

Introduction to Correlational Techniques

*Dr. Tom Pierce
Department of Psychology
Radford University*

One of the mantras you hear repeated over and over again in an undergraduate statistics class is that “you can’t draw a conclusion about cause and effect from a correlation”. For a lot of people, that’s the one thing they remember about a correlation – what you can’t do with it. That’s really a shame because a correlation is one of the most useful tools that psychologists have. Consequently, the goal of this chapter is to introduce the use of correlational tools and to show you what you can do with them.

In this chapter, we’ll concentrate on situations where there are only two variables involved. For example, let’s say that we’ve collected scores for a measure of social support and for a measure of life satisfaction from five Alzheimer’s caregivers. The researcher is interested in testing the idea that higher levels of supportive contact with friends and family members is associated with higher levels of life satisfaction in person who are in highly stressful situations. The scores for the measure of social support can range between 1 and 15. The scores for life satisfaction can range between 1 and 10. For shorthand, let’s refer to the measure of social support as variable X and the measure of life satisfaction as variable Y.

	X	Y
Participant	Social Support	Life Satisfaction
1	4	6
2	6	5
3	8	7
4	10	9
5	12	9
	$\bar{X} = 8.0$	$\bar{Y} = 7.2$
	$S_X = 3.16$	$S_Y = 1.79$

Okay, what do you think? When you look at the scores for social support and life satisfaction, does it look like there’s a relationship between the two variables? Do the scores for the two variables “go together” in some way? Most people are inclined to say “sort of”. When pressed they’ll say that the people who had higher scores on social support also tended to have higher scores on life satisfaction. At the very least at this point you get the feeling that the two variables have something to do with each other. But what? What information about the relationship would the researcher want to be in a position to report? It turns out that there are two basic pieces of information that one

would need to provide in order to adequately describe the relationship between two variables: the direction and the strength of the relationship.

Describing the relationship between two variables

Direction

In the example above it looks like higher scores on the measure of social support tend to go along with having higher scores on the measure of life satisfaction. When higher scores on one variable are associated with higher scores on a second variable, we'd say that there is a **positive relationship** between the two variables.

Now think about the relationship between two other variables: social support and depression. If you had to guess, how would you describe the way in which these variables are related to each other? It would certainly seem reasonable to think that higher scores on social support would be associated with having *lower* scores on depression. In this example, we'd say that there's a **negative relationship** between the two variables because it's a situation where higher scores on one variable tend to go along with having lower scores on the other variable.

Now take a third pair of variables, I.Q. scores and shoe size. What would you say about the direction of this relationship? Is it positive or negative? Unless you're a big fan of the Big Foot-Big Brain theory of intelligence you'd probably say that the answer should be "neither". If, in fact the two variables have nothing to do with each other, as in this case, one's answer about the direction of the relationship is that there is "**no relationship**" between the two variables.

The strength of the relationship between two variables

The strength of the relationship between two variables refers to the degree to which the scores for the two variables go together. One way of talking about that would be in terms of how much information the two variables share. In other words, when you know a person's score on one variable, how much information do you get about what their score is likely to be on the other variable. Do you get practically no information, as in the I.Q. and shoe size example, or do you get a lot of information, as in the relationship between social support and life satisfaction? The more the two variable overlap in the information they're giving you, the stronger the relationship is said to be between them.

Assessing the relationship between two variables using a scatterplot

So far we've talked about the relationship between two variables in verbal terms, which is perfectly fine, but another way to think about the relationship is by considering the pattern that one sees in a type of graph known as a scatterplot. The idea behind a scatterplot is pretty simple. The X-axis of the graph displays possible scores that participants might have on one variable (logically enough, the one labeled as variable X). The Y-axis displays possible scores that participants might have on the other variable

(variable Y). For each subject, we can locate a point in the scatterplot that represents the combination of their scores on both variables X and Y. For example, Figure X.1 shows the position of Subject #1 in the scatterplot. Figure X.2 displays the location of all five subjects in the data set.

Figure X.1

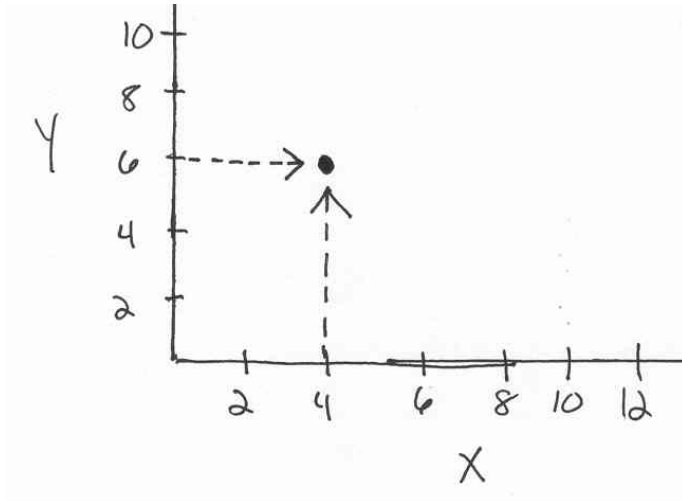
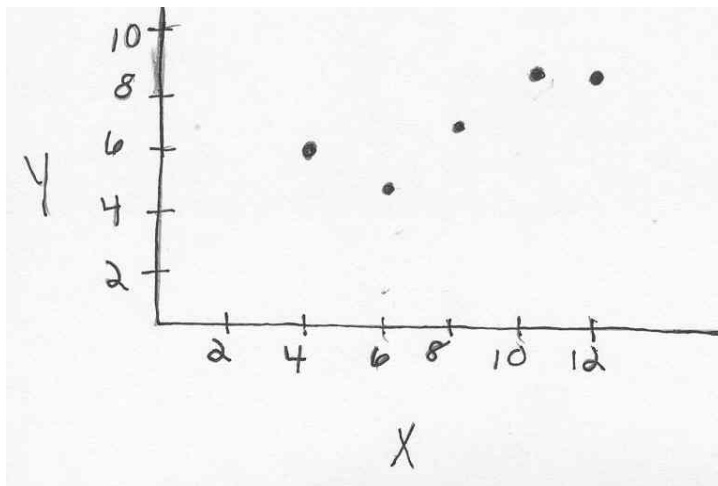


Figure X.2



Direction of the relationship. The interesting thing about a scatterplot is that it provides a visual source of information for thinking about the relationship between the two variables. In our little data set, for example, your eye detects a pattern when you inspect the points going from left to right across the graph. It appears that the points are getting higher as we move from left to right. Another way of thinking about it is to imagine what a line would look like if you were to draw the line through the points in the scatterplot that runs the closest, on average, to those points. In this case, the line would be going up

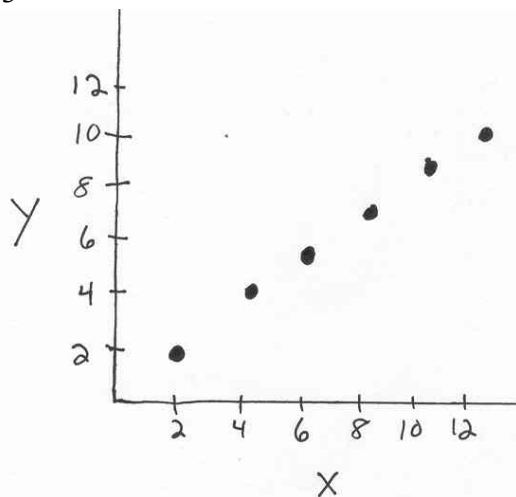
as you go from left to right. When this is the case, the graph tells us that we're dealing with a **positive relationship**. If the pattern we saw in the scatterplot were one where this best-fitting line was going down as we go from left to right, we'd say that we're dealing with a **negative relationship**.

If there were **no relationship** between variables X and Y the points in the scatterplot would just be scattered in a more or less random way around the area of the graph. Figure X.4 shows a hypothetical scatterplot of the relationship between shoe size and I.Q. scores. In this scatterplot someone with huge feet is just as likely to have a high I.Q. as a low I.Q.. Also, someone with teeny tiny feet is just as likely to have a high I.Q. as a low I.Q. Intelligence (at least, as measured by this particular test) doesn't seem to have anything to do with shoe size. If you were asked to draw a line through these points in the scatterplot a line that's going up as you move from left to right doesn't seem any better or worse than a line that's going down. About the best you can do is to draw a flat line through those points.

Strength of the relationship. In the section above on this topic we talked about the strength of the relationship between two variables in terms of the degree to which the two variables overlap in terms of the information they provide. In a visual sense, when statisticians talk about the strength of the relationship they're talking about the degree to which the points in the scatterplot can be fit by the pattern of a straight line. Statisticians like talking in terms of a straight line because (a) it's the simplest kind of pattern that you could have between two variables and (b) because it turns out that most relationships that exist between variables are fit very well by the simple pattern of a straight line.

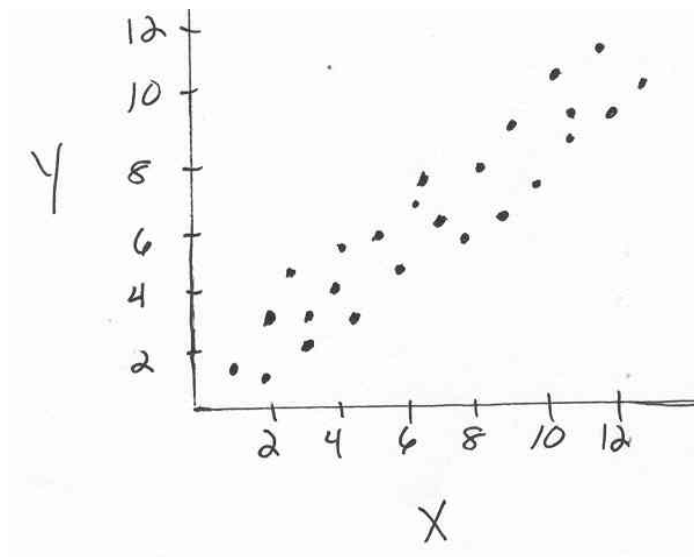
The strongest relationship you could have between two variables is a **perfect relationship**. In terms of the scatterplot, you'd know you were dealing with a perfect relationship when the points in the scatterplot fall on a perfectly straight line that's going either up or down. To have a perfect relationship means that knowing a person's score on one of the variables would tell you everything you needed to know in order to make a *perfect guess* about their score for the other variable. See Figure X.3.

Figure X.3



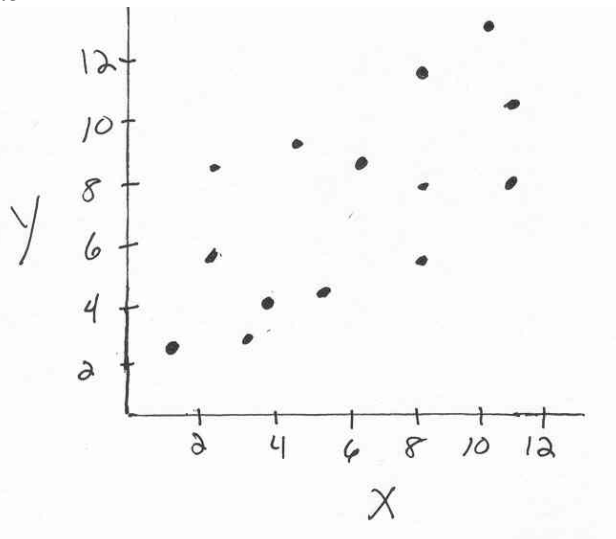
If you run a straight line through the points in the scatterplot and line runs pretty close to those points, but not right through all of them you'd say that you were dealing with a **strong relationship** between the two variables. See Figure X.4.

Figure X.4



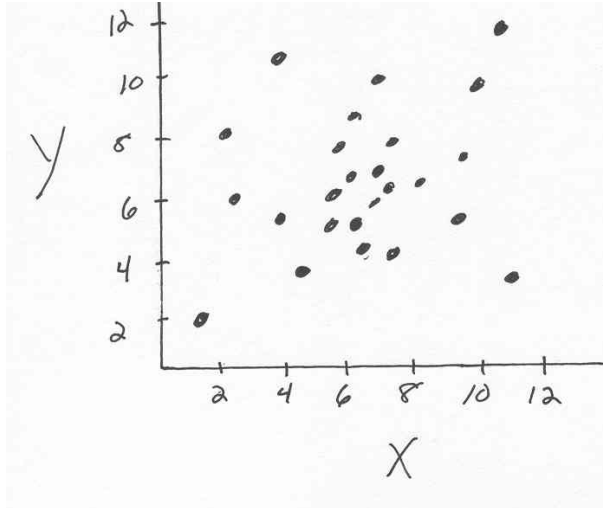
If you've got a scatterplot where, for example, it looks like the points are moving up as you go from left to right, but a line you draw through those points isn't all that close, on average, to those points you'd say that you have a weak relationship between the two variables. See Figure X.5.

Figure X.5



Again, if it looks like the points are scattered randomly around the scatterplot you'd say that there is **no relationship** between the two variables. See Figure X.6.

Figure X.6



Building the Pearson correlation coefficient

Now, there is absolutely nothing wrong with evaluating the direction and strength of the relationship between two variables by looking at a scatterplot. In fact, one of the first things someone should do when conducting correlational research is to look at the scatterplots. But, when the researcher is in the position of trying to convince another researcher that a relationship of a certain type exists between two variables it would be nice to report one number that provides information about both the direction and the strength of the relationship.

By far the most common number that researchers use to describe the relationship between two variables is the Pearson Correlation Coefficient; otherwise known as the Pearson r . Now, to start with, you of course remember a number of things about the Pearson r . For example, you'll recall that...

- Values for r can range between -1 and +1.
- A negative value for r means that there is a negative relationship between X and Y .
- A positive value for r means that there is a positive relationship between X and Y .
- A value for r of zero means that there is no relationship between X and Y .
- The further away from zero that a correlation is, the stronger the relationship between the two variables.
- Obviously, the furthest away that a correlation coefficient can get in the negative direction is a value of -1. This would reflect a situation where there is a perfect negative relationship between the two variables and the points in the scatterplot would fall on a perfectly straight line that is going down as you go from left to right.
- The furthest away that a correlation coefficient can get in the positive direction is a value for r of +1. This would reflect a situation where there is a perfect positive

relationship between the two variables and the points in the scatterplot would fall on a perfectly straight line that's going up as you go from left to right. See Figure X.3 above.

- In short, the sign of the correlation coefficient tells you about the direction of the relationship and the distance from zero tells you about the strength of the relationship. Pretty efficient for one number!

Okay, you already know a fair bit about the information you get from a Pearson correlation coefficient. And, I'd be willing to bet that your undergraduate course had you calculate r using one of a number of unpleasant looking calculator formulas (and I'm sure you were a better person for having done it). But what I bet you didn't cover was the fact that these equations represent the end result of building a better mousetrap – in this case the mousetrap has the job of “capturing” the relationship between two variables. It turns out that we can think about calculating a Pearson correlation coefficient as a series of steps and that each step allows the researcher to answer a kind of question that the previous step couldn't handle.

Sum of Cross-Products

Here goes. The place to start in thinking about where a correlation coefficient comes from is with the scatterplot. Below is the scatterplot for the social support and life satisfaction data we talked about above. If you remember, the mean for social support (variable X) was 8.0. The mean for life satisfaction (variable Y) was 7.2.

In the next graph (X.7), you'll notice that that the scatterplot is divided into four quadrants. The scale for variable X is divided in half at the mean for that variable. The scale for Y is divided in half based on the mean for variable Y.

Now, let's just look at the X-axis for a moment. Let's say that we convert the raw scores for variable X to deviation scores. In other words, you subtract the mean for variable X (8.0) from each raw score. Obviously, what'll happen is that anytime you get a raw score for X that's less than 8.0 you'll end up with a deviation score that's a negative number. And anytime you get a raw score for X that's greater than 8.0 you'll get a deviation score that's a positive number.

Now look at the Y-axis. Let's say we convert raw scores for variable Y to deviation scores. Here, anytime you have a raw score for Y that's less than 7.2 you get a deviation score that's a negative number. Anytime you take a raw score for Y that's greater than 7.2 you'll end up with a deviation score that's a positive number. Notice that I've put plus-signs and negative signs along both axes to remind us of where the positive and negative deviation scores are for both variables.

Now, look at the data point for subject 1 in the data set. That point is located in the bottom-left quadrant of the scatterplot. See Figure X.X. If you take the deviation score for that person on variable X and then multiply it by the deviation score for that person on variable Y, will you get a positive number or a negative number? Well, you know that

as far as variable X goes, you're dealing with a negative number (because 4 is less than the mean of X of 8.0). You also know that as far as variable Y goes, you're dealing with a negative number (because 6 is less than the mean of Y of 7.2). So, you're taking one negative number and multiplying it by another negative number. What do you get. A positive number! When you multiply a person's deviation score for variable X by their deviation score for variable Y you end up with a value that statisticians refer to as a **cross-product**.

Here's where the four quadrants come in. We'll talk about them one at a time.

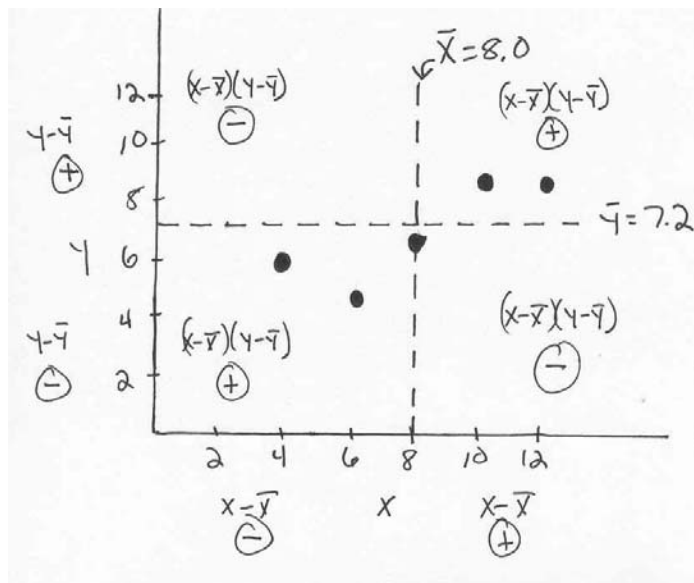
Bottom-left. All the points in the scatterplot that fall in the bottom-left quadrant are going to have cross-products that are **positive** numbers (negative numbers multiplied by negative numbers).

Top-right. All the points that fall in the top-right quadrant are also going to have **positive** numbers for their cross-products. This is because you're taking positive deviation scores for X and multiplying them by positive deviation scores for Y.

Top-left. The points in the top-left quadrant are going to have negative cross-products because a negative number is what you end up with when you multiply a negative deviation score for X by a positive deviation score for Y.

Bottom-right. The points in the bottom-right quadrant are also going to have negative cross-products because you're taking positive deviation scores for X and multiplying them by negative deviation scores for Y.

Figure X.7 shows the four quadrants of the scatterplot and the signs that are associated with the cross-products for points contained in each quadrant.



One simple way of thinking about whether the relationship between the two variables is positive or negative is to **see if looks like there are more point in the positive quadrants or the negative quadrants**. In our example, all of the points fall in the positive quadrants and none of them fall in the negative quadrants. When the number of positive cross-products is greater than the number of negative cross-products, it usually


means that a positive relationship exists. When the negative cross-products outnumber the positive cross-products, it's almost always the case that a negative relationship exists.

Now, let's get the actual numbers for the cross-products. We're going to get the deviation scores for both X and Y and then multiply these numbers together for each person.

Calculation of the Sum of Cross-Products

Participant	X Social Support	Y Life Satisfaction	X- \bar{X}	Y- \bar{Y}	(X- \bar{X})(Y- \bar{Y}) Cross-Product
1	4	6	-4.0	-1.2	4.8
2	6	5	-2.0	-2.2	4.4
3	8	7	0.0	-0.2	0.0
4	10	9	+2.0	+1.8	3.6
5	12	9	+4.0	+1.8	7.2
		$\bar{X} = 8.0$ $S_X = 3.16$	$\bar{Y} = 7.2$ $S_Y = 1.79$		20.0

$$\text{Sum of Cross-Products} = \sum(X - \bar{X})(Y - \bar{Y})$$



The cross-product for each person is displayed in the column at the far right. Now, to get one number that describes the relationship X and Y all you have to do is to add these cross-products up! And the number you end up with is 20. Logically enough, the name for this measure is the **Sum of Cross-Products**.

Basically, if the sum of cross-products is a positive number this means that the weight of points in the positive quadrants is greater than the weight of points in the negative quadrants. So *if the sum of cross-products is a positive number there is a positive relationship between the two variables. If the sum of cross-products is a negative number there is a negative relationship between the two variables.* This property of the sum of cross-products is at the root of why the Pearson Correlation Coefficient ends up being a positive number or a negative number. If the sum of cross-products is zero this means that the weight of points in the positive quadrants is balanced out by the weight of points in the negative quadrants. In this situation there is said to be no relationship between the two variables. The further the sum of cross-products is from zero the stronger the relationship there is between the two variables.

The Covariance

The Sum of Cross-Products is a perfectly good number to use in measuring the direction and strength of relationship between two variables. However, there are some kinds of questions that this measure just can't help with. For example, let's say that you're asked to determine whether the relationship between social support and life satisfaction is stronger in males or females. You've got a sample of 10 males and a sample of 20 females. You calculate the sum of cross-products for the 10 males and get a value of 40. You calculate the sum of cross-products for the 20 females and get a value of 80. We said before that the further the sum of cross-products is from zero the stronger the relationship between the two variables. Eighty is further away from zero than 40, so obviously the relationship is stronger in females than in males. Right? NO! It's the same problem we ran into in interpreting a sum of squares in Chapter 1. Think about it. Twenty numbers got added up to give us the sum of cross-products for females of 80. Ten numbers got added up to give us the sum of cross-products for males of 40. You can't compare those two sums of cross-products to each other because ***you can't compare one sum to another sum when they're based on different numbers of values***. The sum of the cross-products can't handle this type of situation. But there's a simple adjustment that we can make to the sum of cross-products that can handle it.

You can't compare one sum to another sum when they're based on different numbers of values, but ***you can compare one mean to another mean, even if the sample sizes for these means are different from each other***. If you take the sum of the values that got added up (in this case, cross-products) and then divide that sum by the number of values that got added up you get a mean. In this case you'd get the mean of the cross-products. For the male and females groups we were just talking about we get...

$$\text{For males, } \frac{\text{Sum of Cross-Products}}{N} = \frac{40}{10} = 4.0$$

$$\text{For females, } \frac{\text{Sum of Cross-Products}}{N} = \frac{80}{20} = 4.0$$

The mean cross-product for the 10 males is 4.0 and the mean cross-product for the 20 females is 4.0! These two numbers are comparable to each other. So it turns out that the answer to original questions is that the strength of the relationship between the two variables is the same for males and females.

For statisticians, the name for the mean of a bunch of cross-products is the **covariance**. If you want to compare one group to another group in terms of the strength of the relationship between two variables and the two groups have different sample sizes, you can't use the sum of cross-products, but you can use the covariance. The equation for calculating the covariance is...

$$\text{Covariance} = \frac{\text{Sum of Cross-Products}}{N - 1}$$

If you remember, in our example using the data from five participants for the variables of social support and life satisfaction the sum of cross-products was 20. Taking this number and converting it to a covariance we get:

$$\text{Covariance} = \frac{\text{Sum of Cross-Products}}{N - 1} = \frac{20}{4} = 5.0$$

The covariance describing the relationship between social support and life satisfaction is 5.0.

The Pearson Correlation Coefficient

The covariance solves a problem that the sum of cross-products can't handle. It turns out that the Pearson r is able to solve a problem that the covariance can't handle. It allows a researcher to answer a kind of question that the covariance can't handle.

Let's say you're in the following situation: you want to know whether the strength of the relationship between social support and life satisfaction is different from the strength of the relationship between social support and depression. In other words, there are two different pairs of variables and you want to know which pair has the stronger relationship. Okay, that kind of makes sense. Now, think about what could happen if you decided to use the covariance as your measure of the strength of relationship. What you want is a situation where the *only* thing that influences the size of the measure you're using is the strength of the relationship between two variables; that way the relationship with the bigger number is the stronger relationship. But is that the case with the covariance? It turns out that the answer is no.

What are factors that influence the size of the covariance. One of the things that influences the size of the covariance is the strength of the relationship. Great, that's what we want. BUT, there's something else that can make the covariance a larger or smaller number that has nothing to do with the strength of the relationship. Think about it. What's happening when you calculate the sum of the cross-products. You're multiplying the deviation scores for one variable by the deviation scores for the other variable. If the range of scores for a variable is really small – say the scores range from 0 to 1.0. This is going to give you small numbers for the deviation scores for this variable. When you multiply these small deviation scores for one variable by the deviation scores for the other variable you're going to get smaller numbers than if the range of scores for the variable was really large and had really large deviation scores. So, the other thing that

influences the size of the covariance is the amount of variability you find in either of the variables involved. It's possible that the covariance between variables X and Y is a larger number than the covariance between variables X and Z not because the strength of the relationship is greater, but because the standard deviations for the combination of variables X and Z are larger than the standard deviations for the combination of variables X and Y.

So how can we get around this situation? How can we come up with a number that is only influenced by the strength of the relationship between two variables. It turns out to be simple. If the covariance between variables X and Y can be made larger or smaller because of the variability in these measures, we can negate this by taking the variability in variables X and Y into account. The equation for the Pearson r takes the covariance and divides it by the number you get when you multiply the standard deviation for variable X by the standard deviation for variables Y. So if the covariance is made a larger number because of larger variabilities in either X or Y, you balance this out by dividing by the larger number you get when you multiply the standard deviation for X by the standard deviation for Y.

$$\text{Pearson } r = \frac{\text{Covariance}}{(S_X)(S_Y)} = \frac{5.0}{(3.16)(1.79)} = \frac{5.0}{5.65} = .885$$

So, we would say that the correlation between the two variables is .885.

Interpreting the Pearson Correlation Coefficient

You know that negative correlation coefficients indicate the presence of negative relationships and that positive correlation coefficients indicate the presence of positive relationships. A correlation of zero is what you get when there is no relationship between the two variables and the further from zero the correlation is the stronger the relationship between the two variables. The sign of the correlation (positive or negative) has nothing to do with how strong the relationship between the two variables is. A correlation of +.65 is equally strong as a correlation of -0.65.

The Squared Pearson Correlation

But here's a different question. Does a correlation of +.30 mean that the relationship between the two variables is 30% of the way to a perfect relationship? In other words, in percentage terms, how strong is the relationship between the two variables? Are the two variables measuring the same thing to the extent of 30%? At first glance it would kind of make sense to say "Yes". After all, .30 is 30% of the way between zero and 1.0, the value for a perfect positive correlation. But it turns out that a correlation coefficient can't be

interpreted in this way. The percentage of the distance between zero and 1.0 or -1.0 does not describe the strength of the relationship in percentage terms. Bummer.

However, there is something simple we can do to a correlation coefficient that will give us this kind of information. Just take the correlation coefficient and *square it*. If we take the correlation we got just now of .885 and square it we get .78.

$$r^2 = (.885)^2 = .78$$

It turns out that, in some respects, the squared correlation between two variables is actually more useful and easy to interpret than is the original correlation. ***A squared correlation represents the proportion of overlap between two variables.*** In our example, if you take that proportion and multiply it by 100 you get 78%. This value lets you say that the two variables overlap by 78% -- or that they share the same information to the extent of 78% -- or that the two variables are measuring the same thing to the extent of 78%. You get the picture. The squared correlation is often reported in describing the strength of the relationship between two variables because people are much more comfortable in thinking about things in percentage terms.

So, to recap, if your instructor asks you if a correlation of +.30 means that the two variables overlap by 30% your answer should be “uh uh”. You should stall for time by asking them to repeat the question, figure out that .30 squared is .09, and then state with a look of clear intellectual superiority that the two variables overlap by 9%. Now turn to a friend, roll your eyes and say “duh”.

Testing the significance of a correlation coefficient

Okay, so you can compute a correlation coefficient or get a program like SPSS or SAS to give it to you. You know how to interpret it. BUT, is it real? Just because you get a correlation coefficient of +.2 or +.3, does that mean that there really is a relationship between the two variables out there in the real world?

What? What do you mean is it real? I mean, like, the correlation's not zero, right? If it's not zero then that means that there's at least some relationship between the two variables. That's true, if you're only talking about the people in the sample that gave you the data. But most researchers don't want to limit their conclusions to just the participants that make up the sample. We want to be able to talk about the relationship between the two variables in the *population*. Remember, what's the point of using a sample? It's to let you make a best guess about what's going on in the entire population.

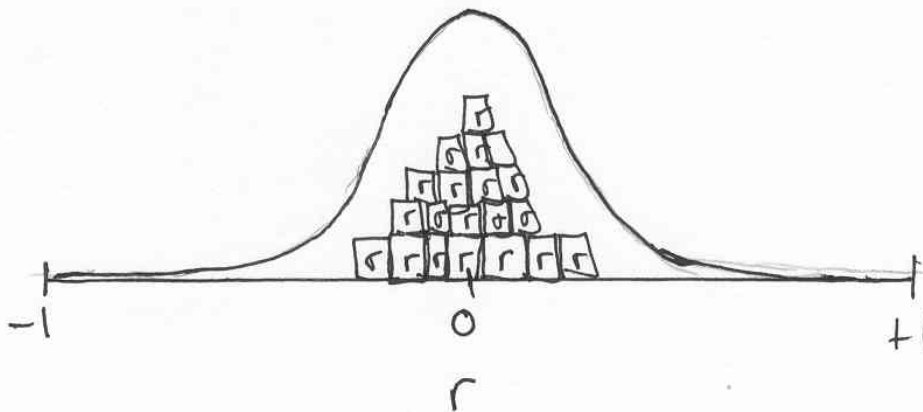
The thing you have to ask yourself is whether it's possible that the correlation between the two variables is something that could have just shown up by chance. When a researcher tests the significance of a correlation coefficient they're getting information about the odds that the correlation they got from their sample was only different from zero by chance. If you think about it, a correlation coefficient is a sample statistic, just like a sample mean. Just like the job of a sample mean is to give you an estimate of the

mean of a population, the job of a sample correlation coefficient to give you an estimate of the correlation for everyone in the population. By the way, the symbol for the correlation for a population is the Greek letter “ ρ ” (Rho).

You know that the value for r from your sample is aiming at some population value for ρ . The null hypothesis in testing the significance is that there is nothing going on in terms of a relationship between the two variables. ***The null hypothesis is that the correlation coefficient in the population is zero.*** Using an alpha level of .05, ***if we can show that the odds are less than 5% that this null hypothesis is true we are justified in rejecting this null hypothesis and accepting an alternative hypothesis that the correlation coefficient in the population is significantly different from zero.***

So how can we figure out whether the odds are less than 5% that the null hypothesis is true in this case. The answer in this situation is the same kind of answer we came up with when trying to see whether a sample mean is different from the mean of a population. We need to find out what sample correlation coefficients look like when they’re estimates of a population correlation of zero. We know that these estimates don’t have to be perfect; they just have to be unbiased (i.e., no more likely to underestimate than to overestimate the population correlation of zero). Using laws of probability, statisticians can tell us how sample correlations will fall on the scale when they’re all estimates of a population correlation of zero. Once we know where these correlations fall we’ve got something to compare our sample mean to. Displayed below in Figure X.8 is a hypothetical frequency distribution of correlation coefficients that are all estimates of a population correlation of zero – in other words, that were all collected when the null hypothesis is true.

Figure X.8



Now we’ve got something to compare our sample correlation coefficient to. We’ve got to decide whether our correlation coefficient of .89 belongs with this collection of other sample correlations or not. If we decide it does we’re saying that the null hypothesis is true. If we become convinced that it doesn’t belong in that set we’re deciding that our correlation must have been collected when the null hypothesis is false. Because we’ve already decided to use an alpha level of .05 we’re able to say that if a correlation falls

among the outer 5% of correlations that make up the curve then the odds must be less than 5% that it belongs in that collection. By the way, the name of this collection of sample correlations is the *sampling distribution of the correlation coefficient*. It's a collection of a very large number of sample correlation coefficients obtained when the null hypothesis is true.

Statisticians can tell us what the shape of this distribution of correlations looks like when the null hypothesis is true. For one thing, they can tell us that the shape of this pile of correlations depends on the sample size used to compute each correlation. Just like with the sampling distribution of the mean the smaller the sample size the more spread out the correlations tend to be around zero. This means that the smaller the sample size the flatter and more spread out the curve will be. This pushes the beginning of the outer 5% of the area under this curve further away from zero. In short, the smaller the sample size the larger a correlation coefficient has to be to say that it's significant.

Table XX gives us critical values for r for different sample sizes. To know which row of the table to look in you have to know the number of degrees of freedom for the test. When testing the significance of a correlation the number of degrees of freedom is equal to $N-2$. So, for our example we've got five subjects in the sample or three degrees of freedom. We're interested in knowing whether our correlation is significantly *different* from zero, so we're doing a two-tailed test. The critical values for r for a two-tailed test with an alpha level of .05 and three degrees of freedom are +.878 and -.878. Our correlation of .89 is further away from zero than that so we can say that it is significantly different from zero! We could report this result in APA format using a sentence along the lines of "The correlation between social support and life satisfaction was statistically significant, $r(3) = .89, p < .05$."

Experimental versus correlational research

Okay, we're willing to bet that the correlation of .89 is not a fluke. The odds are that there really is a relationship between social support and life satisfaction. Does that allow me to say that having higher levels of social support causes people to have higher levels of life satisfaction? It seems like it should. I mean, that seems like a pretty likely scenario. And in fact it's very possible that this is the case. But, unfortunately, a correlation coefficient does not give you enough information to draw this type of conclusion. After all, ***you can't draw a conclusion about cause and effect from a correlation***. You already know that. But why not? Why can you draw a conclusion about a cause and effect relationship using data from an experiment, but not from a correlational study?

There are a couple of ways of approaching this question. First of all let's think about what an experiment has going for it. In a simple experiment a researcher might randomly assign each subject to one of two groups. At that point there's nothing that makes the subject in one group different from the subjects in the other group. Now the researcher does something to make the subjects in one group different from another one. The researcher has changed the conditions in one way – they've manipulated an independent

variable. And then they compare the means for the two groups on a second variable, the dependent variable. If the researcher does a t-test and finds a significant differences between the two groups there is only one possible explanation for this difference. The only way in which the two groups are different from each other is in terms of the one difference that the researcher put there – the independent variable. So the independent variable must be responsible for – must have caused – the difference between the two groups on the dependent variable.

In addition to that, in order to establish that one variable is causing an effect on another variable the researcher has to be able to know which variable occurred first. After all, it seems pretty reasonable to expect that causes happen before effects. To know which variable is the cause of an effect you've got to be able to know which variable occurred first. That's not a problem with an experiment because we know that the researcher changed the conditions first (they manipulated the independent variable) and then they saw what happened (they measured people's scores on the dependent variable).

Before → After
Cause → Effect
I.V. → D.V

However, with a correlational study it is impossible to know which variable came first because both variables are measured at the same time. In our example, it's possible that some subjects had higher level of social support (variable X) first and that later on this caused them to have higher levels of life satisfaction (variable Y).

Social Support (X) → Life Satisfaction (Y)
Cause → Effect

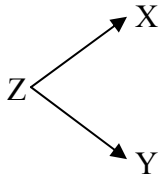
There is nothing in our data to rule out this possibility. However, there is also nothing in our data to rule out the possibility that higher scores for life satisfaction cause people to have more extensive supportive social networks.

Life Satisfaction (Y) → Social Support (X)
Cause → Effect

X could be having a causal effect on Y or Y could be having a causal effect on X and there's no way to tell which of these scenarios is responsible for the relationship between the two variables. In this sense, a correlation coefficient is nothing more than a descriptive statistic. It describes the relationship between the two variables, but it doesn't explain the source of the relationship – it's not able to show you why that relationship is there.

And to make matters worse, there's yet another reason for why you can't draw a causal conclusion from a correlation coefficient. It might be that neither scenario described above represents the cause for our statistically significant correlation. It might be that X

doesn't act causally on Y and that the same time that Y doesn't act causally on X. The cause of the relationship may be due to the causal influence of a third variable (Z) on both X and Y at the same time. In other words, the relationship between X and Y isn't there because X was causing higher or lower scores on Y or because Y was causing higher or lower scores on X. X and Y may not be acting causally on each other at all. The third variable may be influencing the scores for variable X at the same time that it's influencing the scores for variable Y.



The problem with this “third variable” scenario is that the researcher may not have measured scores for this third variable, in which case the researcher has no way of knowing the true cause of the relationship between X and Y. This issue is often referred to as the **Third Variable Problem** in interpreting a correlation coefficient.

So, there are really two reasons for why a researcher can't draw a conclusion about cause and effect from a correlation. First, the researcher isn't able to figure out which variable is the cause and which one is the effect. Second, you can never rule out the possibility that a third variable is the true cause of the relationship between X and Y.

Examples of “third variables”.

I've got two examples of third variables that I use in my lectures. They're both a little on the morbid side. Oh well.

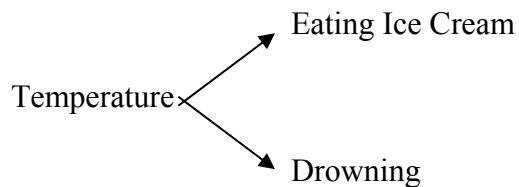
In the first example, there is a strong positive correlation between the amount of ice cream eaten in a given day at the beach and the number of drownings on those same days. That's a little dark. But, where does this significant positive correlation come from? There are several possibilities. Let's say that the amount of ice cream eaten is variable X and that the number of drownings is variable Y. It's possible that your grandmother was right and that eating ice cream causes every muscle in your body to seize up the moment you hit the water, thus causing you to sink to the bottom and drown. In this case variable X is having a causal effect on variable Y.

Eating Ice Cream → Drowning

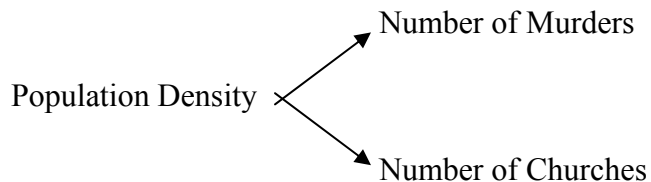
Alternatively, one might try to argue that the causal relationship goes the other way around and that drowning causes people to eat more ice cream. In this case variable Y is said to exert a causal effect on variable X. That seems a little less likely in this case.

Drowning → Eating Ice Cream

Third, it's possible that a third variable is causing there to be higher scores on both the amount of ice cream eaten and the number of drownings at the same time. In this case a likely third variable would be the temperature of the day. Think about it. On hot days more people want to eat ice cream and on hot days there will be more people going into the water to swim, thus resulting in more people drowning. Ice cream eating and drowning don't act causally on each other at all. The third variable, the temperature of the day, is influencing the scores for both the other variables at the same time.



In the second example, there is a strong positive correlation between the number of churches there are in a given neighborhood and the number of murders there are in those same neighborhoods. Let's say that the number of churches is variable Y and that the number of murders is variable X. It's possible that variable X is having a causal effect on variable Y. In other words, it's possible that committing an act of murder (X) causes people to go to church (out of guilt, perhaps), resulting in a greater need for churches (Y). Alternatively, it's possible that going to church (Y) causes people to go out and commit murder (X). Both of these scenarios seem unlikely. However, as you might have surmised by now, there's a third variable that's responsible for higher scores on both of these variables occurring at the same time. The third variable in this case is the density of the population in these neighborhoods. In densely populated neighborhoods there tend to be more murders and densely populated neighborhoods also tend to have more churches.



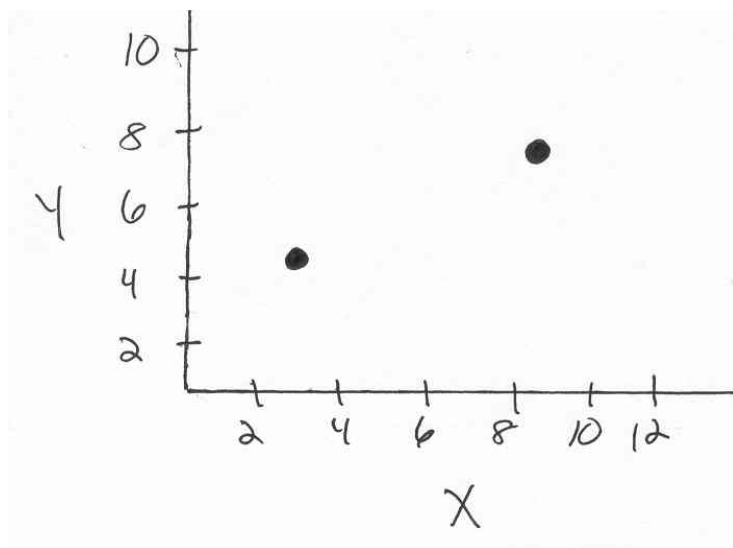
The variables of murder rate and the number of churches don't act causally on each other at all. The population density is driving the scores for both of the other variables higher at the same time.

Experiments and correlational studies answer different kinds of questions. An experiment can provide information about whether changes in one variable cause changes in the scores for another variable. A correlational study is able to describe the relationship between two variables. The important thing for the researcher is to understand the conceptual limitations of the conclusions they're entitled to draw from the study they've conducted.

The Adjusted Pearson Correlation

Got a strange question for you: What's the fewest number of participants you could have and still compute a correlation coefficient? *Hmmm. How about two.* Okay, let's say you collect data from two subjects on two variables. From the scatterplot you determine that the relationship between X and Y is positive. What's the correlation between X and Y? *You mean tell you the number, the actual correlation coefficient?* Yup. What's the actual number. *How are we supposed to know that?! We don't even know the scores for those variables.* You don't need them! Think! Think about the scatterplot for the two variables. Let's say it looks like this...

Figure X.9



There are two points in the scatterplot. The one on the right is higher than the one on the left so we've got a positive relationship. Back towards the beginning of the chapter we talked about the relationship between the strength of the relationship between two variables and the pattern you see in a scatterplot of those two variables. We talked about how easy it looked like it would be to draw a straight line through the points in the scatterplot. The closer the points in a scatterplot come to falling on a straight line the closer the strength of the relationship comes to being perfect. In a sense, you can think of **a correlation coefficient as a straight line detector**. The closer the correlation is to +1.0 or -1.0 then closer the points in the scatterplot come to falling on a straight line.

So what happens when you go to draw a straight line through the points in this scatterplot? That straight line runs right through the only two points that are there! A straight line provides a perfect fit to the data and that corresponds to a situation where the correlation between the two variables is perfect. We know that the correlation between the two variables in this case has to be +1.0.

When there are only two subjects in the data set there are only three possible values that a correlation coefficient can take: +1.0 if the line is going up, 0.0 if the line is flat, and -1.0

if the line is going down. This seems kind of strange doesn't it? Doesn't it seem like a correlation ought to be able to detect a relationship that falls somewhere between no relationship at all and perfect? It turns out that there is a bias associated with the Pearson correlation coefficient. Remember, the job of a sample correlation coefficient is to give you an unbiased estimate of the correlation between the two variables in the population. By unbiased we mean that the sample correlation is no more likely to overestimate than to underestimate the actual correlation in the population. The Pearson correlation, especially at low sample sizes, is more likely to be too high than it is to be too low. When the sample size is as small as it could be – two subjects – the correlation jumps from zero to perfect. At sample sizes greater than this the problem isn't as severe, but the sample correlation still gets pushed further away from zero than it really should be.

Fortunately, there is a fairly simple adjustment a researcher can make to a correlation that will give them a number that corrects for this bias. The equation below is an equation for an adjusted correlation.

$$\text{Adjusted correlation} = \sqrt{1 - \frac{(1 - r^2)(N-1)}{N-2}}$$

The adjustment part of this equation comes from the ratio of N-1 to N-2. If you leave these elements out of the equation you end up with the same correlation you started with (except that the result always comes out as a positive number). The fact that the degree of adjustment comes from the ratio of N-1 to N-2 tells us that there isn't all that much to worry about when the sample size is decently large. For example, if there were 100 subjects in the sample the ratio of 99 to 98 is so close to 1.0 that the original correlation will hardly be "adjusted" at all. However, with small sample sizes the degree of adjustment is enough to make the procedure worth considering.

If you recall, the correlation we calculated was .885. Applying the equation for an adjusted correlation we get

$$\text{Adjusted correlation} = \sqrt{1 - \frac{(1 - .78^2)(5-1)}{5-2}} = .865$$

This value of .865 is the best guess we can make as to the strength of the correlation coefficient in the population.

End remarks

Correlational studies are very common in psychology and the behavioral sciences and are highly influential in the development of hypotheses to explain human behavior. In light of the ethical limitations placed on behavioral scientists in manipulating the conditions

under which data are collected, correlational research is often the only source of empirical information on a given topic. Furthermore, as we'll see in the next chapter, establishing a relationship between two variables is an essential first step in using the scores from one variable to predict a person's score on another, unmeasured, variable.