**In Neil Salkind (Ed.), *Encyclopedia of Research Design.***
**Thousand Oaks, CA: Sage. 2010**

# Coefficients of

# Correlation, Alienation and Determination

**Hervé Abdi · Lynne J. Williams**

## 1 Overview

The coefficient of correlation evaluates the similarity of two sets of measurements (*i.e.,* two dependent variables) obtained on the same observations. The coefficient of correlation indicates the amount of information common to the two variables. This coefficient takes values between $-1$ and $+1$ (inclusive). A value of $+1$ shows that the two series of measurements are measuring the same thing. A value of $-1$ indicates that the two measurements are 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.

Hervé Abdi
The University of Texas at Dallas

Lynne J. Williams
The University of Toronto Scarborough

Address correspondence to:
Hervé Abdi
Program in Cognition and Neurosciences, MS: Gr.4.1,
The University of Texas at Dallas,
Richardson, TX 75083–0688, USA
***E-mail:*** herve@utdallas.edu http://www.utd.edu/∼herve

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. We present three different approaches which 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 on computer simulations to derive empirical approximations of sampling distributions; and (3) the non-parametric *permutation* (*a.k.a. randomization*) test which evaluates the likelihood of the actual data against the set of all possible configurations of these data. In addition to obtaining $p$ values, confidence intervals can be computed using Fisher's $Z$-transform or the more modern, computationally based and non-parametric, Efron's Bootstrap.

Note that 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."

## 2 Notations and definition

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 to its mean. The sum of these cross-products, denoted $SCP_{WY}$, is computed as:

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

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

The average value of the $SCP_{WY}$ is called the covariance [just as the variance, the covariance can be computed by dividing by $S$ or

$(S-1)$]:

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

The covariance reflects the association between the variables but it is expressed in the original units of measurement. In order 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

$$r_{W.Y} = \frac{\text{cov}_{WY}}{\sigma_W \sigma_Y} \ . \tag{3}$$

Rewriting the previous formula, gives a more practical formula:

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

where $SCP$ is the sum of the cross-product and $SS_W$ (resp. $SS_Y$) is the sum of squares of $W$ (resp. $Y$).

## 3 Correlation computation: an example

We illustrate the computation for the coefficient of correlation with the following data, describing the values of $W$ and $Y$ for $S = 6$ subjects:

$$W_1 = 1 \quad W_2 = 3 \quad W_3 = 4 \quad W_4 = 4 \quad W_5 = 5 \quad W_6 = 7$$

$$Y_1 = 16 \quad Y_2 = 10 \quad Y_3 = 12 \quad Y_4 = 4 \quad Y_5 = 8 \quad Y_6 = 10$$

*Step 1: Computing 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

$$SCP_{YW} = \sum_s (Y_s - M_Y)(W_s - M_W)$$

$$
\begin{aligned}
&= (16-10)(1-4) + (10-10)(3-4) + (12-10)(4-4) \\
&\quad + (4-10)(4-4) + (8-10)(5-4) + (10-10)(7-4) \\
&= (6 \times -3) + (0 \times -1) + (2 \times 0) + (-6 \times 0) \\
&\quad + (-2 \times 1) + (0 \times 3) \\
&= -18 + 0 + 0 + 0 - 2 + 0 \\
&= -20 \ . \qquad (5)
\end{aligned}
$$

*Step 2: Computing the sums of squares*

The sum of squares of $W_s$ is obtained as

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

The sum of squares of $Y_s$ is

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

*Step 3: Computing $r_{W.Y}$*

The coefficient of correlation between $W$ and $Y$ is equal to

$$
r_{W.Y} = \frac{\sum_{s} (Y_s - M_Y)(W_s - M_W)}{\sqrt{SS_Y \times SS_W}} = \frac{SCP_{WY}}{\sqrt{SS_W SS_Y}}
$$

$$= \frac{-20}{\sqrt{80 \times 20}} = \frac{-20}{\sqrt{1600}} = -\frac{20}{40}$$

$$= -.5 \; . \tag{8}$$

We can interpret this value of $r = -.5$ as an indication of a negative linear relationship between $W$ and $Y$.

## 4 Some 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*, for our example, it is equal to $1 - r^2_{WY} = .75$.

# 5 Interpreting correlation

## 5.1 Linear and nonlinear relationship

The coefficient of correlation measures only *linear* relationships between two variables and will miss *non-linear* relationships. For example, Figure 1 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.

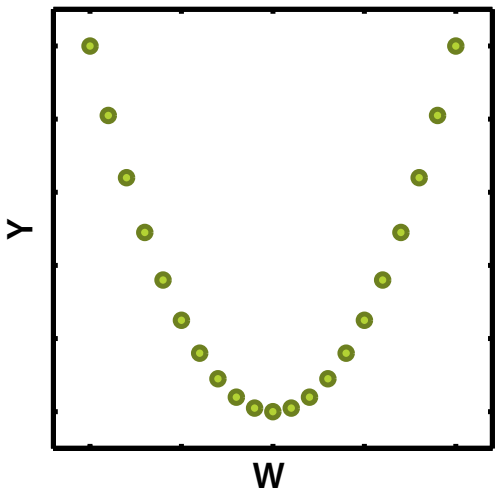


**Figure 1:** A perfect nonlinear relationship with a 0 correlation ($r_{W.Y} = 0$).

## 5.2 The 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 2, a single extremely deviant observation (often called an "outlier") can dramatically influence the value of $r$.

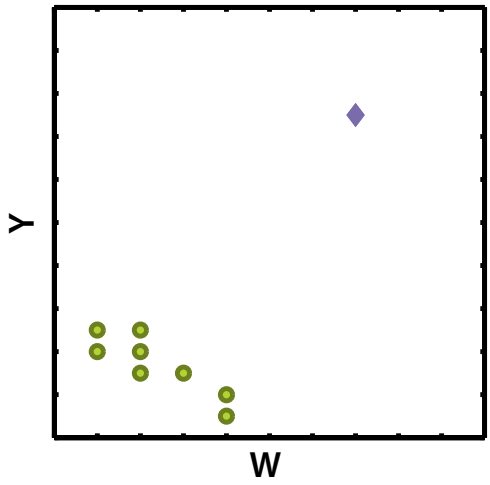


**Figure 2:** The dangerous effect of outliers on the value of the coefficient of correlation. The correlation of the set of points represented by the circles is equal to $-.87$, when the point represented by the diamond is added to the set, the correlation is now equal to $+.61$. This shows that an outlier can determine the value of the coefficient of correlation.

## 5.3 Geometric interpretation: The coefficient of correlation is a cosine

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").

## 5.4 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, are 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 cites: the larger a city, the larger the number of churches, schools and alcoholics, etc. In this example, the correlation between number of churches/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).

## 6 Testing the significance of $r$

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

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

For our example, we find that

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

In order to perform a statistical test, the next step is to evaluate the sampling distribution of the $F$. 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 $\alpha$-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 (non-parametric) permutation test which evaluates the likelihood of the actual configuration of results among all other possible configurations of results.

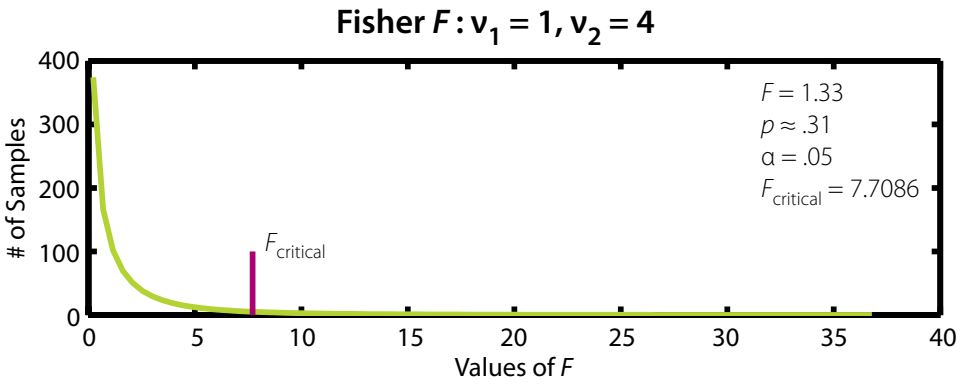


**Figure 3:** The Fisher distribution for $\nu_1 = 1$ and $\nu_2 = 4$, along with the $\alpha = .05$ critical value of $F = 7.7086$.

## 6.1 Finding the probability for $F$: Classical approach

In order to analytically derive the sampling distribution of $F$, several assumptions need to be made: (1) the error of measurement is added to the true measure; (2) the error is independent of the measure; and (3) the mean error is normally distributed, has a mean of zero, and a variance of $\sigma_e^2$. When theses 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 3, has $\nu_1 = 1$ and $\nu_2 = S - 2 = 6 - 2 = 4$ and gives the sampling distribution of $F$. Using 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 statistical packages will routinely provide this value). Such a $p$ value does not lead to rejecting $H_0$ at the usual level 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 the critical value leads to reject the Null Hypothesis at the chosen $\alpha$ level, whereas an $F$ value smaller than the critical value leads 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.

## 6.2 Finding the probability for $F$: 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 (this approach owes its name to the long standing association between probability and gaming: Monte-Carlo is notorious for its casinos).

With the Monte-Carlo approach, we generate a large number of random samples of observations (*e.g.,* $1,000$ or $10,000$) and compute for each sample $r$ and $F$. In order to generate these samples, we need to specify the shape of the population from which these samples are obtained. Here we decided to 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 4 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 (over 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 4).

Figure 4 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 4 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, we can obtain very large values of $r^2$ and $F$.

From now on, we will focus 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.

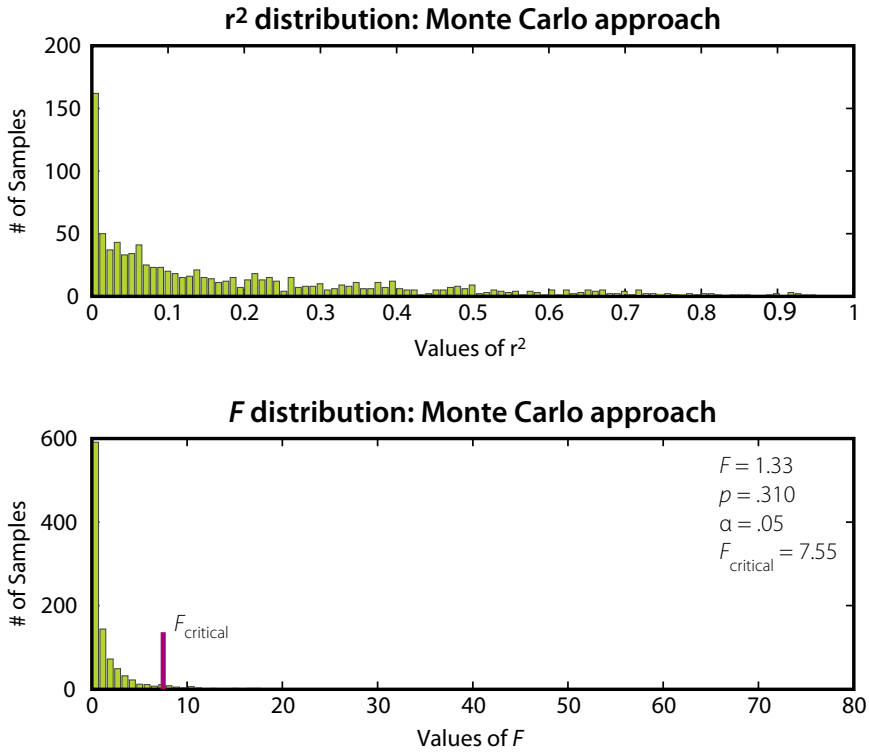


**Figure 4:** Histogram of values of $r^2$ and $F$ computed from 1,000 random samples when the null hypothesis is true. The histogram shows the *empirical* distribution of $F$ and $r^2$ *under* the null hypothesis.

Equivalently, a value of $F$ larger than the $\alpha$ level critical value leads 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 7.5500 (see Figure 4). Because the computed value of $F = 1.33$ is not larger than the 7.5500, we fail to reject the Null Hypothesis.

## 6.3 Finding the probability for $F$: 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) and the classical approach assumes a normal distribution. An alternative way to look at a Null Hypothesis test is to evaluate if 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 also sometimes a *randomization* test.

This *non-parametric* approach originated with Student and Fisher (1935, see also Pitman, 1937, 1938) who developed the (now standard) $F$ approach because it was possible then to compute one $F$ but very impractical to compute all the $F$'s for all possible permutations (it seems that both Student and Fisher spent some inordinate amount of time doing so, though, in order to derive a "feel" for the sampling distribution they were looking for). If Fisher could have had access to modern computers, it is likely that permutation tests would be the standard procedure.

So, in order 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 to all the values we could have obtained by permuting the values of the sample. For our example, we have 6 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:

$$W_1 = 1 \quad W_2 = 3 \quad W_3 = 4 \quad W_4 = 4 \quad W_5 = 5 \quad W_6 = 7$$

$$Y_1 = 8 \quad Y_2 = 10 \quad Y_3 = 16 \quad Y_4 = 12 \quad Y_5 = 10 \quad Y_6 = 4$$

(*Nota Bene,* we just need to permute *one* of the two series of numbers, here we permuted $Y$). This permutation gives a value of $r_{W.Y} = -.30$ and of $r^2_{W.Y} = .09$ . We computed the value of $r_{W.Y}$ for the remaining 718 permutations. The histogram is plotted in Figure 5,

where, for convenience, we have also plotted the histogram of the corresponding $F$ values.

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

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

Interestingly 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 5, where we have plotted the permutation histogram for $F$, with Figure 3, where we have plotted the Fisher distribution.

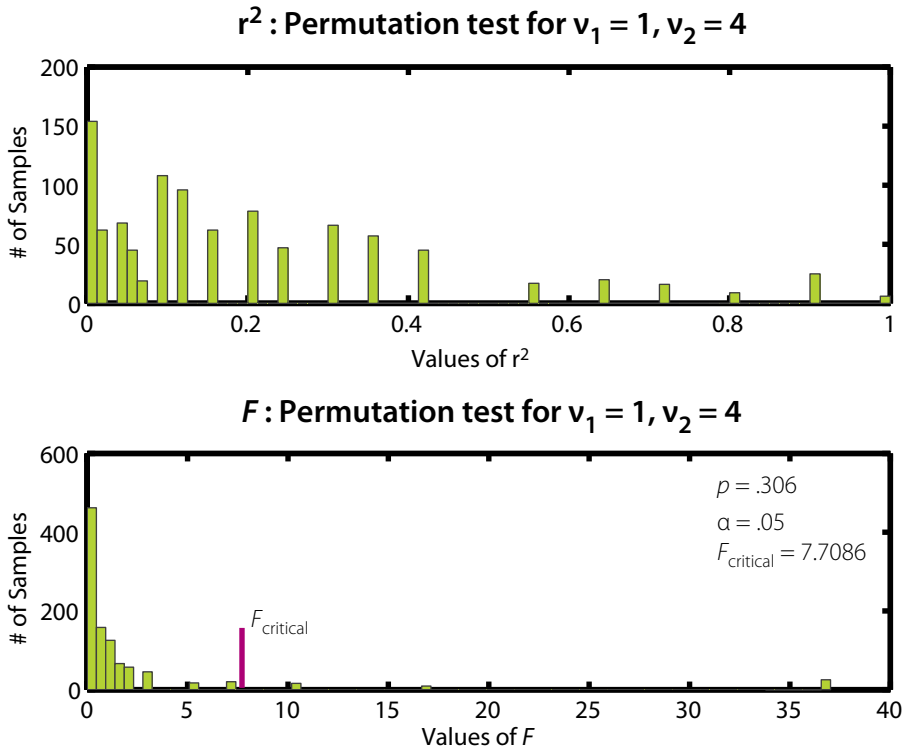


**Figure 5:** Histogram of $F$ values computed from the $6! = 720$ possible permutations of the $6$ scores of the example.

When the number of observations is small (as it is the case for this example with 6 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 when 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*).

# 7 Confidence intervals

## 7.1 Classical approach: Fisher's $Z$ transform

The value of $r$ computed from a sample is an estimation of the correlation of the population from which this 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. In order 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 the Fisher's $Z$-transform, creates a new $Z$-variable whose a 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.

### *7.1.1 Fisher's $Z$ transform*

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

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

where ln is the *natural* logarithm.

The inverse transformation, which gives $r$ from $Z$, is obtained using the following formula:

$$r = \frac{\exp\left\{2 \times Z\right\} - 1}{\exp\left\{2 \times Z\right\} + 1} \ . \tag{11}$$

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

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

### 7.1.2 How to transform $r$ to $Z$: an example

We will illustrate the computation of the confidence interval for the coefficient of correlation using the previous example where we computed a coefficient of correlation of $r = -.5$ on a sample made of $S = 6$ observations. The procedure can be decomposed into seven steps which are detailed below.

*Step 1.* Before doing any computation we need to choose an $\alpha$ level that will correspond to the probability of finding the population value of $r$ in the confidence interval. Suppose we chose the value $\alpha = .05$. This means that we want to obtain a confidence interval such that there is a 95% chance $[(1 - \alpha) = (1 - .05) = .95]$ 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:
- $Z_{\alpha=.10} = 1.645$ (for $\alpha = .10$).
- $Z_{\alpha=.05} = 1.960$ (for $\alpha = .05$).
- $Z_{\alpha=.01} = 2.575$ (for $\alpha = .01$).
- $Z_{\alpha=.001} = 3.325$ (for $\alpha = .001$).

*Step 3.* Transform $r$ into $Z$ using Equation 10. For the present example, with $r = -.5$, we find that $Z = -0.5493$.

*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 upper limits for $Z$ as:

$$\text{Lower Limit } = Z_{\text{lower}} = Z - Q = -0.5493 - 1.1316 = -1.6809$$

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

*Step 6.* Transform $Z_{\text{lower}}$ and $Z_{\text{upper}}$ into $r_{\text{lower}}$ and $r_{\text{upper}}$ This is done using Equation 11. For the present example, we find that

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

$$\text{Upper Limit } = r_{\text{upper}} = .5243$$

As you can see, 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 the conclusion we reached with the Null Hypothesis tests).

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

Finally, current statistical practice recommends to use confidence intervals routinely because this approach is more informative than Null Hypothesis testing.

## 7.2 Modern approach: Efron's Bootstrap

A modern Monte Carlo approach for deriving confidence intervals was proposed by Efron (1979, 1981, see also Efron & Tibshirani, 1993). This approach, called the *bootstrap*, is probably the most important advance for inferential statistics in the second part of the $20th$ Century.

The idea 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 in order to estimate the sampling distribution of a statistic, we just need to create bootstrap samples obtained by drawing observations *with replacement*[1] 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

[1] When we draw an observation with replacement, each observation is put back into the sample after it has been drawn, therefore an observation can be drawn several times.

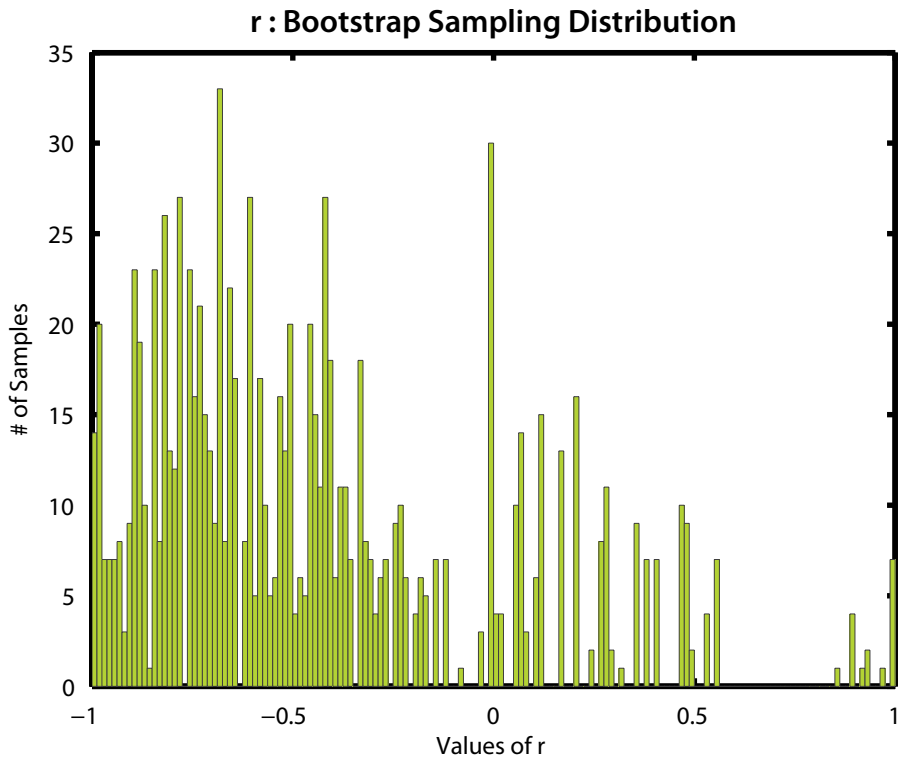


**Figure 6:** Histogram of $r_{W.Y}$ values computed from $1,000$ bootstraped samples drawn with replacement from the data from our example.

are missing and some are repeated as a consequence of drawing with replacement):

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

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

$$W_1 = 5 \quad W_2 = 1 \quad W_3 = 4 \quad W_4 = 3 \quad W_5 = 4 \quad W_6 = 7$$

$$Y_1 = 8 \quad Y_2 = 16 \quad Y_3 = 12 \quad Y_4 = 10 \quad Y_5 = 12 \quad Y_6 = 10 \ .$$

This bootstraped 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 6. From this figure, it is obvious that the value of $r_{W.Y}$ varies a lot with such a small sample (in fact, it covers 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 bootstraped distribution corresponding to the 2.5th and the 97.5th percentile. 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 upon 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.

## 8 Estimating the population correlation: shrunken and adjusted $r$

The coefficient of correlation is a *descriptive* statistic which always overestimates the population correlation. This problem is similar to the problem of the estimation of the variance of a population from a sample. In order to obtain a better estimate of the population, the value $r$ needs to be corrected. The corrected value of $r$ 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) and we denote it by $\tilde{r}^2$. There are several correction formulas available, the one most often used estimates the

value of the population correlation as

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

For our example, this gives:

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

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

## 9 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, see Abdi, 2007); when both variables are dichotomous (i.e., they take only the values 0 and 1), we obtain the Phi coefficient of correlation; and when only one of the two variables is dichotomous, we obtain the point biserial coefficient.

## Further readings

1. Abdi, H. (2007). Kendall rank correlation. In N.J. Salkind (Ed.): *Encyclopedia of Measurement and Statistics*. Thousand Oaks (CA): Sage. pp. 508-510.
2. Abdi, H., Edelman, B., Valentin, D., & Dowling, W.J. (2009). *Experimental Design and Analysis for Psychology*. Oxford: Oxford University Press.
3. Cohen, J., & Cohen, P. (1983) *Applied multiple regression / correlation analysis for the social sciences*. Hillsdale (NJ): Erlbaum.
4. Darlington, R.B. (1990). *Regression and linear models*. New York: McGraw-Hill.
5. Edwards, A.L. (1985). *An introduction to linear regression and correlation*. New York: Freeman.
6. Pedhazur, E.J. (1997) *Multiple regression in behavioral research*. New York: Harcourt-Brace.