Correlations

The Pearson Product-Moment Correlation Coefficient (hereafter: *correlation coefficient*) is a very popular and useful statistic. It provides a measure of both the *direction* and the *strength* of the ***linear*** relationship between two variables. The direction is given by the sign; the strength is given by the square of the correlation coefficient.

The correlation coefficient (r) is both a descriptive and an inferential statistic. It describes the relationship between two paired variables (X and Y) in a given sample (r_{XY}); it also provides a very good estimate for this relationship in the population from which the paired sample was taken (ρ_{XY}). In other words, like the mean, the same formula is used for the descriptive stat and the best guess for the population. Note, however, that r is only unbiased when ρ is zero, while r^2 over-estimates ρ^2 when r is near zero, especially when the sample-size is small. (More on these issues later, when we get to measures of association. There is a simple "fix" for at least one of these problems.)

Calculating the Product-Moment Correlation Coefficient

There are a variety of formulae for calculating the correlation coefficient. They all produce the same value, but are more or less useful in different situations. In general, the correlation coefficient equals the covariance between the two variables, divided by the product of their standard deviations:

$$r_{XY} = COV_{XY} / (S_X \, S_Y)$$

You may use either the observed covariance and observed standard deviations (i.e., those calculated using N), or the estimated population covariance and estimated standard deviations (i.e., those calculated using $N{-}1$), but the three must match. The reason that you can use either version of these values is because the Ns or $N{-}1$s will "cancel" as they appear the same number of times in the numerator as in the denominator. Thus, the value of r does not depend on N.

Another nice property of the correlation coefficient is that it has no units. The units of COV_{XY} are the product of the units of X and the units of Y. The units of S_X are the units of X and the units of S_Y are units of Y. Therefore, the units in the numerator of the formula for the correlation coefficient are the product of the X and Y units, and the units of the denominator are also the product of the X and Y units, so r_{XY} has no units.

Finally, because the most-extreme possible value for COV_{XY} is the product of S_X and S_Y, the most-extreme value for r_{XY} is ±1.00. So the correlation coefficient is a unitless measure with fixed extremes. This is exactly the sort of thing that is particularly useful when one intends to make comparisons across samples that involve completely different dependent measures.

Other Formulae for Sub-types of the Correlation Coefficient

All of the formulae for the correlation coefficient produce the same answer, but some are more easy to use when doing the calculations by hand. For example, if your data are already in the form of *z*-scores (which, by definition, have a mean of zero and a standard deviation of one), then the following can make things much easier:

$$r_{XY} = (\Sigma\, z_X\, z_Y) / N$$

In words: the correlation coefficient is (also) the mean product of *z*-scores. (Yes, this formula has an N in it, but it's effectively cancelled by the Σ, so, as always, the size of r doesn't depend on N.)

Other short-cut formulae are available when the two variables are not both continuous, quantitative variables. For example, when one variable is continuous and quantitative and the other is dichotomous (e.g., when correlating sex with anxiety), the **point-biserial** formula may be used:

$$r_{XY} = [(\bar{Y}_1 - \bar{Y}_0) \surd (p_1\, p_0)] / S_Y$$

where $\bar{Y}_1$ and $\bar{Y}_0$ are the mean values of the quantitative variable for the two levels of the dichotomous variable (which is here coded as having values 1 and 0), p_1 and p_0 are the proportions in the two levels (so they must sum to 1.00), and S_Y is the standard deviation of the quantitative variable ignoring the dichotomous variable. (Note that it's traditional to use 1 and 0 [as opposed to, say, 1 and 2], because dichotomous variables are often yes vs no or true vs false or succeed vs fail and, more importantly, **must** be coded this way for fancier forms of regression.)

Another short-cut is useful when both variables are dichotomous. The **Phi Coefficient** (r_ϕ) is:

$$r_\phi = (BC\text{-}AD) / \surd\, [(A+B)(C+D)(A+C)(B+D)]$$

where A, B, C, and D are the frequencies in the four cells:
(with cells A & B vs C & D [rows] being the levels of one variable,
and cells A & C vs B & D [columns] being the levels of the other variable)

A	B
C	D

☞ Note: all of the different formulae for r_{XY} produce the same answer. They differ only in some specific steps of calculation. Put another way: the alternative formulae (other than the original, $COV_{XY} / S_X\, S_Y$) have merely been simplified, not radically altered.

☞ All correlation coefficients are "reversible" in that $r_{XY} = r_{YX}$. However, regression coefficients are not reversible, so we should all get in the habit of including the subscripts starting now.

Testing the Significance of a Correlation Coefficient

Before going into the details of significance testing, recall the following: in order to conduct *any* kind of inferential test, it is necessary to set up some sort of hypothetical sampling distribution for the value of interest and this often involves making assumptions (e.g., about shape). In the case of the univariate *t*-test for the mean, for example, the hypothetical sampling distribution was created by making an assumption about its shape (i.e., that it is always normal) and then using the estimated population standard deviation (extracted from the sample) to set the spread. Estimating the population standard deviation has $N-1$ degrees of freedom because the standard deviation requires an estimate of the mean, so the univariate *t*-test has N–1 df. In contrast, in the case of the correlation coefficient, we have to estimate two population standard deviations (one for X and one for Y), plus the covariance. To do this, we need estimates of both means, so the degrees of freedom for a correlation is N–2.

In general, there are two ways to use an observed correlation coefficient to test the null hypothesis that $\rho=0$. One way is to simply look up the critical value of r in a table or get it from the web; many text-books include such a table. The other way is to convert the value of r to a value of *t*, which, again, has N–2 df:

$$t = (r \sqrt{\text{df}}) / \sqrt{(1 - r^2)}$$

The fact that a correlation coefficient can be directly transformed into a *t*-statistic has some deep and interesting implications.

⊕ To test the significance of a correlation under SPSS, use **Statistics...Correlate...Bivariate...** and then push the two variables over to the dependent variable window and click OK. Note that SPSS reports N, not N–2, so you have to subtract two to find the degrees of freedom. Note, also, that because SPSS has no idea what α you use, it reports a *p*-value to three decimal places. Finally, when you use **Statistics...Correlate...Bivariate...** you won't be given the value of *t*. If the journal wants this, too, for some reason, you need to use the formula above (or a different SPSS procedure for getting the correlation).

At this point, stop and use SPSS to get the correlation between the two variables in Demo13A.sav. The output will say that r = .965, p = .000, and N = 32. This should be re-expressed as r = .97, *t*(30) = 20.15, p < .001.

Fisher's z′ Transformation (Revisted) and Some Other Tests of Correlations

Because the value of a correlation coefficient is "trapped" between ±1.00, it clearly isn't normal. Therefore, it would be a clear violation of the assumptions of most inferential tests to use raw correlation coefficients as the dependent measure in a *t*-test or an ANOVA. Fortunately, as already mentioned in lecture, there is an already-worked-out transformation for converting correlation coefficients into a new set of values that are close enough to normal for our purposes. This is Fisher's z′ ("zee prime") Transformation:

$$z' = \tfrac{1}{2} [\mathbf{ln} (1+r) - \mathbf{ln} (1-r)] = \tfrac{1}{2} \mathbf{ln} [(1+r) / (1-r)]$$

where **ln** stands for the natural logarithm. Thus, if you have calculated separate correlation coefficients for each of N subjects, and now want to know if the mean value of individual *r*s (in the population) is equal to *V*, then you can convert the *r*s to *z*′s and use the univariate *t*-test.

☞ Note the wording of the above: the univariate *t*-test on a set of z′s concerns the **mean** correlation coefficient in the population, where each member of the population has their own *r* and these can vary (a lot); it does not concern the correlation coefficient *across* the members of the population. As an example of the first situation, consider the correlation between mood and weather (where each person in the population can have a different value, such that some people are greatly affected by weather, and others are not); the z′ test would be used to test the idea that people, *on average*, are affected by weather. As an example of the second, consider the correlation between anxiety and depression (which are often assumed to be relatively stable within people, so each is only measured once in a given person); the standard *r* test would be used to check for a linear relationship between the two variables across people. This second case wouldn't necessarily involve the use of Fisher's z′; you can just use the one value of *r* that you have.

However, you do need the z′ transform when you want to test a null hypothesis other than $\rho=0$. In this case, you use a *z*-test (not a typo: no "prime" mark), because the hypothetical sampling distribution of z′ is known to be the same as the standard normal (and doesn't require any estimating). The value of *z* is given by:

$$z = (z'_O – z'_H) / \sigma_{z'}$$

where z'_O is the transform of the observed value of *r*, z'_H is the transform of the value of *r* under the null hypothesis, and $\sigma_{z'}$ is the standard error of z′, which is given by:

$$\sigma_{z'} = 1 / \surd (N–3)$$

for reasons that I can only half explain. (Note the use of the lower-case sigma, instead of an *s*, to represent the standard error; this is a way of saying that there is no error in estimating this value, which is why we are using a *z*-test and not a *t*-test.) If the value of *z* is outside the range of ±1.96, it is significant at the $\alpha=.05$ level.

Similarly, to test whether two independent correlation coefficients are different from each other (e.g., to test the null hypothesis that $\rho_{AB} = \rho_{CD}$, where A, B, C, and D are four *different* variables), we can use:

$$z = (z'_{AB} – z'_{CD}) / \sigma_{z'-z'}$$

where z'_{AB} and z'_{CD} are the transforms of the two *r*s and $\sigma_{z'-z'}$ is the standard error of their difference, which is given by:

$$\sigma_{z'-z'} = \surd \{ [1/(N_{AB}–3)] + [1/(N_{CD}–3)] \}$$

Again, because this is a z-test, anything outside the range of ±1.96 implies a significant difference.

⊕ SPSS will not perform any of these latter tests for you. (SPSS doesn't even have a built-in gizmo for Fisher's z′.) Therefore, here are some practice examples:

A sample with $N = 25$ had $r = .65$. What is *t*?

A sample with $\boldsymbol{N} = 50$ had $r = .85$. Can the null hypothesis that $\rho = .70$ be rejected?

A sample with $N = 60$ produced $r_{AB} = .40$. A completely separate sample with $N = 35$ produced $r_{CD} = .60$. Can we reject the null hypothesis that $\rho_{AB} = \rho_{CD}$?

A sample with $N = 25$ had $r = .65$. What is t? $t = (.65 \sqrt{} 23) / \sqrt{} (1 – .65^2)$ therefore, $t(23) = 4.10$

A sample with $\boldsymbol{N} = 50$ had $r = .85$. Can the null hypothesis that $\rho = .70$ be rejected?

$z'_O = 1.2562$, $z'_H = 0.8673$, $\sigma_{z'} = 0.1459$ $z = 2.67$ therefore, yes, reject H_0

A sample with $N = 60$ produced $r_{AB} = .40$. A completely separate sample with $N = 35$ produced $r_{CD} = .60$. Can we reject the null hypothesis that $\rho_{AB} = \rho_{CD}$?

$z'_{AB} = 0.4236$, $z'_{CD} = 0.6931$, $\sigma_{z'-z'} = 0.2209$ $z = -1.22$ therefore, no, retain H_0