Chapter 13

STRENGTH OF RELATIONSHIPS: CONTINUOUS DATA

"Variance" is a technical word for "variability." It is a measure of spread about the mean. The variance of a population is the mean of the squared deviations from the mean:

$$\sigma^2 = \frac{\sum (x_i - \overline{x})^2}{n} = \frac{\sum d_x^2}{n}$$

Q: Why square the deviations?

A: If you just add the deviations up without squaring them, the total will always be zero, and this won't tell you anything at all. Squaring them removes the minus signs, among other things.

To get the *standard deviation*, you take the *square root of the variance*This compensates for the fact that you squared the deviations before adding them up.

If you are talking about how many cups of coffee per day people consume, the standard deviation is in units of cups-per-day. So if the mean is 5 cups per day, the standard deviation is 2 cups, and the distribution is normal, 68% of everyone drinks between 3 and 7 cups per day.

If you also have age as a variable and you find that the standard deviation of age is 5 years, you will have a problem if you want to compare age with coffee consumption. For one thing, the measures of spread (standard deviation) for the two variables are in different units. For coffee, the units are cups of coffee per day. For age, the units are years of life. How do you compare cups of coffee per day to years of life?

I'm going to make two changes to the equation for variance. The first change allows you to get a measure of how much of the variance in one variable is shared by another variable, or, in other words, how closely associated or tied together are the two variables. The second change will get rid of the units the variables are measured in (cups, years, ... whatever) by changing the original units into standardized units. This change will make it possible to use the measure of shared variance to compare the strength of the relationship between different pairs of variables.

The First Change: Covariance

If you want to measure how the two variables vary together—"covariance"— you make a simple change in the equation for variance. Instead of calculating the mean of the *squared* deviation scores like this:

$$\text{var}_x = \frac{\sum (x_i - \bar{x})^2}{n} = \frac{\sum d_x^2}{n} = \frac{\sum d_x\, d_x}{n}$$

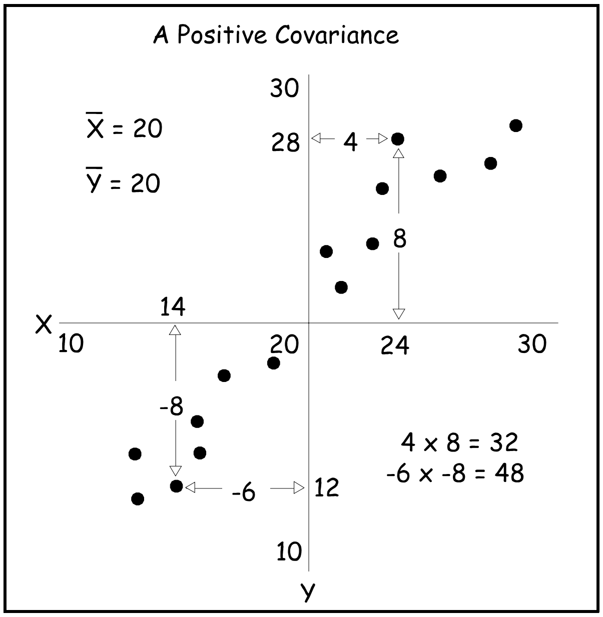
you calculate the mean of the *cross-products* of the deviation scores. That is, instead of multiplying each deviation score by itself, you multiply it by the corresponding deviation score of the second variable:

$$\text{cov}_{xy} = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{n} = \frac{\sum d_x d_y}{n}$$

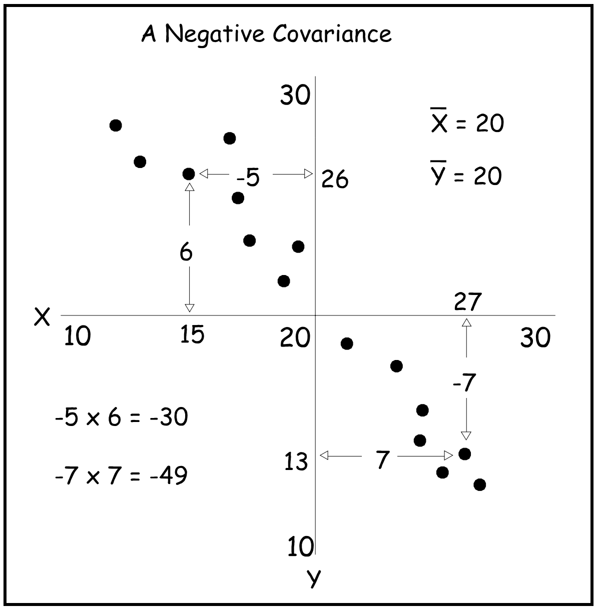
Look closely at the equation for variance and compare it with the one for covariance. The only difference is that variance multiplies each deviation score by itself, while covariance multiplies the deviation score of the *first* variable by the corresponding deviation score of the*second*variable. Everything else is the same.

One consequence of combining the two variables' deviation scores like this is that the sum of the products may be either negative or positive.

If the first variable is *below* the mean whenever the second one is *also* below the mean, both deviation scores will be negative, and the product of the two will be positive. If the first variable is *above* the mean whenever the second one is also above the mean, both deviation scores will be positive, and the product of the two will be positive. The sum of the products will thus be positive, which means the covariance will be positive. An example of this kind of situation is shown in the drawing on the left.


A Positive Covariance

On the other hand, if the first variable is *below* the mean when the second one is *above* the mean, or the first variable is *above* the mean when the second one is *below* the mean, the products of the deviation scores will be negative, and the covariance will be negative, as you can see in the drawing below.


A Negative Covariance

A positive covariance means that the two variables co-vary (vary together) in a positive way: when one is high, so is the other; when one is low, so is the other. A negative covariance means that the variables covary in a negative or inverse way: when one variable is high, the other one is low; when one is low, the other is high. For any given pair of variables measured in a particular way, the larger the covariance is, the stronger the relationship between the variables is.

An example:

Imagine that you are a toy dealer doing some market research on various toys. This time you are particularly interested in xylophones and yo-yo's. You want to see whether there is a relationship between the number of xylophones and the number of yo-yo's a typical child owns. You happen to be studying covariance in a statistics course you are taking, and you see this as an opportunity to put your knowledge to good use.

You decide to count the number of xylophones and yo-yo's owned by each child in your neighborhood and to call the variables "X" (number of xylophones) and "Y" (number of yo-yo's), and to calculate the covariance of the two variables.

Here is the equation for covariance:

$$\text{cov}_{xy} = \frac{\sum d_x d_y}{n}$$

In this equation, d_x and d_y are deviation scores for the two variables. The data and all relevant calculations are summarized in the table below.

i	*X*	d_x	*Y*	d_y	$d_x d_y$
1	2	-3.4	1	-4.9	16.660
2	2	-3.4	3	-2.9	9.860
3	3	-2.4	6	.1	-.240
4	4	-1.4	5	-.9	1.260
5	5	-.4	3	-2.9	1.160
6	6	.6	8	2.1	1.260
7	7	1.6	6	.1	.160
8	7	1.6	10	4.1	6.560
9	8	2.6	8	2.1	5.460
10	10	4.6	9	3.1	14.260
Tot	54	0.0	59	0.0	56.40

$$\bar{x} = 5.4 \qquad \bar{y} = 5.9 \qquad n = 10$$

$$\sum d_x d_y = 56.40$$

$$\text{cov}_{xy} = \frac{\sum d_x d_y}{n} = \frac{56.40}{10} = 5.640$$

The covariance is 5.640, which indicates that there is a positive relationship between the two variables. The more xylophones a child in your neighborhood owns, the more yo-yo's the child is likely to own. You could compare this covariance with one calculated with the same variables next year and see whether the relationship between the two variables has become stronger or weaker, but you couldn't use it to compare the relationship between xylophone and yo-yo ownership with the one between age and coffee consumption. The covariance doesn't tell how strong the relationship is in terms of how many of the children in the neighborhood fit the pattern nor in terms of how many more yo-yo's a child with three xylophones is likely to have than a child with only one or two.

An easier way to calculate covariance

Instead of using this formula for covariance,

$$\text{cov}_{xy} = \frac{\sum d_x d_y}{n}$$

you could use this one:

$$\text{cov}_{xy} = \frac{\sum x\,y - \frac{\sum x \sum y}{n}}{n}$$

The advantage of this *computational* form of the equation is that you do not have to calculate sample means and deviation scores first. It takes less work, there is less rounding error, and you are more likely to get the correct answer. Try it with the data shown in the table on the left and see how it compares. Compare the above equation with the computational formula for variance and make note of the differences and similarities of the two:

$$\text{var}_x = \frac{\sum x^2 - \frac{\left(\sum x\right)^2}{n}}{n}$$

The Second Change: Correlation

The units you use to measure amount of coffee consumed and the person's age are easy to understand —cups and years. But covariance is measured in strange units. If you used covariance to assess the relation between age and number of cups of coffee consumed per day, you would multiply each person's deviation score for amount of coffee (*cups*) by the person's deviation score for age (*years*). The covariance units for this example would be something like "cup-years"—a strange unit indeed. Think about the units you'd be using in the xylophone & yo-yo example.

If you had another variable that might be related to coffee consumption, say height, you could calculate the covariance of the relation between height and cups of coffee per day. But you could not com-pare this to the covariance of coffee and age to see which relationship is stronger, because the second covariance would be in units of "cup-inches" or "cup-centimeters" which, sadly, are not comparable to "cup-years." This is the first problem with covariance.

The second problem with covariance is that it is strongly influenced by the variance of the variables you are working with. A variable whose values are spread far away from the mean will have large deviation scores, while one whose values are clustered closely around the mean will have small deviation scores. This means that you will be likely to get larger covariances when your variables have higher levels of dispersion than when the dispersion is low, regardless of the strength of the relationship between the variables. In other words, the covariance may, to an extent, be a function of the dispersion of your variables as much as a measure of the strength of the relationship between the variables. You can eliminate this problem if you can standardize your variables so they always have the same level of dispersion. This turns out to be an easy thing to do.

If you turn both height and number of cups into standard scores (*z*-scores), the resulting variables will both have means of 0.0 and standard deviations of 1.0. *Z*-scores are in units of "standard deviations," and they can be compared to one another, unlike the original variables. You standardize your variables (convert deviation scores to *z*-scores) by dividing by the standard deviation.

So you make an adjustment to covariance. Instead of using the sum of the cross-products of the *deviation* scores, you use the sum of the cross-products of the *z-scores*. Although using the deviation scores centers your data and removes the effect of the mean (by subtracting it from each value), converting the deviation scores to *z*-scores removes both the effects of the original units of measurement and the degree of dispersion of the variable. You change from

$$\frac{\sum d_{x_i} d_{y_i}}{n} \quad \text{to} \quad \frac{\sum z_{x_i} z_{y_i}}{n}$$

You could achieve the same result by dividing the covariance by the product of the standard deviations of the variables (calculated with *n*). This does the same as converting the deviation scores to *z*-scores:

$$\frac{\text{cov}_{xy}}{s_x s_y} = \frac{\frac{\sum d_x d_y}{n}}{s_x s_y} = \frac{\sum d_x d_y}{n s_x s_y}$$

and . . .

$$\frac{\sum d_x d_y}{n s_x s_y} = \frac{\sum \frac{d_x}{s_x} \frac{d_y}{s_y}}{n} = \frac{\sum z_x z_y}{n}$$

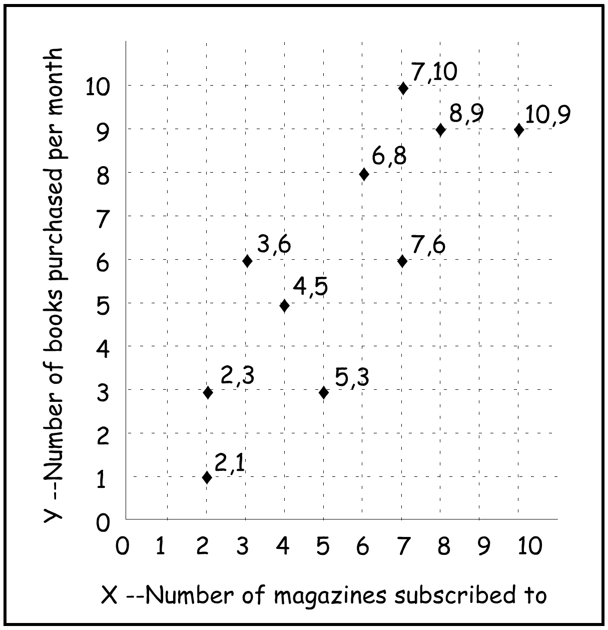
Changing from deviation scores to *z*-scores produces an important result. The value you get now is the *Pearson product-moment correlation coefficient*, more commonly known as the "correlation." The correlation can have values between -1.0 and 1.0. If the correlation is 1.0, the two variables are perfectly correlated with one another. They are in effect interchangeable, for when you know the value of one, you also know the value of the other. If the correlation is -1.0, they are perfectly correlated, but the relationship is negative—when one is high, the other is low. If the correlation is 0.0, there is no relationship at all between the two variables; knowing the value of one tells you nothing about the value of the other.

You can compare any correlation with any other correlation, for they are all in the same units. The correlation coefficient is one of the most commonly used descriptive statistics. It provides an elegant and very useful summary description of the strength of the relationship between a pair of variables.

Why correlation is better than covariance

- Correlation is better because *you can compare any correlation to any other correlation and see which is stronger*. You cannot do this with covariance. This works because correlation uses *z*-scores instead of deviation scores. Doing this replaces the original units in which the variables were measured with standard deviations. This replacement a) eliminates the strange hybrid units you get with covariance and b) removes the effect of the variance of the original variables, resulting in a standardized, absolute measure of the strength of the relationship, bounded by -1.0 and 1.0.

- If you square a correlation, you get r^2, a measure of how much of the variance in one variable is explained by the other variable. You cannot do this with covariance. This measure, *the coefficient of determination*, ranges from 0.0 to 1.0.



Calculating *r*

Say you have a group of ten people. For each of them you know how many magazines they subscribe to (X) and how many books they purchase per month (Y). The data for this example is plotted on the right. You would like to know if the people who subscribe to a lot of magazines also purchase a lot of books. This is an appropriate situation to illustrate the use of the correlation. Three ways to calculate *r* are demonstrated below. The first is the definitional formula based on *z*-scores, the second uses deviation scores, and the third is the computational formula.

1) r **based on** *z***-scores**

The Pearson product moment correlation coefficient is defined as

$$r_{xy} = \frac{\sum z_x z_y}{n}$$

In this equation, z_x and z_y are *z*-scores for the two variables *x* and *y*; *n* is the sample size. While this is the simplest complete definition of the correlation, it is probably the most difficult one to use, because it uses *z*-scores, each of which involves a subtraction and a division. First you calculate the means, then the

Important Notes:

1. When using the computational formula, keep life simple—*don't use (n-1) in either the numerator or the denominator.*

2. When using the formula based on deviation scores, if you use (*n-1*) to calculate the standard deviations in the denominator, you must also use (*n-1*) in the denominator. If you use *n* to calculate the standard deviation, you must also use *n* in the denominator.

3. When using the formula that divides covariance by the product of the standard deviations, you must use *n* to calculate the standard deviations and covariance.

4. When using the formula based on *z*-scores, if you use (*n-1*) to calculate the standard deviations for these scores, you must also use (*n-1*) in the denominator. If you use *n* to calculate the standard deviation, you must also use *n* in the denominator.

deviation scores and the standard deviations. Next you divide the deviation scores by the standard deviations to get the *z*-scores which you multiply to get the cross-products in the last column. The sum of the cross-products goes in the numerator of the equation for *r*. As you can see in the example here, it is difficult to do these calculations without introducing rounding errors.

i	*X*	d_x	z_x	*Y*	d_y	z_y	$z_x z_y$
1	2	-3.4	-1.340	1	-4.9	-1.767	2.367
2	2	-3.4	-1.340	3	-2.9	-1.046	1.401
3	3	-2.4	-.946	6	.1	.036	-.034
4	4	-1.4	-.552	5	-.9	-.325	.179
5	5	-.4	-.158	3	-2.9	-1.046	.165
6	6	.6	.236	8	2.1	.757	.179
7	7	1.6	.630	6	.1	.036	.022
8	7	1.6	.63	10	4.1	1.478	.932
9	8	2.6	1.025	8	2.1	.757	.776
10	10	4.6	1.813	9	3.1	1.118	2.026
Tot	54	0.0		59	0.0		8.014

$$\bar{x}=5.4 \quad s_x=2.5377 \quad \bar{y}=5.9 \quad s_y=2.7731 \quad n=10$$

$$r_{xy}=\frac{\sum z_x z_y}{n}=\frac{8.014}{10}=.8014$$

2) *r* based on deviation scores

The formula below is a bit more complicated than the one that uses *z*-scores, but it is a lot easier to use. With this method you use deviation scores instead.

$$r_{xy}=\frac{\sum d_x d_y}{n\, s_x\, s_y}=\frac{\text{cov}_{xy}}{s_x\, s_y}$$

Here, d_x, d_y, s_x, and s_y are deviation scores and standard deviations for the two variables. This approach and the one that uses *z*-scores should give you the same answer. This one is easier to calculate than the one with *z*-scores, but it is still a lot of work and prone to errors because it needs deviation scores. Note that you must use *n* and not (*n–1*) for the standard deviations with this method.

i	*X*	d_x	*Y*	d_y	$d_x d_y$
1	2	-3.4	1	-4.9	16.660
2	2	-3.4	3	-2.9	9.860
3	3	-2.4	6	.1	-.240
4	4	-1.4	5	-.9	1.260
5	5	-.4	3	-2.9	1.160
6	6	.6	8	2.1	1.260
7	7	1.6	6	.1	.160
8	7	1.6	10	4.1	6.560
9	8	2.6	8	2.1	5.460
10	10	4.6	9	3.1	14.260
Tot	54	0.0	59	0.0	56.40

$$\bar{x}=5.4 \quad s_x=2.5377 \quad \bar{y}=5.9 \quad s_y=2.7731 \quad n=10$$

$$r_{xy}=\frac{56.40}{10\times 2.5377\times 2.7731}=.8014$$

The above calculations show that it is easier to work with deviation scores than with *z*-scores if you already have the standard deviations. You can see, though, that the result is the same.

3) the computational equation for *r*

A computational form of the equation for correlation, somewhat similar to the one for standard deviation, makes it even easier to calculate *r*. Although the equation looks a lot more complicated than the original one, you will find it *much* easier to use. Once again, the advantage of the computational equation is that it requires fewer calculations that usually produce fractions and messy numbers that have to be rounded. The result is greater accuracy and fewer errors.

Here is the computational equation for *r*:

$$r_{xy} = \frac{n\sum xy - \sum x \sum y}{\sqrt{n\sum x^2 - (\sum x)^2}\sqrt{n\sum y^2 - (\sum y)^2}}$$

Besides the raw data, this equation requires five sums which are fairly easy to calculate. If your raw data values are whole numbers, there will be no fractions in any of these sums making this an easy task to complete:

$\sum x$ — the sum of the *x* values

$\sum y$ — the sum of the *y* values

$\sum x^2$ — the sum of the squared *x* values

$\sum y^2$ — the sum of the squared *y* values

$\sum xy$ — the sum of the cross-products of the *x* and *y* values

i	*X*	*X*²	*Y*	*Y*²	*XY*
1	**2**	4	**1**	1	2
2	**2**	4	**3**	9	6
3	**3**	9	**6**	36	18
4	**4**	16	**5**	25	20
5	**5**	25	**3**	9	15
6	**6**	36	**8**	64	48
7	**7**	49	**6**	36	42
8	**7**	49	**10**	100	70
9	**8**	64	**8**	64	64
10	**10**	100	**9**	81	90
Tot	54	356	59	425	375

$$r_{xy} = \frac{(10\times 375) - (54\times 59)}{\sqrt{(10\times 356) - 54^2}\sqrt{(10\times 425) - 59^2}} = .8014$$

Compare the numbers in the above table with the ones for the first two methods for calculating *r*. Note that the numbers used with this method are all whole numbers up to the point where you plug them in the equation.

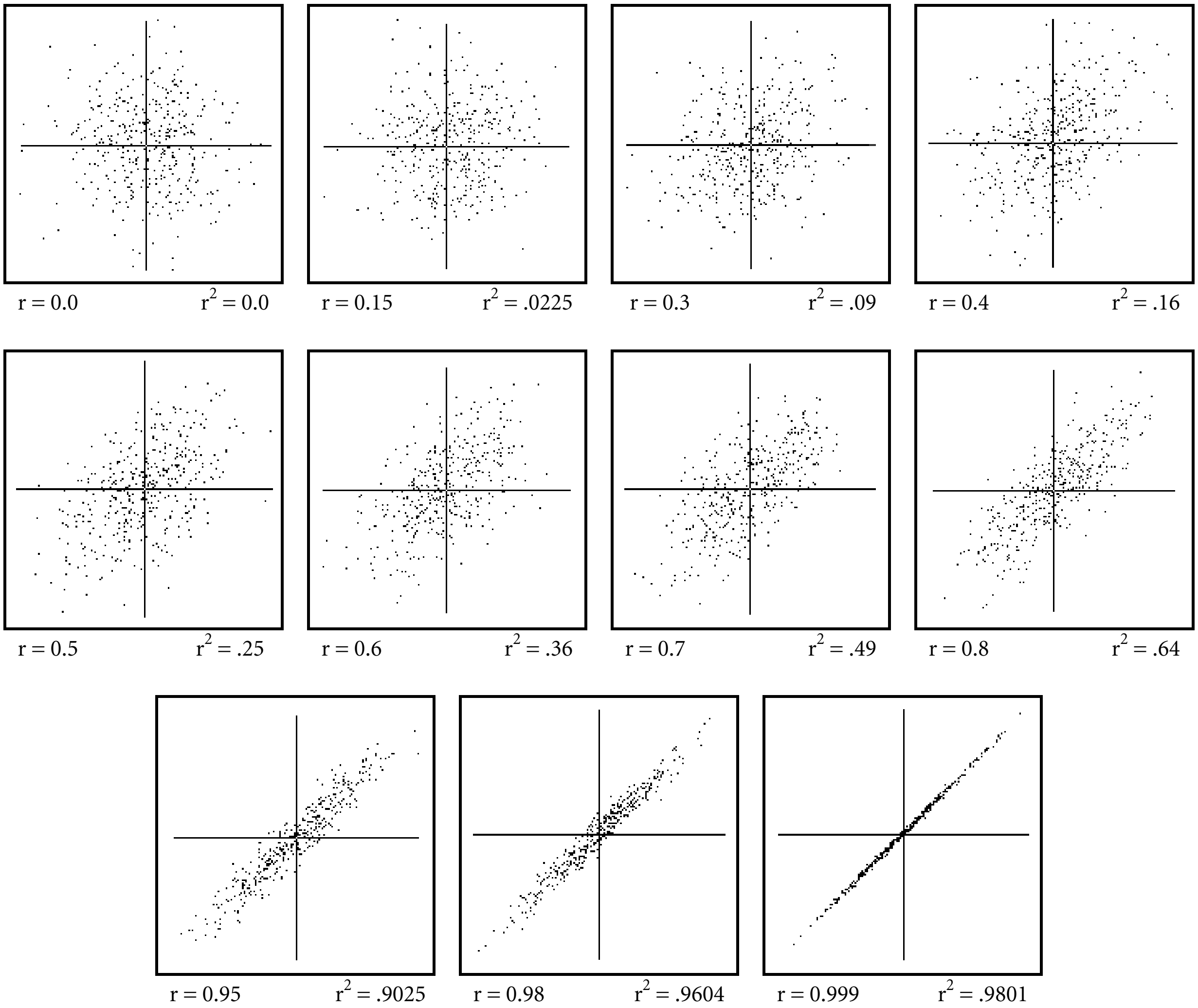
I once asked a class of 35 students to calculate correlations with the data in the examples here using the first method. I counted 26 (!!) different answers. Most of the variation was due to differences in the way the students rounded the intermediate values in their calculations. Even after they received careful instructions on rounding procedures, their results deviated from the correct answer by as much as ten percent. In contrast, almost all students who used the computational formula got very close to the correct answer. For them, most differences were either the result of overzealous rounding of the final answer or the product of arithmetic errors.

A Gallery of Correlations

On the next page are scatterplots for data with correlations ranging from 0.0 to 0.999 so you can get an idea of what various correlations look like. In each of these plots, both variables are normally distributed and have the same variance and mean. The only thing that differs from plot to plot is the correlation between the variables.

You probably will have difficulty seeing any difference between the scatterplots for correlations of 0.0, 0.15, and 0.3. With a correlation of 0.4, the pattern in the scatterplot is noticeable but certainly not something to write home about. The r^2 values tell you that these correlations account for 0%, 2.25%, 9%, and 16% of the variance. The scatterplot for $r = 0.4$ shows you how little 16% of the variance is.

A gallery of correlations



r = 0.0 $r^2 = 0.0$

r = 0.15 $r^2 = .0225$

r = 0.3 $r^2 = .09$

r = 0.4 $r^2 = .16$

r = 0.5 $r^2 = .25$

r = 0.6 $r^2 = .36$

r = 0.7 $r^2 = .49$

r = 0.8 $r^2 = .64$

r = 0.95 $r^2 = .9025$

r = 0.98 $r^2 = .9604$

r = 0.999 $r^2 = .9801$

Important Terms and Concepts

coefficient of determination
computational equation
correlation
covariance
cross-products
deviation scores
negative relationship

Pearson's *r*
positive relationship
r^2
standard scores
variance
z-scores