12.6: The Bivariate Correlation Formula

The correlation coefficient is used to summarize the relationship between two quantitative variables in a dataset using a number ranging from -1.00 to 1.00. Recall that variables, by definition, vary. Covariance, therefore, refers to the extent to which two variables vary together in a patterned way. When a correlation is perfect, it means that X and Y have a perfect pattern of covariance such that the amount of covariance is equal to the total of the individual variability of the two variables. Thus, the correlation formula is assessing the proportion of shared variance to unshared variance. When r is stronger, it means a greater proportion of the total variance is shared. When the correlation coefficient is 1.00 it means that all of the variance is shared between the variables. When the correlation coefficient is .00 it means that all the variance is attributed to the two variables separately and is, thus, not indicative of a pattern or relationship between the variables. To summarize, we can understand the formula's main construction and outcomes as follows:

$$r = \frac{\text{how greatly } X \text{ and } Y \text{ vary together}}{\text{how greatly } X \text{ and } Y \text{ vary separately}} = \text{ ratio of covariance to random variance}$$

When this formula is written out to summarize the computational elements, it is written as follows:

$$r = \frac{\Sigma(X - \bar{X})(Y - \bar{Y})}{\sqrt{\Sigma(X - \bar{X})^2}\sqrt{\Sigma(Y - \bar{Y})^2}}$$

The numerator of the formula focuses on deviations of each variable (which are a central part of variance) and their connections. The denominator of the formula looks at the deviations for each variable separately. When put together, the result indicates how much of the variance was shared verses separate. The greater the magnitude of final result, the stronger the covariance and, thus, the stronger the relationship is between the two variables.

Notice that the sum of squares within formulas have, once again, made their way into an inferential formula. You can see $\Sigma(X - \bar{X})^2$ in the left side of the denominator which is the sum of squared deviations for the X values. You can also see $\Sigma(Y - \bar{Y})^2$ in the right side of the denominator which is the sum of squared deviations for the Y values. Thus, the correlation formula can also be written as follows by replacing those two SS formulas with the symbol for their respective SSs, as follows:

$$r = \frac{\Sigma(X - \bar{X})(Y - \bar{Y})}{\sqrt{SS_X}\sqrt{SS_Y}}$$

Formula Components

Now that we have taken some time to understand the basic construction of the correlation formula, let's focus on how to actually use it, starting with identifying all of its parts.

In order to solve for r, we need three things:

$\Sigma(X - \bar{X})(Y - \bar{Y})$ = the sum of the products of the deviations from X and Y

$\Sigma(X - \bar{X})^2$ = sum of squared deviations for the X-values

$\Sigma(Y - \bar{Y})^2$ = sum of squared deviations for the Y-values

Formula Steps

The steps are shown in order and categorized into two sections:

A. preparation and
B. solving.

Preparation steps for correlation include finding the mean for X and the mean for Y. Then, these are used in section B to find deviations needed for the three main formula components listed above. Once those components are known and plugged into the formula, order of operations is followed to yield the obtained value (known as r) for the formula. Follow these steps, in order, to find r:

Section A: Preparation

1. Find $\bar{X}$ (the mean for the X-variable scores)
2. Find $\bar{Y}$ (the mean for the Y-variable scores)

> **Note**
>
> Though we do not need to find n for this version of the formula, it is good to make note of what it is because it will be used to find the degrees of freedom (df) later.

Section B: Solving

The values from the preparatory steps must now be used to find the three main components of the correlation formula before the r-value can be computed.

1. Find $\sqrt{\Sigma(X-\bar{X})^2}$
   a. Find $(X-\bar{X})$ by subtracting the mean of X from each x-value (these are the deviations for each X)
   b. Find $\Sigma(X-\bar{X})^2$ by squaring each deviation and then summing those values (which yields the sum of squares for the X-variable, also known as SS_X)
   c. Square root the SS_X.
2. Find \(\sqrt{\Sigma(Y-\bar{Y})^2\)
   a. Find $(Y-\bar{Y})$ by subtracting the mean of Y from each y-value (these are the deviations for each Y)
   b. Find $\Sigma(Y-\bar{Y})^2$ by squaring each deviation and then summing those values (which yields the sum of squares for the Y-variable, also known as SS_Y)
   c. Square root the SS_Y.
3. Find $\Sigma(X-\bar{X})(Y-\bar{Y})$
   a. Find $(X-\bar{X})(Y-\bar{Y})$ by multiplying each deviation from X by its deviation from Y to get the product of the deviations for each case
   b. Find $\Sigma(X-\bar{X})(Y-\bar{Y})$ by summing the product of deviations for each case (i.e. sum the results from step 3a). This is the numerator for the formula.
4. Find the denominator by multiplying the square root of SS_X (which is the result of step 1c) by the square root of SS_Y (which is the result of step 2c)
5. Divide the numerator (which is the result of step 3b) by the denominator (which is the result of step 4).

Reading Review 12.2

1. What is the focus of the numerator of the correlation formula?
2. What is the focus of the denominator of the correlation formula?
3. Which two descriptive statistics should be found in preparation for using the correlation formula?
4. What are the steps to calculating SS_X?
5. What are the steps to calculating SS_Y?

Example of How to Test a Hypothesis Using Correlation

Let's test the hypothesis and Data Set 12.1, which was introduced earlier in this chapter. We supposed a researcher collected data from 10 college students to test the hypothesis that sleep would positively relate to quiz scores. Assume that Data Set 12.1 includes data from the aforementioned sample. Let's follow the steps in hypothesis testing using these data.

Data Set 12.1. Hours of Sleep and Quiz Scores (n = 10)

Participant Number	Sleep Hours	Quiz Score
1	7	92
2	8	88
3	9	96
4	6	70

Participant Number	Sleep Hours	Quiz Score
5	6	79
6	4	64
7	5	75
8	10	98
9	3	53
10	7	85

Steps in Hypothesis Testing

In order to test a hypothesis, we must follow these steps:

1. **State the hypothesis.**

A summary of the research hypothesis and corresponding null hypothesis in sentence and symbol format are shown below. However, researchers often only state the research hypothesis using a format like this: *It is hypothesized that hours of sleep will positively relate to quiz scores.* The format shown in the table below could also be used. Because the hypothesis is directional a one-tailed is needed.

Directional Hypothesis for a Bivariate Correlation

Research hypothesis	Hours of sleep will be positively related to quiz scores.	$H_A : r_{xy} > 0$
Null hypothesis	Hours of sleep will not be positively related to quiz scores.	$H_0 : r_{xy} \leq 0$

2. **Choose the inferential test (formula) that best fits the hypothesis.**

The relationship between two quantitative variables is being tested so the appropriate test is a bivariate correlation.

3. **Determine the critical value.**

In order to determine the critical value for a bivariate correlation, three things must be identified:

1. the alpha level,
2. the degrees of freedom (df), and
3. whether the hypothesis is directional (requiring a one-tailed test) or non-directional (requiring a two tailed test).

The alpha level is often set at .05 unless there is reason to adjust it such as when multiple hypotheses are being tested in one study or when a Type I Error could be particularly problematic. The default alpha level can be used for this example because only one hypothesis is being tested and there is no clear indication that a Type I Error would be especially problematic. Thus, alpha can be set to 5%, which can be summarized as α = .05.

The df must also be calculated. The df is calculated as the sample size minus the number of variables being tested. In bivariate correlation, there are always 2 variables being tested so the formula is $df = n - 2$. The sample size in Data Set 12.1 is 10. Thus, the df for Data Set 12.1 is as follows:

$$df = n - 2$$
$$df = 10 - 2$$
$$df = 8$$

The hypothesis is directional because it specified that the expected correlation would be positive. Thus, this hypothesis requires a one-tailed test of significance.

The alpha level, df, and determination of whether the hypothesis requires a one-tailed or two-tailed test of significance are used to locate the critical value from the test. The full tables of the critical values for *r* are located in Appendix G. Below is an excerpt of

the section of the *r*-tables that fits the current hypothesis and data. Under the conditions of an alpha level of .05, df = 8, and using a one-tailed test, the critical value is.549.

Critical Values Table

Degrees of Freedom	**one-tailed test**		
	alpha level:	**α = 0.05**	**α = 0.01**
	8	.549	.716

The critical value represents the value which must be exceeded in order to declare a result significant. The obtained value (which is called *r* in correlation) is the magnitude of evidence present. Because the correlation is directional and states the correlation will be positive, the obtained value must both be in the hypothesized direction (i.e. indicate a positive relationship) and exceed the critical value in magnitude. Thus, in order for the result to significantly support the hypothesis is needs to be positive and exceed the critical value of .549.

4. **Calculate the test statistic.**

A test statistic can also be referred to as an obtained value. The formula needed to find the test statistic *r* for this scenario is as follows:

$$r = \frac{\Sigma(X-\bar{X})(Y-\bar{Y})}{\sqrt{\Sigma(X-\bar{X})^2}\sqrt{\Sigma(Y-\bar{Y})^2}}$$

Section A: Preparation

Start each inferential formula by identifying and solving for the pieces that must go into the formula. For bivariate correlation, this preparatory work is as follows:

1. Find $\bar{X}$ (the mean for the X-variable scores)

This value is found using Data Set 12.1 and is summarized as $\bar{X}$ = 6.50

2. Find $\bar{Y}$ (the mean for the Y-variable scores)

This value is found using Data Set 12.1 and is summarized as $\bar{Y}$ = 80.00

Now that the pieces needed for the formula have been found, we can move to Section B.

Section B: Solving

The values from the preparatory steps can now be plugged into the correlation formula and used to find the *r*-value. Much of the work involves finding deviations, which were reviewed in detail in Chapter 4. Therefore, a summary table will be used to show deviations for the two variables in this section.

1. Find $\sqrt{\Sigma(X-\bar{X})^2}$
   a. Find $((X-\bar{X})\,)$ by subtracting the mean of X from each x-value
   b. Find $\Sigma(X-\bar{X})^2$ by squaring each deviation and then summing those values, yielding SS_X
   c. Square root SS_X.
2. Find $\sqrt{\Sigma(Y-\bar{Y})^2}$
   a. Find $(Y-\bar{Y}\,)$ by subtracting the mean of Y from each y-value
   b. Find $\Sigma(Y-\bar{Y})^2$ by squaring each deviation and then summing those values, yielding SS_Y
   c. Square root SS_Y.
3. Find $\Sigma(X-\bar{X})(Y-\bar{Y})$
   a. Find $(X-\bar{X})(Y-\bar{Y})$ by multiplying each deviation from X by its deviation from Y to get the product of the deviations for each case
   b. Find $\Sigma(X-\bar{X})(Y-\bar{Y})$ by summing the product of deviations for each case (i.e. sum the results from step 3a). This is the numerator for the formula.

4. Find the denominator by multiplying the square root of SS_X (which is the result of step 1c) by the square root of SS_Y (which is the result of step 2c)
5. Divide the numerator (which is the result of step 3b) by the denominator (which is the result of step 4).

Correlation Computations for Data Set 12.1.

Steps 1 Through 3							
Deviation Steps		**1a**	**1b**		**2a**	**2b**	**3a**
	Sleep Hours (X)	Deviation $(X-\bar{X})$	Dev. Squared $(X-\bar{X})$	Quiz Scores (Y)	Deviation $(Y-\bar{Y})$	Dev. Squared $(Y-\bar{Y})$	$(X-\bar{X})(Y-\bar{Y})$
	7	0.50	0.25	92	12	144	6.00
	8	1.50	2.25	88	8	64	12.00
	9	2.50	6.25	96	16	256	40.00
	6	-0.50	0.25	70	-10	100	5.00
	6	-0.50	0.25	79	-1	1	0.50
	4	-2.50	6.25	64	-16	256	40.00
	5	-1.50	2.25	75	-5	25	7.50
	10	3.50	12.25	98	18	324	63.00
	3	-3.50	12.25	53	-27	729	94.50
	7	0.50	0.25	85	5	25	2.50
Summation Steps	$\bar{X}$ = 6.50		**Step 1c** $\Sigma(X-\bar{X})$ = 42.50	$\bar{Y}$ = 80.00		**Step 2c** $\Sigma(Y-\bar{Y})$ = 1,924.00	**Step 3b** $\Sigma(X-\bar{X})(Y-\bar{Y})$ = 271.00
Step 4							
Find the denominator	$\sqrt{\Sigma(X-\bar{X})^2}\sqrt{\Sigma(Y-\bar{Y})^2} = \sqrt{42.50}\sqrt{1924} = 285.9545\ldots$						
Step 5							
Put the pieces together to find r and then round to the hundredths place	$r = \frac{\sum(X-\bar{X})(Y-\bar{Y})}{\sqrt{\sum(X-\bar{X})^2}\sqrt{\sum(Y-\bar{Y})^2}}$ $r = \frac{271.00}{285.9545\ldots}$ $r = .9477\ldots$ $r \approx .95$						

The obtained value for this test is .95 when rounded to the hundredths place. This is a very strong, positive correlation.

5. **Apply a decision rule and determine whether the result is significant.**

Assess whether the obtained value for r exceeds the critical value in the proper direction as follows:

a. Check the direction. The hypothesis stated the relationship would be positive. The result was positive. Thus, the direction hypothesized is supported.
b. Check the magnitude. The critical value is .549. The obtained r-value is .95. The obtained r-value exceeds (i.e. is greater in magnitude than) the critical value. Thus, the magnitude is sufficient to support the hypothesis.
c. Decide whether the evidence is sufficient to support the hypothesis

The criteria has been met for both direction and magnitude. Thus, the result significantly supports the hypothesis.

> **Note**
>
> If the hypothesis had not been directional, the only comparison needed before concluding that the hypothesis was supported would be the magnitude of the obtained r-value compared to the critical value.

6. **Calculate the effect size and/or other relevant secondary analyses.**

When it is determined that the result is significant, effect sizes should typically be computed. However, in correlation a secondary analysis is typically given rather than an effect size. The secondary computation for correlation is known as the coefficient of determination and, thus, this will be the focus of step 6 for correlation.

The **coefficient of determination** is the percent of variation in the Y-variable that is accounted for by variance in the X-variable. It can also be described as a calculation of how well a model using one variable (X) can be used to estimate the other (Y). These refer to two ways of interpreting and describing the same thing with the former more applicable to correlation and the latter more applicable to regression. We will focus on Regression in the next chapter (Chapter 13). As the focus of this chapter is correlation, we will use the interpretation language that is most applicable to correlation.

The symbol and the formula for the coefficient of determination are the same and are written as follows:

$$r^2$$

To calculate this, the obtained r-value is squared and often reported as a percent. The greatest r^2 can be is 1.00 (or 100.00% when presented as a percent). This would occur if there was a perfect correlation and could be interpreted and reported as follows:

Approximately 100.00% of the variance in Y is accounted for by variance in X.

The lowest r^2 can be is 0.00 (or 0.00% when presented as a percent). This would occur if there was no correlation (r = 0.00) and could be interpreted and reported as follows:

Approximately 0.00% of the variance in Y is accounted for by variance in X.

However, these two extreme correlation coefficients are rare and reporting a coefficient of determination is only warranted when there is a significant result for the r-value. Our result with Data Set 12.1 was r = .9477… which was significant. Thus, the coefficient of determination is warranted and would be computed as follows:

$$r^2 = (0.9477\ldots)^2$$
$$r^2 = 0.8981\ldots$$

This result is quite large and can be reported as a percent and interpreted as follows for Data Set 12.1:

Approximately 89.81% of the variance in quiz scores was accounted for by variance in hours of sleep.

7. **Report the results in American Psychological Associate (APA) format.**

Results for inferential tests are often best summarized using a paragraph that states the following:

a. the hypothesis and specific inferential test used,
b. the main results of the test and whether they were significant,
c. any additional results that clarify or add details about the results,
d. whether the results support or refute the hypothesis.

Following this, the results for our hypothesis with Data Set 12.1 can be written as shown in the summary example below.

> **APA Formatted Summary Example**
>
> A bivariate correlation was used to test the hypothesis that hours of sleep would positively relate to quiz scores. Consistent with the hypothesis, sleep was positively related to quiz scores, $r(8) = .95, p < .05$. Approximately 89.81% of the variance in quiz scores was accounted for by variance in hours of sleep.

As always, the APA-formatted summary provides a lot of detail in a particular order. For a brief review of the structure for the APA-formatted summary of the test results, see the summary below.

Anatomy of the Evidence String

The following breaks down what each part represents in the evidence string for the correlation results in the APA-formatted paragraph above:

Symbol for the test	Degrees of Freedom	Obtained Value	r-Value
r	(8)	= .95,	r < .05.

Reading Review 12.3

1. How is df calculated for bivariate correlation?
2. What is $\bar{Y}$ and how is it calculated?
3. What information is needed to find the critical value for a bivariate correlation?
4. What does the coefficient of determination estimate?
