# COEFFICIENT OF CORRELATION

*Hervé Abdi*[*]

The coefficient of correlation, typically denoted by the letter *r*, evaluates the similarity of two sets of measurements (i.e., two dependent variables) obtained from the same sample. To do so, the coefficient of correlation quantifies the amount of information, or shared variance, common to the two variables.

The idea of correlation is rather old but the modern approach and definition of the coefficient of correlation was initiated by Francis Galton (in an evolutionary context) in the end of the 19th century and formalized by Karl Pearson in the early 20th century. The sampling distribution of this coefficient was mostly derived a few years later by Ronald Fisher and this was the source of a lifelong enmity between these two giants of statistics.

The correlation coefficient takes values between –1 and +1 (inclusive). A value of +1 shows that the two series of measurements are measuring the same thing, whereas a value of –1 indicates that the two measurements are still measuring the same thing, but one measurement varies inversely to the other. A value of 0 indicates that the two series of measurements have nothing in common. It is important to note that the coefficient of correlation measures only the *linear* relationship between two variables and that its value is very sensitive to outliers.

The squared correlation gives the proportion of common variance between two variables and is also called the *coefficient of determination.* Subtracting the coefficient of determination from unity gives the proportion of variance not shared between two variables. This quantity is called the *coefficient of alienation.*

The significance of the coefficient of correlation can be tested with an *F* or a *t* test. This entry presents three different approaches that can be used to obtain *p* values: (1) the classical approach, which relies on Fisher's *F* distributions; (2) the Monte Carlo approach, which relies

[*] In Bruce Frey (Ed.), *The SAGE Encyclopedia of Research Design.* Thousand Oaks, CA: Sage. 2022

Address Correspondence to Hervé Abdi,
Program in Cognition and Neurosciences, MS: Gr.4.1,
The University of Texas at Dallas,
Richardson, TX 75080—3021 USA
**E-mail: herve@utdallas.edu https://personal.utdallas.edu/~herve.**

on computer simulations to derive empirical approximations of sampling distributions; and (3) the nonparametric permutation (also known as randomization) test, which evaluates the likelihood of the actual data against the set of all possible configurations of these data. In addition to *p* values, confidence intervals can be computed using Fisher's *Z* transform or the more modern, computationally based, and nonparametric Efron's bootstrap.

The coefficient of correlation always overestimates the intensity of the correlation in the population and needs to be "corrected" in order to provide a better estimation. The corrected value is called *shrunken* or *adjusted.*

This entry also presents variations of the correlation coefficients (often called nonparametric measures of correlation) that can be used with ordinal or nominal data.

## Notations and Definition

Suppose we have $S$ observations, and for each observation $s$, we have two measurements, denoted $W_s$ and $Y_s$, with respective means denoted $M_W$ and $M_Y$. For each observation, we define the cross-product as the product of the deviations of each variable from its mean. The sum of these cross-products, denoted $SCP_{WY}$, is computed as

$$SCP_{WY} = \sum_{s}^{S} (W_s - M_W)(Y_s - M_Y). \qquad (1)$$

The sum of the cross-products reflects the association between the variables. When the deviations have the same sign, they indicate a positive relationship, and when they have different signs, they indicate a negative relationship.

The average value of the $SCP_{WY}$ is called the covariance (just like the variance, the covariance can be computed by dividing by $S$ or by $S - 1$):

$$\text{cov}_{WY} = \frac{SCP}{\text{Number of Observations}} = \frac{SCP}{S}. \qquad (2)$$

The covariance reflects the association between the variables, but it is expressed in the original units of measurement. To eliminate the units, the covariance is normalized by division by the standard deviation of each variable. This defines the coefficient of correlation, denoted $r_{W,Y}$, which is equal to

$$r_{W,Y} = \frac{\text{cov}_{WY}}{\sigma_W \sigma_Y}, \qquad (3)$$

where $\sigma_W$ (respectively, $\sigma_Y$) denotes the standard deviation of $W$ (respectively, $Y$), which is the square root of the variances computed as:

$$\sigma_W^2 = \frac{1}{S}\sum_{s}^{S}(W_s - M_W)^2 \quad \text{and} \quad \sigma_W^2 = \frac{1}{S}\sum_{s}^{S}(W_s - M_W)^2 \ . \qquad (3a)$$

Rewriting Equation 3 gives a more practical formula:

$$r_{W.Y} = \frac{SCP_{WY}}{\sqrt{SS_W SS_Y}}. \qquad (4)$$

where *SCP* is the sum of the cross-product and $SS_W$ and $SS_Y$ are the sum of squares of *W* and *Y,* respectively.

## Correlation Computation: An Example

The computation for the coefficient of correlation is illustrated with the following data, describing the values of *W* and *Y* for *S*=6 subjects:

$$W_1 = 1, W_2 = 3, W_3 = 4, W_4 = 4, W_5 = 5, W_6 = 7$$
$$Y_1 = 16, Y_2 = 10, Y_3 = 12, Y_4 = 4, Y_5 = 8, Y_6 = 10.$$

**Step 1**
Compute the sum of the cross-products. First compute the means of *W* and *Y:*

$$M_W = \frac{1}{S}\sum_{s=1}^{S} W_s = \frac{24}{6} = 4 \text{ and}$$
$$M_Y = \frac{1}{S}\sum_{s=1}^{S} Y_s = \frac{60}{6} = 10.$$

The sum of the cross-products is then equal to

$$\begin{aligned} SCP_{WY} &= \sum_{s}^{S}(Y_s - M_Y)(W_s - M_W) \\ &= (16-10)(1-4) \\ &\quad + (10-10)(3-4) \\ &\quad + (12-10)(4-4) \\ &\quad + (4-10)(4-4) \\ &\quad + (8-10)(5-4) \\ &\quad + (10-10)(7-4) \qquad (5) \\ &= (6\times -3)(0\times -1) \\ &\quad + (2\times 0)(-6\times 0) \\ &\quad + (-2\times 1) + (0\times 3) \\ &= -18 + 0 + 0 + 0 - 2 + 0 \\ &= -20. \end{aligned}$$

**Step 2**

Compute the sums of squares. The sum of squares of $W_s$ is obtained as

$$\begin{aligned} SS_W &= \sum_{s=1}^{S} (W_s - M_W)^2 \\ &= (1-4)^2 + (3-4)^2 + (4-4)^2 \\ &\quad + (4-4)^2 + (5-4)^2 + (7-4)^2 \\ &= (-3)^2 + (-1)^2 + 0^2 + 0^2 \\ &\quad + 1^2 + 3^2 \\ &= 9+1+0+0+1+9 \\ &= -18+0+0+0-2+0 \\ &= 20. \end{aligned} \tag{6}$$

The sum of squares of $Y_s$ is

$$\begin{aligned} SS_Y &= \sum_{s=1}^{S} (Y_s - M_Y)^2 \\ &= (16-10)^2 + (10-10)^2 \\ &\quad + (12-10)^2 + (4-10)^2 + (8-10)^2 \\ &\quad + (10-10)^2 \\ &= 6^2 + 0^2 + 2^2 + (-6)^2 + (-2)^2 + 0^2 \\ &= 36+0+4+36+4+0 \\ &= 80. \end{aligned} \tag{7}$$

**Step 3**

Compute $r_{W.Y}$. The coefficient of correlation between $W$ and $Y$ is equal to

$$\begin{aligned} r_{W.Y} &= \frac{\sum_s (Y_s - M_Y)(W_s - M_W)}{\sqrt{SS_Y \times SS_W}} = \frac{SCP_{WY}}{\sqrt{SS_W SS_Y}} \\ &= \frac{-20}{\sqrt{80 \times 20}} = \frac{-20}{\sqrt{1600}} = -\frac{20}{40} \\ &= -.5. \end{aligned} \tag{8}$$

This value of $r = .5$ can be interpreted as an indication of a negative linear relationship between $W$ and $Y$—a conclusion confirmed by the visual examination of Figure 1.

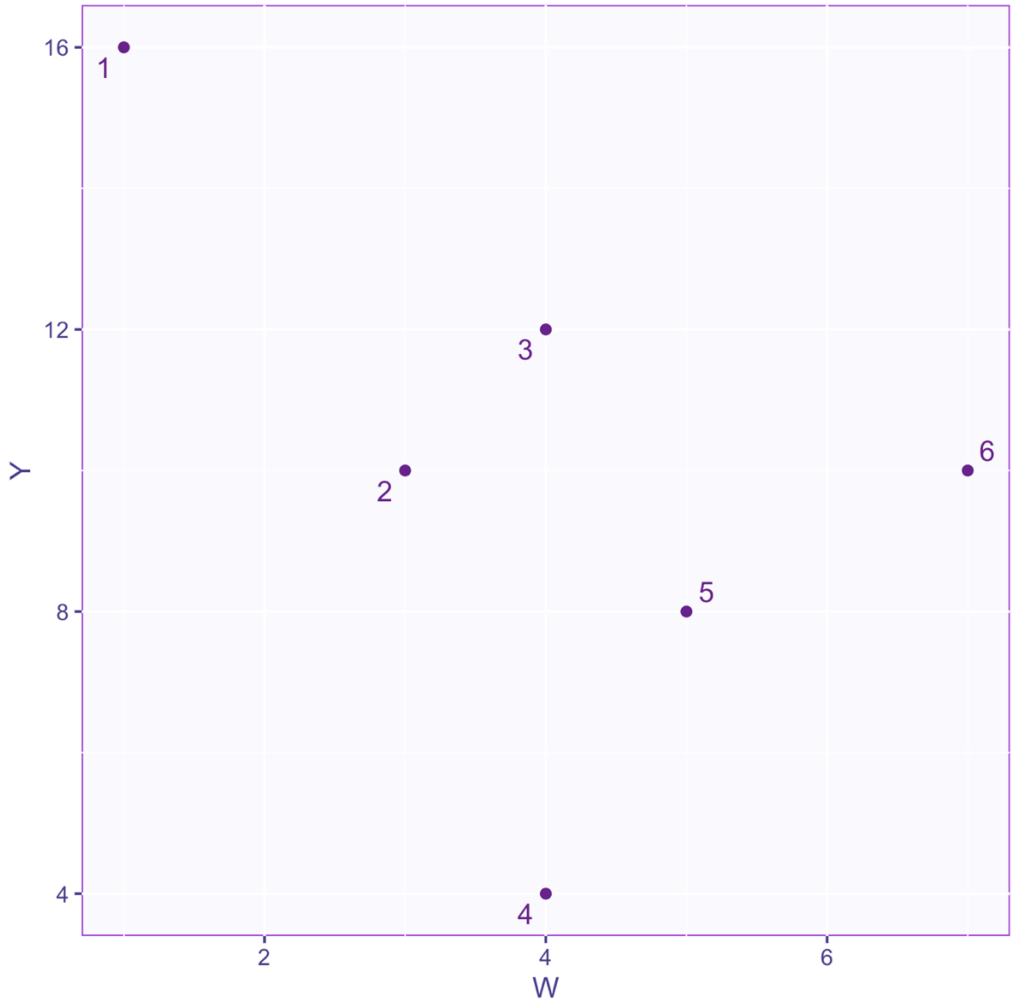


**Figure 1.** The Scatter-plot of variables *W* and *Y*. The plot shows that values of *W* and *Y* are negatively correlated because large values of one variable are paired with small value of the other variable and vice versa.

## Properties of the Coefficient of Correlation

The coefficient of correlation is a number without unit. This occurs because dividing the units of the numerator by the same units in the denominator eliminates the units. Hence, the coefficient of correlation can be used to compare different studies performed using different variables.

The magnitude of the coefficient of correlation is always smaller than or equal to 1. This happens because the numerator of the coefficient of correlation (see Equation 4) is always smaller than or equal to its denominator (this property follows from the Cauchy–Schwartz

inequality). A coefficient of correlation that is equal to +1 or 1 indicates that the plot of the observations will have all observations positioned on a line.

The squared coefficient of correlation gives the *proportion of common variance* between two variables. It is also called the *coefficient of determination.* In our example, the coefficient of determination is equal to $r^2_{WY} = .25$. The proportion of variance not shared between the variables is called the *coefficient of alienation,* and for our example, it is equal to $1 - r^2_{WY} = .75$.

## Interpreting Correlation

The ubiquity of the coefficient of correlation in applied statistics and science is such that it is sometimes incorrectly interpreted. So it is worth stressing that: 1) the coefficient of correlation measures only *linear* relationships, 2) that it is very sensitive to the effect of outliers (this is why the computation of a coefficient of correlation should always go with a graphical display of the relationship between the variables of interest), 3) that the correlation between two variables cannot be directly used to conclude that there is a *causal* relationship between these variables. These themes are developed below.

### Linear and Nonlinear Relationship

The coefficient of correlation measures only linear relationships between two variables and will miss nonlinear relationships. For example, Figure 2 displays a perfect nonlinear relationship between two variables (i.e., the data show a U-shaped relationship with *Y* being proportional to the square of *W*), but the coefficient of correlation is equal to 0.

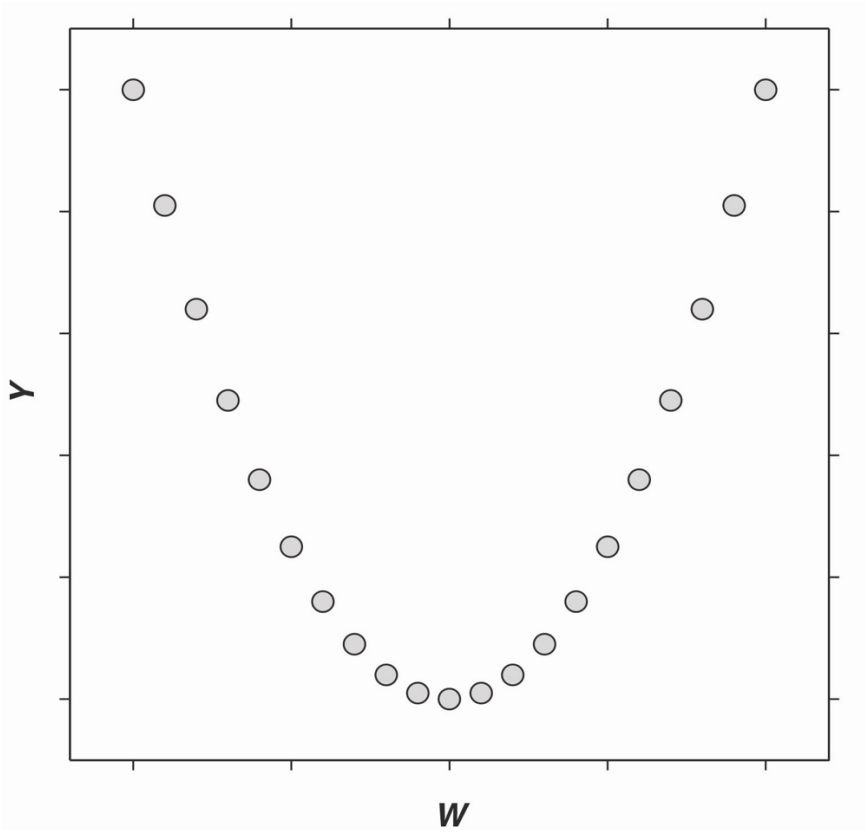


**Figure 2** A Perfect Nonlinear Relationship With a 0 Correlation ($r_{W:Y} = 0$)

**Effect of Outliers**

Observations far from the center of the distribution contribute a lot to the sum of the cross-products. In fact, as illustrated in Figure 3, a single extremely deviant observation (often called an *outlier*) can dramatically influence the value of *r.* Another famous example of the sensitivity of the coefficient of correlation to specific observations was provided, in 1973, by the statistician Francis Anscombe who created four very different bivariate distributions all having a coefficient of correlation equal to .82.

**Geometric Interpretation**

Each set of observations can also be seen as a *vector* in an *S* dimensional space (one dimension per observation). Within this framework, the correlation is equal to the *cosine* of the angle between the two vectors after they have been centered by subtracting their respective mean. For example, a coefficient of correlation of $r = .50$ corresponds to a 150-degree angle. A coefficient of correlation of 0 corresponds to a right angle, and therefore two uncorrelated variables are called *orthogonal* (which is derived from the Greek word for right angle).

**Correlation and Causation**

The fact that two variables are correlated does not mean that one variable causes the other one: *Correlation is not causation.* For example, in France, the number of Catholic churches in a city, as well as the number of schools, is highly correlated with the number of cases of cirrhosis of the liver, the number of teenage pregnancies, and the number of violent deaths. Does this mean that churches and schools are sources of vice and that newborns are murderers? Here, in fact, the observed correlation is due to a third variable, namely the size of the cities: the larger a city, the larger the number of churches, schools, alcoholics, and so forth. In this example, the correlation between number of churches or schools and alcoholics is called a *spurious* correlation because it reflects only their mutual correlation with a third variable (i.e., size of the city).

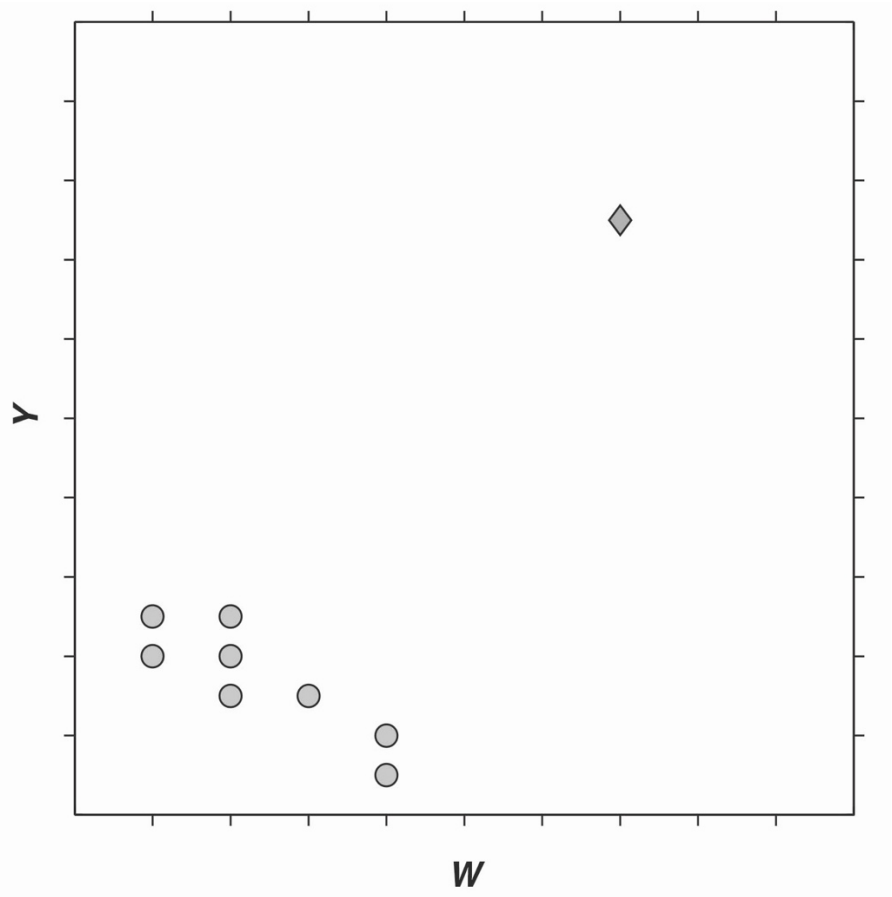


**Figure 3** The Dangerous Effect of Outliers on the Value of the Coefficient of Correlation

*Note:* The correlation of the set of points represented by the circles is equal to –. 87. When the point represented by the diamond is added to the set, the correlation is now equal to

+.61—a change that shows that a single outlier can determine the value of the coefficient of correlation.

## Testing the Significance of *r*

A null hypothesis test for *r* can be performed using an *F* statistic obtained as

$$F = \frac{r^2}{1-r^2} \times (S-2). \qquad (9)$$

For our example, we find that

$$F = \frac{.25}{1-.25} \times (6-2) =$$
$$\frac{.25}{.75} \times 4 = \frac{1}{3} \times 4 = \frac{4}{3} = 1.33.$$

To perform a statistical test, the next step is to evaluate the sampling distribution of the *F*-statistic. This sampling distribution provides the probability of finding any given value of the *F* criterion (i.e., the *p* value) under the null hypothesis (i.e., when there is no correlation between the variables). If this *p* value is smaller than the chosen level (e.g., .05 or .01), then the null hypothesis can be rejected, and *r* is considered significant. The problem of finding the *p* value can be addressed in three ways: (1) the classical approach, which uses Fisher's *F* distributions; (2) the Monte Carlo approach, which generates empirical probability distributions; and (3) the (nonparametric) permutation test, which evaluates the likelihood of the actual configuration of results among all other possible configurations of results.

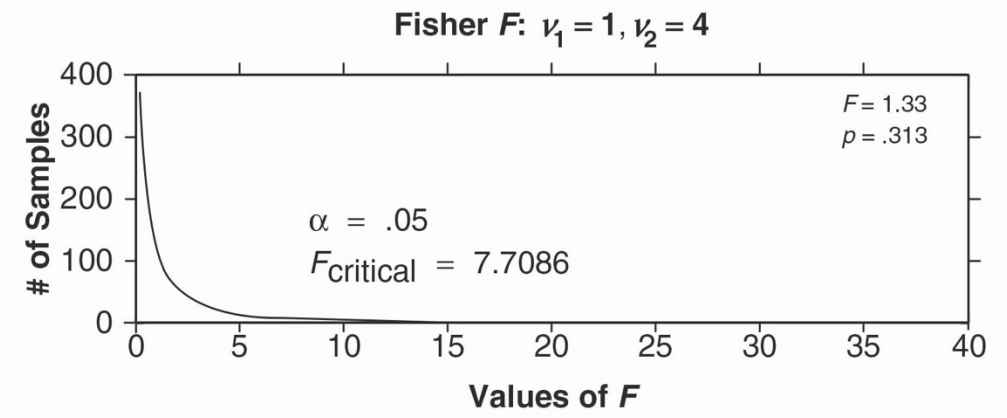


**Figure 4** The Fisher Distribution for $\nu_1 = 1$ and $\nu_2 = 4$, Along With $\alpha = .05$

*Note:* Critical value of $F = 7.7086$.

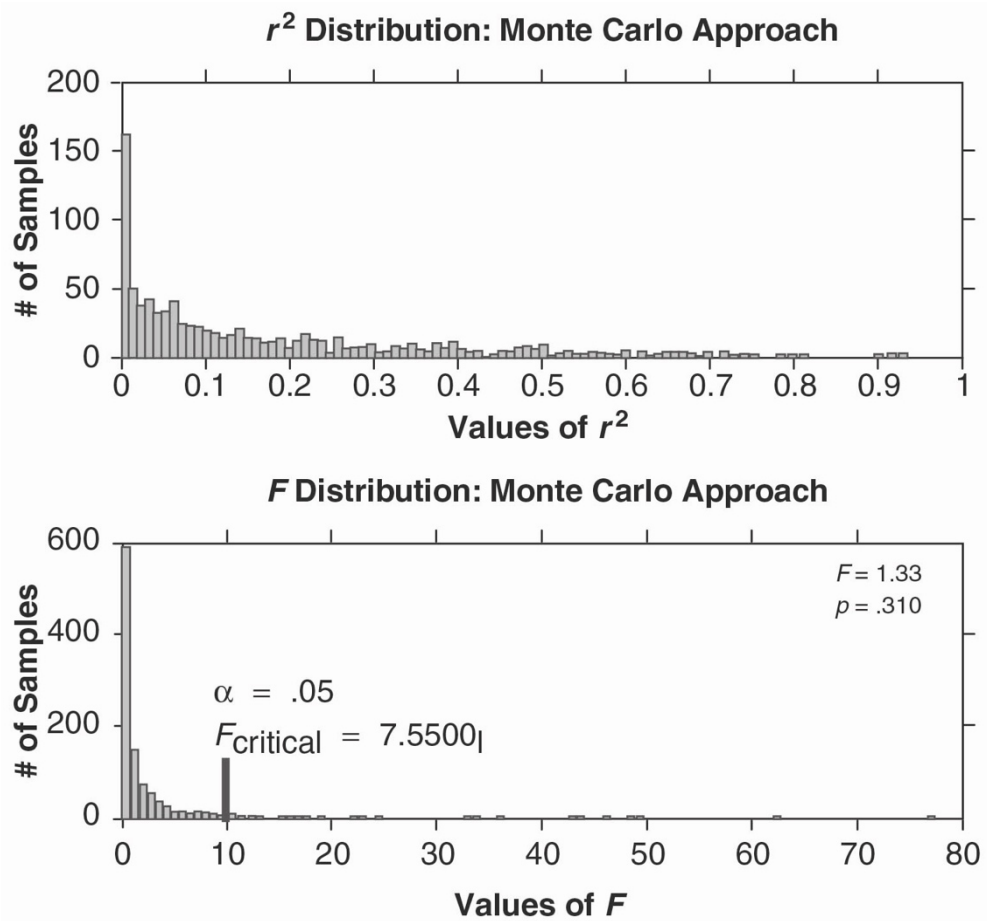


**Figure 5** Histogram of Values of $r^2$ and $F$ Computed From 1,000 Random Samples When the Null Hypothesis Is True

*Note:* The histograms show the *empirical* distribution of $F$ and $r^2$ under the null hypothesis.

### Classical Approach

To analytically derive the sampling distribution of $F$, several assumptions need to be made: (a) the error of measurement is added to the true measure; (b) the error is independent of the measure; and (c) the mean error is normally distributed, has a mean of zero, and has a variance of $\sigma_e^2$. When these assumptions hold and when the null hypothesis is true, the $F$ statistic is distributed as a Fisher's $F$ with $\nu_1 = 1$ and $\nu_2 = S - 2$ degrees of freedom. (Incidentally, an equivalent test can be performed using $t = \sqrt{F}$, which is distributed under $H_0$ as a Student's distribution with $\nu = S - 2$ degrees of freedom).

For our example, the Fisher distribution shown in Figure 4 has $\nu_1 = 1$ and $\nu_2 = S - 2 = 6 - 2 = 4$ and gives the sampling distribution of $F$. The use of this distribution will show that the probability of finding a value of $F = 1.33$ under $H_0$ is equal to $p \approx .313$ (most current statistical packages will routinely provide this value). Such a $p$ value does not lead to rejecting

$H_0$ at the usual levels of $\alpha = .05$ or $\alpha = .01$. An equivalent way of performing a test uses critical values that correspond to values of $F$ whose $p$ value is equal to a given $\alpha$ level. For our example, the critical value (found in tables available in most standard textbooks) for $\alpha = .05$ is equal to $F(1,4) = 7.7086$. Any $F$ with a value larger than the critical value leads to rejection of the null hypothesis at the chosen $\alpha$ level, whereas an $F$ value smaller than the critical value leads one to fail to reject the null hypothesis. For our example, because $F = 1.33$ is smaller than the critical value of 7.7086, we cannot reject the null hypothesis.

**Monte Carlo Approach**

A modern alternative to the analytical derivation of the sampling distribution is to empirically obtain the sampling distribution of $F$ when the null hypothesis is true. This approach is often called a Monte Carlo approach.

With the Monte Carlo approach, we generate a large number of random samples of observations (e.g., 1,000 or 10,000) and compute $r$ and $F$ for each sample. To generate these samples, we need to specify the shape of the population from which these samples are obtained. Let us use a normal distribution (this makes the assumptions for the Monte Carlo approach equivalent to the assumptions of the classical approach). The frequency distribution of these randomly generated samples provides an estimation of the sampling distribution of the statistic of interest (i.e., $r$ or $F$). For our example, Figure 5 shows the histogram of the values of $r^2$ and $F$ obtained for 1,000 random samples of 6 observations each. The horizontal axes represent the different values of $r^2$ (top panel) and $F$ (bottom panel) obtained for the 1,000 trials, and the vertical axis the number of occurrences of each value of $r^2$ and $F$. For example, the top panel shows that 160 samples (of the 1,000 trials) have a value of $r^2 = .01$, which was between 0 and .01 (this corresponds to the first bar of the histogram in Figure 5).

Figure 5 shows that the number of occurrences of a given value of $r^2$ and $F$ decreases as an inverse function of their magnitude: The greater the value, the less likely it is to obtain it when there is no correlation in the population (i.e., when the null hypothesis is true). However, Figure 5 shows also that the probability of obtaining a large value of $r^2$ or $F$ is not null. In other words, even when the null hypothesis is true, very large values of $r^2$ and $F$ can be obtained.

From this point on, this entry focuses on the $F$ distribution, but everything also applies to the $r^2$ distribution. After the sampling distribution has been obtained, the Monte Carlo procedure follows the same steps as the classical approach. Specifically, if the $p$ value for the criterion is smaller than the chosen $\alpha$ level, the null hypothesis can be rejected. Equivalently,

a value of $F$ larger than the $\alpha$-level critical value leads one to reject the null hypothesis for this $\alpha$ level.

For our example, we find that 310 random samples (out of 1,000) had a value of $F$ larger than $F = 1.33$, and this corresponds to a probability of $p = .310$ (compare with a value of $p = .313$ for the classical approach). Because this $p$ value is not smaller than $\alpha = .05$, we cannot reject the null hypothesis. Using the critical-value approach leads to the same decision. The empirical critical value for $\alpha = .05$ is equal to 7.55 (see Figure 5). Because the computed value of $F = 1.33$ is not larger than the critical value of 7.55, we do not reject the null hypothesis.

## Permutation Tests

For both the Monte Carlo and the traditional (i.e., Fisher) approaches, we need to specify the shape of the distribution under the null hypothesis. The Monte Carlo approach can be used with any distribution (but we need to specify which one we want, but more of the time the normal distribution is chosen), and the classical approach assumes a normal distribution. An alternative way to look at a null hypothesis test is to evaluate whether the pattern of results for the experiment is a rare event by comparing it to all the other patterns of results that could have arisen from these data. This is called a *permutation* test or sometimes a *randomization* test.

This nonparametric approach originated with William Gosset (better known under his nom de plume of *Student*) and Fisher, who developed the (now standard) $F$ approach because it was possible then to compute one $F$ but very impractical to compute the $F$s for all possible permutations. If Student and Fisher could have had access to modern computers, it is likely that permutation tests would nowadays be the standard procedure.

So, to perform a permutation test, we need to evaluate the probability of finding the value of the statistic of interest (e.g., $r$ or $F$) that we have obtained, compared with all the values we could have obtained by permuting the values of the sample. For our example, we have six observations, and therefore there are

$$6! = 6 \times 5 \times 4 \times 3 \times 2 = 720$$

different possible patterns of results. Each of these patterns corresponds to a given permutation of the data. For instance, here is a possible permutation of the results for our example:

$$W_1 = 1;\ W_2 = 3;\ W_3 = 4;\ W_4 = 4;\ W_5 = 5;\ W_6 = 7$$
$$Y_1 = 8;\ Y_2 = 10;\ Y_3 = 16;\ Y_4 = 12;\ Y_5 = 10;\ Y_6 = 4$$

(Note that we need to permute just one of the two series of numbers; here we permuted *Y*). This permutation gives a value of $r_{WY} = -.30$ and of $r^2_{WY} = .09$. We computed the value of $r_{W.Y}$ for the remaining 718 permutations. The histogram is plotted in Figure 6, where, for convenience, we have also plotted the histogram of the corresponding *F* values.

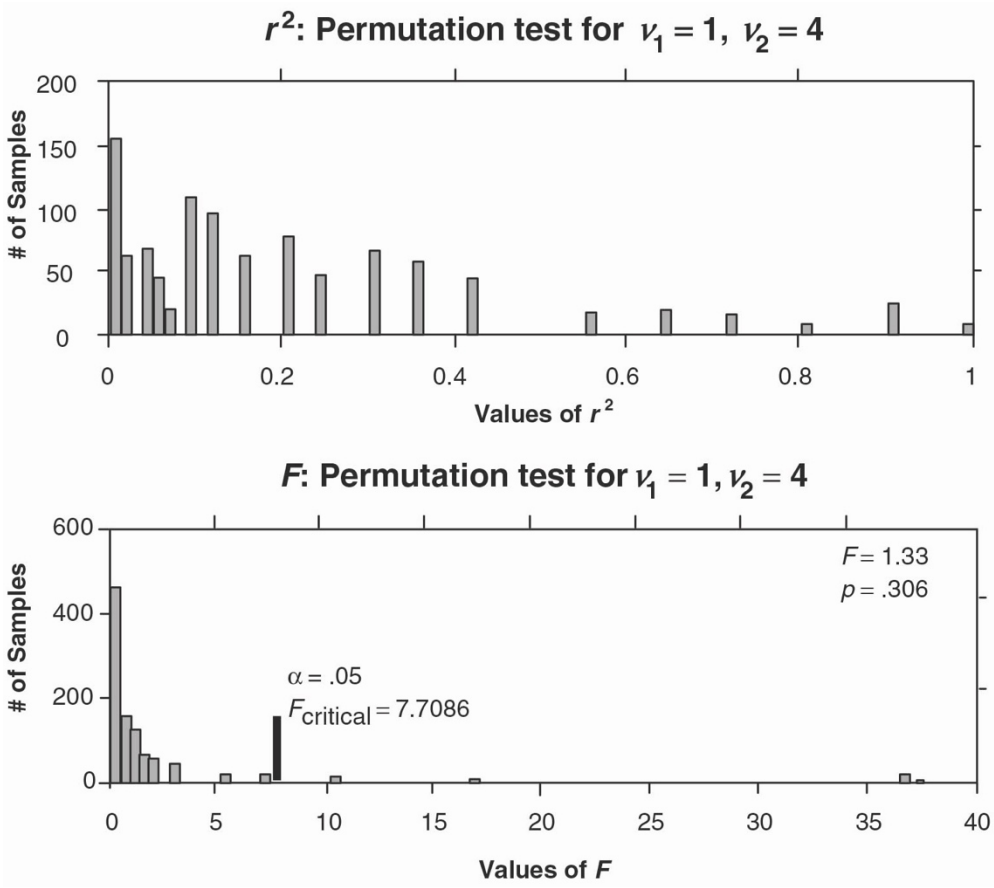


**Figure 6** Histogram of *F* Values Computed From the 6! = 720 Possible Permutations of the Six Scores of the Example

For our example, we want to use the permutation test to compute the probability associated with $r^2_{WY} = .25$. This is obtained by computing the proportion of $r^2_{WY}$ larger than .25. We counted 220 $r^2_{WY}$ out of 720 larger or equal to .25; this gives a probability of

$$p = \frac{220}{720} = .306.$$

It is interesting to note that this value is very close to the values found with the two other approaches (cf. Fisher distribution $p$ = .313 and Monte Carlo $p$ = .310). This similarity is confirmed by comparing Figure 6, where we have plotted the permutation histogram for *F,* with Figure 4, where we have plotted the Fisher distribution.

When the number of observations is small (as is the case for this example with six observations), it is possible to compute all the possible permutations. In this case we have an *exact* permutation test. But the number of permutations grows very fast as the number of observations increases. For example, with 20 observations the total number of permutations is close to $2.4 \times 10^{18}$ (this is a very big number). Such large numbers obviously prohibit computing all the permutations. Therefore, for samples of large size, we approximate the permutation test by using a large number (say 10,000 or 100,000) of random permutations (this approach is sometimes called a *Monte Carlo permutation test*).

## Confidence Intervals

A null hypothesis test for a correlation coefficient computed on a (random) sample of observations obtained from a given population only evaluates if the population value of the correlation is zero but this null hypothesis test informs neither about the *size* of the correlation in the population nor about the *expected size* of the correlation if the original study were to be replicated. These concerns are addressed by computing confidence intervals as they provide a *range* of likely values of the correlation in the population or for replications of the original studies. These intervals can be theoretically computed from analytical approaches (which, in general, require statistical assumptions such as normality of the scores in the population) or, more recently, by computational approaches such as the bootstrap (that require modern computation resources).

### Classical Approach

The value of $r$ computed from a sample is an estimation of the correlation of the population from which the sample was obtained. Suppose that we obtain a new sample from the same population and that we compute the value of the coefficient of correlation for this new sample. In what range is this value likely to fall? This question is answered by computing the confidence interval of the coefficient of correlation. This gives an upper bound and a lower bound between which the population coefficient of correlation is likely to stand. For example, we want to specify the range of values of $r_{W.Y}$ in which the correlation in the population has a 95% chance of falling.

Using confidence intervals is more general than a null hypothesis test because if the confidence interval excludes the value 0 then we can reject the null hypothesis. But a confidence interval also gives a range of probable values for the correlation. Using confidence intervals has another big advantage: We can act as if we could accept the null hypothesis. To do so, we first compute the confidence interval of the coefficient of correlation and look at the

largest magnitude it can have. If we consider that this value is small, then we can say that even if the magnitude of the population correlation is not zero, it is too small to be of interest.

Conversely, we can give more weight to a conclusion if we show that the smallest possible value for the coefficient of correlation will still be large enough to be impressive.

The problem of computing the confidence interval for *r* has been explored (once again) by Student and Fisher. Fisher found that the problem was not simple but that it could be simplified by transforming *r* into another variable called *Z.* This transformation, which is called Fisher's *Z* transform, creates a new *Z* variable whose sampling distribution is close to the normal distribution. Therefore, we can use the normal distribution to compute the confidence interval of *Z,* and this will give a lower and a higher bound for the population values of *Z.* Then we can transform these bounds back into *r* values (using the inverse *Z* transformation), and this gives a lower and upper bound for the possible values of *r* in the population.

**Fisher's *Z* Transform**

Fisher's *Z* transform is applied to a coefficient of correlation *r* according to the following formula:

$$Z = \frac{1}{2}\left[\ln\left(1+r\right) - \ln\left(1-r\right)\right], \quad (10)$$

where ln is the *natural* logarithm.

The inverse transformation, which gives *r* from *Z,* is obtained with the following formula:

$$r = \frac{\exp\left\{2\times Z\right\}-1}{\exp\left\{2\times Z\right\}+2}, \quad (11)$$

where $\exp\{x\}$ means to raise the number *e* to the power *x* (*i.e.*, $exp\{x\} = e^x$ and *e* is Euler's constant, which is approximately 2.71828). Most hand calculators can be used to compute both transformations.

Fisher showed that the new *Z* variable has a sampling distribution that is normal, with a mean of 0 and a variance of $S - 3$. From this distribution we can compute directly the upper and lower bounds of *Z* and then transform them back into values of *r.*

**Example**

The computation of the confidence interval for the coefficient of correlation is illustrated using the previous example, in which we computed a coefficient of correlation of $r = .5$ on a sample

made of $S = 6$ observations. The procedure can be decomposed into six steps, which are detailed next.

*Step 1*

Before doing any computation, we need to choose an α level that will correspond to the probability of finding the population value of $r$ in the confidence interval. Suppose we chose the value α = .05. This means that we want to obtain a confidence interval such that there is a 95% chance, or (1 – α) = (1 – .05) = .95, of having the population value being in the confidence interval that we will compute.

*Step 2*

Find in the table of the normal distribution the critical values corresponding to the chosen $\alpha$ level. Call this value $Z_\alpha$. The most frequently used values are

$$Z_{\alpha=.10} = 1.645 \ (\alpha = .10)$$
$$Z_{\alpha=.05} = 1.960 \ (\alpha = .05)$$
$$Z_{\alpha=.01} = 2.575 \ (\alpha = .01)$$
$$Z_{\alpha=001} = 3.325 \ (\alpha = 001).$$

*Step 3*

Transform $r$ into $Z$ using Equation 10. For the present example, with $r = .50$, we find that $Z = .5493$.

*Step 4*

Compute a quantity called $Q$ as

$$Q = Z_\alpha \times \sqrt{\frac{1}{S-3}}.$$

For our example we obtain

$$Q = Z_{.05} \times \sqrt{\frac{1}{6-3}} = 1.960 \times \sqrt{\frac{1}{3}} = 1.1316.$$

*Step 5*

Compute the lower and

$$\text{Lower Limit} = Z_{\text{lower}} = Z - Q$$
$$= -0.5493 - 1.1316 = -1.6809$$
$$\text{Upper Limit} = Z_{\text{upper}} = Z + Q$$
$$= -0.5493 + 1.1316 = 0.5823.$$

*Step 6*

Transform $Z_{\text{lower}}$ and $Z_{\text{upper}}$ into $r_{\text{lower}}$ and $r_{\text{upper}}$. This is done with the use of Equation 11. For the present example, we find that

$$\text{Lower Limit} = r_{\text{lower}} = -.9330$$
$$\text{Upper Limit} = r_{\text{upper}} = .5243.$$

**Figure 7** Histogram of $r_{W:Y}$ Values Computed From 1,000 Bootstrapped Samples Drawn With Replacement From the Data From Our Example

The range of possible values of $r$ is very large: the value of the coefficient of correlation that we have computed could come from a population whose correlation could have been as low as $r_{\text{lower}} = -.9330$ or as high as $r_{\text{upper}} = .5243$. Also, because zero is in the range of possible values, we cannot reject the null hypothesis (which is also the conclusion reached with the null hypothesis tests).

It is worth noting that because the $Z$ transformation is nonlinear, the confidence interval is *not* symmetric around $r$.

Finally, current statistical practice recommends the routine use of confidence intervals because this approach is more informative than null hypothesis testing.

## Efron's Bootstrap

A modern Monte Carlo approach for deriving confidence intervals was proposed by Bradley Efron. This approach, called the *bootstrap,* was probably the most important advance for inferential statistics in the second part of the 20th century.

The idea behind the bootstrap is simple but could be implemented only with modern computers, which explains why it is a recent development. With the bootstrap approach, we treat the sample as if it were the population of interest in order to estimate the sampling distribution of a statistic computed on the sample. Practically this means that to estimate the sampling distribution of a statistic, we just need to create bootstrap samples obtained by drawing observations with replacement (whereby each observation is put back into the sample after it has been drawn) from the original sample. The distribution of the bootstrap samples is taken as the population distribution. Confidence intervals are then computed from the percentile of this distribution.

For our example, the first bootstrap sample that we obtained comprised the following observations (note that some observations are missing and some are repeated as a consequence of drawing with replacement):

$$s_1 = \text{observation } 5,$$
$$s_2 = \text{observation } 1,$$
$$s_3 = \text{observation } 3,$$
$$s_4 = \text{observation } 2,$$
$$s_4 = \text{observation } 3,$$
$$s_6 = \text{observation } 6.$$

This gives the following values for the first bootstrapped sample obtained by drawing with replacement from our example:

$$W_1 = 5;\ W_2 = 1;\ W_3 = 4;\ W_4 = 3;\ W_5 = 4;\ W_6 = 7$$
$$Y_1 = 8;\ Y_2 = 16;\ Y_3 = 12;\ Y_4 = 10;\ Y_5 = 12;\ Y_6 = 10.$$

This bootstrapped sample gives a correlation of $r_{W.Y} = -.73$.

If we repeat the bootstrap procedure for 1,000 samples, we obtain the sampling distribution of $r_{W.Y}$ as shown in Figure 7. This figure shows that the values of $r_{W.Y}$ vary a lot with such a small sample (in fact, these values cover the whole range of possible values, from −1 to +1). In order to find the upper and the lower limits of a confidence interval, we look for the corresponding percentiles. For example, if we select a value of $\alpha = .05$, we look at the values of the bootstrapped distribution corresponding to the 2.5th and the 97.5th percentiles. In our example, we find that 2.5% of the values are smaller than –.9487 and that 2.5% of the values are larger than .4093. Therefore, these two values constitute the lower and the upper limits of the 95% confidence interval of the population estimation of $r_{W.Y}$ (cf. the values obtained with Fisher's $Z$ transform of –.9330 and .5243). Contrary to Fisher's $Z$ transform approach, the bootstrap limits are not dependent on assumptions about the population or its parameters (but it is comforting to see that these two approaches concur for our example). Because the value of 0 is in the confidence interval of $r_{W.Y,}$ we cannot reject the null hypothesis. This shows once again that the confidence interval approach provides more information than the null hypothesis approach.

**Shrunken and Adjusted *r***

The coefficient of correlation is a descriptive statistic that *always* overestimates the population correlation. This problem is similar to the problem of the estimation of the variance of a population from a sample. To obtain a better estimate of the population, the value $r$ needs to be

corrected. The corrected value of $r$ (denoted here by $\tilde{r}^2$.) goes under different names: corrected *r,* shrunken *r,* or adjusted *r* (there are some subtle differences between these different appellations, but we will ignore them here). Several correction formulas are available; the one most often used estimates the value of the population correlation as

$$\tilde{r}^2 = 1 - \left[ (1 - r^2) \left( \frac{S-1}{S-2} \right) \right]. \qquad (12)$$

For our example, this gives

$$\tilde{r}^2 = 1 - \left[ (1 - r^2) \left( \frac{S-1}{S-2} \right) \right] = 1 - \left[ (1 - .25) \times \frac{5}{4} \right]$$
$$= 1 - \left[ .75 \times \frac{5}{4} \right] = 0.06.$$

With this formula, we find that the estimation of the population correlation drops from a value of $r = -.50$, to a value of $\tilde{r}^2 = -\sqrt{\tilde{r}^2} = -\sqrt{.06} = -.24$.

## Particular Cases of the Coefficient of Correlation

Mostly for historical reasons, some specific cases of the coefficient of correlation have their own names (in part because these special cases lead to simplified computational formulas). Specifically, when both variables are ranks (or transformed into ranks), we obtain the *Spearman rank correlation coefficient* (a related transformation will provide the *Kendall rank correlation coefficient*); when only one of the two variables is dichotomous and the other one we obtain the *point-biserial coefficient;* when both variables are dichotomous (i.e., they take only the values 0 and 1), we obtain the $\varphi^2$ (squared) *coefficient of correlation;* when both variables are nominal we obtain the (squared coefficient) Cramer's $V^2$ coefficient also called $\varphi'$ coefficient. The $\varphi^2$, $\varphi'$, and $V^2$ coefficients are directly linked to the $\chi^2$ statistic testing independence for a contingency table.

### Spearman Rank Correlation Coefficients

The Spearman rank correlation coefficient, often denoted $\rho$, is obtained as the plain correlation coefficient computed on the rank ordered data (older texts often give a convenient computational formula now made obsolete by computers). As an illustration, with our previous example, the values of *W* and *Y* (reproduced here for convenience):

$$\begin{array}{cccccc} W_1 = 1 & W_2 = 3 & W_3 = 4 & W_4 = 4 & W_5 = 5 & W_6 = 7 \\ Y_1 = 16 & Y_2 = 10 & Y_3 = 12 & Y_4 = 4 & Y_5 = 8 & Y_6 = 10 \end{array}$$

are ranked as:

$$\begin{array}{cccccc} w_1 = 1.0 & w_2 = 2.0 & w_3 = 3.5 & w_4 = 3.5 & w_5 = 5.0 & w_6 = 6.0 \\ y_1 = 6.0 & y_2 = 3.5 & y_3 = 5.0 & y_4 = 1.0 & y_5 = 2.0 & y_6 = 3.5 \end{array} .$$

Note that ties are assigned their average rank (so that the sum of the ranks is the same whether there are ties or not). Using Equation 4, the Spearman coefficient of correlation is obtained from the ranked values as:

$$\rho_{WY} = r_{wy} = \frac{SCP_{wy}}{\sqrt{SS_w \times SS_y}} = \frac{\sum (w_i - M_w)(y_i - M_y)}{\sqrt{\sum (w_i - M_w)^2 \sum (y_i - M_y)^2}} = \frac{-8.5}{\sqrt{17 \times 17}} = -.50 .$$

For this example, Pearson *r* and Spearman $\rho$ give the same value; this is not in general the case, but they often give similar values. Spearman is, however, much less sensitive to extreme values.

*Inferences for Spearman Coefficient of Correlation*

For large number of observations (i.e., more than 10), the significativity of $\rho$ can be tested using Equation 9; for number of observations smaller than 10, exact tables of critical values can be found, for example, in Sidney Siegel's 1956 book *Nonparametric Statistics for the Behavioral Sciences*. For our example, these tables would indicate that, in order to reach significance, the Spearman coefficient of correlation would need to have a magnitude larger than .829. With a value of $\rho$ = –.50, our correlation fails to reach significance—a conclusion similar to the one obtained from the raw data.

Confidence intervals can be computed for Spearman using standard inferential procedures (i.e., Fisher *Z*-transform) or, better, combinatoric or bootstrapped based approaches (see Bishara & Hittner, 2017, for a recent review).

*Kendall Rank Correlation Coefficients*

By contrast with the Spearman correlation coefficient—which is just a variation over Pearson's correlation coefficient—the Kendall rank correlation coefficient (denoted $\tau$) is nonmetric, based on combinatoric, and takes only into account the order between rankings. To do so, the ranks are broken into a set of ordered pairs and a distance (called the symmetric difference distance) is computed by counting the number of different pairs between the two rank orders. This difference is then scaled to fit the [–1 1] range of a coefficient of correlation. Because,

Kendall's $\tau$ reflects the differences between rank orders it is sometimes called a coefficient of disagreement (see, e.g., Siegel, 1956).

So, the first step is to express each variable as a set of ordered pairs of the observations. Here, if we keep the example used to illustrate Spearman correlation and name the observations from *a* to *f*, as shown in the following:

$$\begin{array}{lcccccc} \text{Observation} & a & b & c & d & e & f \\ & w_1 = 1.0 & w_2 = 2.0 & w_3 = 3.5 & w_4 = 3.5 & w_5 = 5.0 & w_6 = 6.0 \\ & y_1 = 6.0 & y_2 = 3.5 & y_3 = 5.0 & y_4 = 1.0 & y_5 = 2.0 & y_6 = 3.5 \end{array} ,$$

we find that variable *w* is equivalent to the following set of pairs where the first element of the pair is strictly preferred to the second element (note that in a tie, the first element is *not* strictly preferred to the second one and so should not be listed):

$$\big\{(a,b),(a,c),(a,d),(a,e),(a,f,),(b,c),(b,d),$$
$$(b,e),(b,f),(c,e),(c,f),(d,e),(d,f),(e,f)\big\} .$$

The same procedure gives the following set of pairs for variable *y:*

$$\big\{(b,a),(c,a),(d,a),(e,a),(f,a),(b,c),(d,b),$$
$$(e,b),(d,c),(e,c),(f,c),(d,e),(d,f),(e,f)\big\} .$$

The next step is to identify the pairs that differ between the two variables. Here this set— denoted $\Delta$— of mismatched pairs is equal to:

$$\big\{(a,b),(a,c),(a,d),(a,e),(a,f),(b,d),(b,e),(b,f),(c,e),(c,f)$$
$$(b,a),(c,a),(d,a),(e,a),(f,a),(d,b),(e,b),(f,b),(e,c),(f,c)\big\} .$$

The cardinal (i.e., number of elements here denoted $d_\Delta$) of this set is called the symmetric difference distance between variables *w* and *y*: in this example it is equal to $d_\Delta = 20$. Kendall's $\tau$ is then obtained by rescaling $d_\Delta$ to fit the [–1 +1] interval of a correlation as:

$$\tau = 1 - \frac{2d_\Delta}{S(S-1)} , \qquad (13)$$

(where *S* is the number of observations). Here, we obtain

$$\tau = 1 - \frac{2d_\Delta}{S(S-1)} = 1 - \frac{2 \times 20}{6 \times 5} = 1 - \frac{4}{3} = -.3333 .$$

*Inferences for Kendall coefficient of correlation*

For values of *N* smaller than 10, exact probabilities can be computed for $\tau$ and tables can be found (e.g., see Siegel, 1956, or Abdi, 2007). Such a table would indicate that for $S = 6$, $\tau$ would need to be at least as large as .7333 to be significant at the $\alpha = .05$ level. And so, with

Kendall's $\tau$ (just like with the other coefficients), variables *W* and *Y* cannot be considered as significantly correlated.

For values of *S* larger than 10, $\tau$ is approximately distributed as a normal distribution with parameters

$$\mu_\tau = 0 \quad \text{and} \quad \sigma_\tau = \sqrt{\frac{2(2S+5)}{9S(S-1)}} \ . \qquad (14)$$

And, therefore, a *Z*-statistic can easily be computed to test Kendall's $\tau$ for *S* larger than 10.

**Cramer's $V^2$ and Phi**

Test of independence for contingency tables are typically conducted by computing a $\chi^2$ of independence. Like most tests, $\chi^2$ can be rewritten as the product of two terms: The first term reflecting the intensity of the effect and the second term reflecting the number of observations. In the case of $\chi^2$, this product can be expressed as

$$\chi^2 = \varphi^2 \times S \qquad (15)$$

($\varphi^2$ is also called the *Inertia* of the contingency table, especially in the context of multivariate approaches tailored for the analysis of contingency tables such as correspondence analysis, see, e.g., Abdi & Béra, 2018). Rewriting Equation 15 shows that:

$$\varphi^2 = \frac{\chi^2}{S} \ . \qquad (16)$$

For an $I$ by $J$ contingency table, the coefficient $\varphi^2$ takes value between 0 and $\min(I, J) - 1$, it is therefore a squared coefficient of correlation for a $2 \times 2$ contingency table (note that, in this particular case, $\varphi^2$ would be equal to the squared Pearson correlation computed between two binary variables). In the general case, to obtain a squared correlation coefficient (i.e., taking values between 0 and 1), the coefficient $\varphi^2$ needs to be rescaled; doing so gives a squared correlation coefficient known as $\varphi'^2$ or Cramer $V^2$ computed as:

$$\varphi'^2 = V^2 = \frac{\varphi^2}{\min(I-1, J-1)} = \frac{\chi^2}{S \times \min(I-1, J-1)} \ . \qquad (17)$$

To illustrate the computation of these coefficients, consider the following $3 \times 6$ contingency table collecting the answers of 260 participants who were asked to associate one (and only one) color to the sound of a vowel:

| | Yellow | Green | Orange | Blue | Red | Violet |
|---|---|---|---|---|---|---|
| ee | 46 | 17 | 2 | 11 | 42 | 5 |
| a | 8 | 7 | 5 | 17 | 30 | 6 |
| ou | 1 | 2 | 15 | 14 | 16 | 16 |

. (18)

The $\chi^2$ of independence for this table is equal to $\chi^2 = 87.75$ with $2 \times 5 = 10$ degrees of freedom (with an associated $p$ value smaller than .0001). The index $\varphi^2$ is then equal to 87.75 / 260 = .337. The squared correlation coefficients derived from $\chi^2$ or $\varphi^2$ are equal to

$$\varphi'^2 = V^2 = \frac{\chi^2}{S \times \min(I-1, J-1)} = \frac{87.75}{260 \times 2} = .169 \ . \qquad (17)$$

Inferences for these squared correlation coefficients are equivalent to inferences based on their $\chi^2$ and so, in this example, these coefficients of correlation would be considered as significantly different from 0.

*Hervé Abdi*

***See also*** Chi-Square Test; Contingency Table Analysis; Coefficient of Concordance; Confidence Intervals; Correspondence Analysis.

## FURTHER READINGS


Abdi, H. (2007). Kendall rank correlation. In N.J. Salkind (Ed.): *Encyclopedia of measurement and statistics*. Thousand Oaks (CA): Sage. pp. 508–510.

Abdi H. & Béra, M. (2018). Correspondence analysis. In R. Alhajj and J. Rokne (Eds.), *Encyclopedia of social networks and mining (2nd Edition).* New York: Springer Verlag.

Abdi H., & Beaton, D. (2022). *Principal component and correspondence analyses using R*. New York: Springer.

Abdi, H., Edelman, B., Valentin, D., & Dowling, W. J. (2009). *Experimental design and analysis for psychology.* Oxford, UK: Oxford University Press.

Anscombe, F. J. (1973). Graphs in Statistical Analysis. *American Statistician*. *27*, 17–21. DOI: 10.1080/00031305.1973.10478966

Bishara, A.J., & Hittner, J.B. (2017) Confidence intervals for correlations when data are not normal. *Behavioral Research Methods.* **49,** 294–309. doi: 10.3758/s13428-016-0702-8

Cohen, J., & Cohen, P. (1983) *Applied multiple regression/correlation analysis for the social sciences.* Hillsdale, NJ: Lawrence Erlbaum.

Darlington, R. B. (1990). *Regression and linear models.* New York: McGraw-Hill.

Edwards, A. L. (1985). *An introduction to linear regression and correlation.* New York: Freeman.

Rodgers, J.L., & Nicewander, W.L. (1988). Thirteen ways to look at the correlation coefficient. *The American Statistician, 42,* 5966. https://doi.org/10.1080/00031305.1988.10475524

Siegel, S. (1956). *Nonparametric statistics for the behavioral sciences.* New York: McGraw-Hill.

Pedhazur, E. J. (1997). *Multiple regression in behavioral research.* New York: Harcourt Brace.