary, just in the opposite manner. When A is 1, B is 10; when A is 2, B is 9, etc. B is a reversed copy of A.

Now, look at the column \(XY\). These are the products we get when we multiply the values of X across with the values of Y. Also, look at the column \(AB\). These are the products we get when we multiply the values of A across with the values of B. So far so good.

Now, look at the `Sums` for the XY and AB columns. Not the same. The sum of the XY products is 385, and the sum of the AB products is 220. For this specific set of data, the numbers 385 and 220 are very important. They represent the biggest possible sum of products (385), and the smallest possible sum of products (220). There is no way of re-ordering the numbers 1 to 10, say for X, and the numbers 1 to 10 for Y, that would ever produce larger or smaller numbers. Don't believe me? Check this out:

```
library(ggplot2)
simulated_sums<-length(0)
for(sim in 1:1000){
    X<-sample(1:10)
    Y<-sample(1:10)
    simulated_sums[sim]<-sum(X*Y)
}
sim_df<-data.frame(sims=1:1000,simulated_sums)
ggplot(sim_df,aes(x=sims,y=simulated_sums))+
  geom_point()+
  theme_classic()+
  geom_hline(yintercept = 385)+
  geom_hline(yintercept = 220)
```

run | restart | restart & run all



The above graph shows 1000 computer simulations. I convinced my computer to randomly order the numbers 1 to 10 for X, and randomly order the numbers 1 to 10 for Y. Then, I multiplied X and Y, and added the products together. I did this 1000 times. The dots show the sum of the products for each simulation. The two black lines show the maximum possible sum (385), and the minimum possible sum (220), for this set of numbers. Notice, how all of the dots are in between the maximum and minimum possible values. Told you so.

"OK fine, you told me so…So what, who cares?". We've been looking for a way to summarize the co-variance between two measures right? Well, for these numbers, we have found one, haven't we. It's the sum of the products. We know that when the sum of the products is 385, we have found a perfect, positive correlation. We know, that when the sum of the products is 220, we have found a perfect negative correlation. What about the numbers in between. What could we conclude about the correlation if we found the sum of the products to be 350. Well, it's going to be positive, because it's close to 385, and that's perfectly positive. If the sum of the products was 240, that's going to be negative, because it's close to the perfectly negatively correlating 220. What about no correlation? Well, that's going to be in the middle between 220 and 385 right.

We have just come up with a data-specific summary measure for the correlation between the numbers 1 to 10 in X, and the numbers 1 to 10 in Y, it's the sum of the products. We know the maximum (385) and minimum values (220), so we can now interpret any product sum for this kind of data with respect to that scale.

> *Pro tip: When the correlation between two measures increases in the positive direction, the sum of their products increases to its maximum possible value. This is because the bigger numbers in X will tend to line up with the bigger numbers in Y, creating the biggest possible sum of products. When the correlation between two measures increases in the negative direction, the sum of their products decreases to its minimum possible value. This is because the bigger numbers in X will tend to line up with the smaller numbers in Y, creating the smallest possible sum of products. When there is no correlation, the big numbers in X will be randomly lined up with the big and small numbers in Y, making the sum of the products, somewhere in the middle.*

Co-variance, the measure

We took some time to see what happens when you multiply sets of numbers together. We found that $\textit{big} * \textit{big} = \text{bigger}$ and $\textit{small} * \textit{small} = \text{still small}$, and $\textit{big} * \textit{small} = \text{in the middle}$. The purpose of this was to give you some conceptual idea of how the co-variance between two measures is reflected in the sum of their products. We did something very straightforward. We just multiplied X with Y, and looked at how the product sums get big and small, as X and Y co-vary in different ways.

Now, we can get a little bit more formal. In statistics, co-variance is not just the straight multiplication of values in X and Y. Instead, it's the multiplication of the deviations in X from the mean of X, and the deviation in Y from the mean of Y. Remember those difference scores from the mean we talked about last chapter? They're coming back to haunt you know, but in a good way like Casper the friendly ghost.

Let's see what this look like in a table:

subject	chocolate	happiness	C_d	H_d	Cd_x_Hd
1	1	1	-3	-3.2	9.6
2	2	2	-2	-2.2	4.4
3	2	3	-2	-1.2	2.4
4	3	3	-1	-1.2	1.2
5	3	3	-1	-1.2	1.2
6	5	5	1	0.8	0.8
7	4	6	0	1.8	0
8	5	5	1	0.8	0.8
9	9	5	5	0.8	4

subject	chocolate	happiness	C_d	H_d	Cd_x_Hd
10	6	9	2	4.8	9.6
Sums	40	42	0	0	34
Means	4	4.2	0	0	3.4

We have computed the deviations from the mean for the chocolate scores (column `C_d`), and the deviations from the mean for the happiness scores (column `H_d`). Then, we multiplied them together (last column). Finally, you can see the mean of the products listed in the bottom right corner of the table, the official the covariance.
The formula for the co-variance is:
\[cov(X,Y) = \frac{\sum_{i}^{n}(x_{i}-\bar{X})(y_{i}-\bar{Y})}{N} \nonumber \]
OK, so now we have a formal single number to calculate the relationship between two variables. This is great, it's what we've been looking for. However, there is a problem. Remember when we learned how to compute just the plain old variance. We looked at that number, and we didn't know what to make of it. It was squared, it wasn't in the same scale as the original data. So, we square rooted the variance to produce the standard deviation, which gave us a more interpretable number in the range of our data. The co-variance has a similar problem. When you calculate the co-variance as we just did, we don't know immediately know its scale. Is a 3 big? is a 6 big? is a 100 big? How big or small is this thing?
From our prelude discussion on the idea of co-variance, we learned the sum of products between two measures ranges between a maximum and minimum value. The same is true of the co-variance. For a given set of data, there is a maximum possible positive value for the co-variance (which occurs when there is perfect positive correlation). And, there is a minimum possible negative value for the co-variance (which occurs when there is a perfect negative correlation). When there is zero co-variation, guess what happens. Zeroes. So, at the very least, when we look at a co-variation statistic, we can see what direction it points, positive or negative. But, we don't know how big or small it is compared to the maximum or minimum possible value, so we don't know the relative size, which means we can't say how strong the correlation is. What to do?

Pearson's r we there yet

Yes, we are here now. Wouldn't it be nice if we could force our measure of co-variation to be between -1 and +1?
-1 would be the minimum possible value for a perfect negative correlation. +1 would be the maximum possible value for a perfect positive correlation. 0 would mean no correlation. Everything in between 0 and -1 would be increasingly large negative correlations. Everything between 0 and +1 would be increasingly large positive correlations. It would be a fantastic, sensible, easy to interpret system. If only we could force the co-variation number to be between -1 and 1. Fortunately, for us, this episode is brought to you by Pearson's \(r\), which does precisely this wonderful thing.
Let's take a look at a formula for Pearson's \(r\):
\[r = \frac{cov(X,Y)}{\sigma_{X}\sigma_{Y}} = \frac{cov(X,Y)}{SD_{X}SD_{Y}} \nonumber \]
We see the symbol \(\sigma\) here, that's more Greek for you. \(\sigma\) is often used as a symbol for the standard deviation (SD). If we read out the formula in English, we see that r is the co-variance of X and Y, divided by the product of the standard deviation of X and the standard deviation of Y. Why are we dividing the co-variance by the product of the standard deviations. This operation has the effect of normalizing the co-variance into the range -1 to 1.

Note

But, we will fill this part in as soon as we can…promissory note to explain the magic. FYI, it's not magic. Brief explanation here is that dividing each measure by its standard deviation ensures that the values in each measure are in the same range as one another.

For now, we will call this mathematical magic. It works, but we don't have space to tell you why it works right now.

> *It's worth saying that there are loads of different formulas for computing Pearson's \(r\). You can find them by Googling them. We will probably include more of them here, when we get around to it. However, they all give you the same answer. And, they are all not as pretty as each other. Some of them might even look scary. In other statistics textbook you will often find formulas that are easier to use for calculation purposes. For example, if*

> *you only had a pen and paper, you might use one or another formula because it helps you compute the answer faster by hand. To be honest, we are not very interested in teaching you how to plug numbers into formulas. We give one lesson on that here: Put the numbers into the letters, then compute the answer. Sorry to be snarky. Nowadays you have a computer that you should use for this kind of stuff. So, we are more interested in teaching you what the calculations mean, rather than how to do them. Of course, every week we are showing you how to do the calculations in lab with computers, because that is important to.*

Does Pearson's \(r\) really stay between -1 and 1 no matter what? It's true, take a look at the following simulation. Here I randomly ordered the numbers 1 to 10 for an X measure, and did the same for a Y measure. Then, I computed Pearson's \(r\), and repeated this process 1000 times. As you can see all of the dots are between -1 and 1. Neat huh.

```
library(ggplot2)
simulated_sums <- length(0)
for(sim in 1:1000){
  X <- sample(1:10)
  Y <- sample(1:10)
  simulated_sums[sim] <- cor(X,Y)
}
sim_df <- data.frame(sims=1:1000,simulated_sums)
ggplot(sim_df, aes(x = sims, y = simulated_sums))+
  geom_point()+
  theme_classic()+
  geom_hline(yintercept = -1)+
  geom_hline(yintercept = 1)+
  ggtitle("Simulation of 1000 r values")
```

run | restart | restart & run all

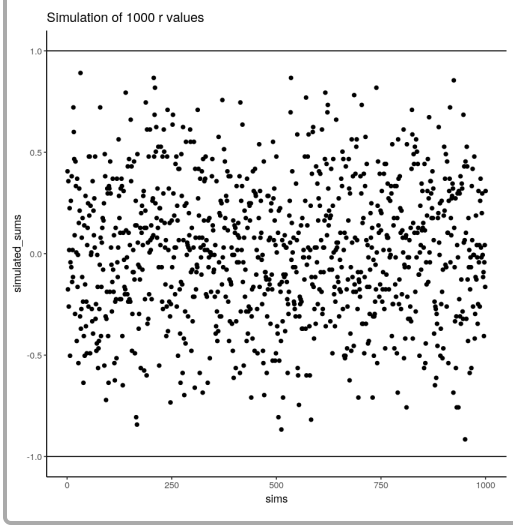


Figure \(\PageIndex{1}\): A simulation of of correlations. Each dot represents the r-value for the correlation between an X and Y variable that each contain the numbers 1 to 10 in random orders. The figure ilustrates that many r-values can be obtained by this random process.