James W. Grice, Ph.D.
Oklahoma State University
November, 2013

Pearson's correlation coefficient

Use when...

Use this inferential statistical test when you wish to examine the linear relationship between two interval or ratio variables. The population correlation coefficient is represented by the Greek letter rho, ρ. Be careful not to confuse rho with the *p*-value. Pearson's *r* ranges from -1 to +1. Values of -1 or +1 indicate perfect negative or positive, respectively, linear relationships. A value of 0 indicates no linear relationship (although the relationship may be non-linear). The correlation coefficient is an inherently standardized statistic and is therefore readily interpretable.

Assumptions

- Random sampling
- Pairs of observations are independent
- Homoscedasticity (even variation in a scatterplot)
- Bivariate normality (examine normality of each variable)
- Variables are continuous
- Relationship between variables is linear
- H_o is true

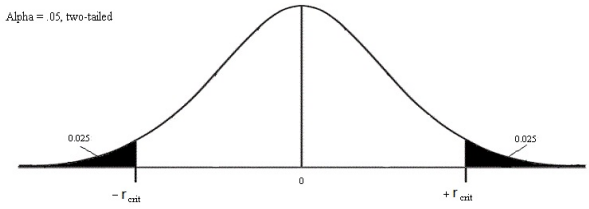
Hypotheses

H_o: ρ = 0

H_A: ρ > 0

or ρ < 0

Sampling Distribution and Critical Values



The *r* distribution is the sampling distribution from which r_{crit} is determined. The darkened area in the distribution to the right is the "rejection region." When r_{obs} falls in the rejection region, the result is "statistically significant", which means that the null hypothesis is rejected. The r_{crit} value is taken from a table of such values or determined using an online calculator. The shape of the *r* distribution changes depending upon the number of people (observations) in the sampling process. Generally speaking, it is a mesokurtic distribution when the null is assumed to be true. When the null hypothesis is false, the sampling distribution becomes skewed. To obtain the correct r_{crit} value, the degrees of freedom value is used. For Pearson's *r*, df = n - 2.

Formulas

There are several formulas that can be used to computer Pearson's *r*. One formula is based on the covariance, which is also an index for the linear association between two continuous variables. Unlike the correlation, however, the covariance is not standardized and can range in value from negative infinity to positive infinity. When the covariance equals zero, then no linear association exists. The formula for the covariance follows:

Observed covariance: $Cov_{xy} = \frac{\sum (x - \bar{x})(y - \bar{y})}{n - 1}$

(x and y are symbols used to denote the two variables)

From the covariance, the correlation can be computed: $r_{obs} \; or \; r_{xy} = \frac{Cov_{xy}}{s_x s_y}$

df = n -2.

For **effect size** you can simply interpret the magnitude of r since it ranges from -1 to +1. Values near -1 or +1 would be considered large effects while values close to 0 would be considered small.

It is also common practice to square the correlation coefficient, r^2. This is sometimes referred to as the coefficient of determination. It is like eta-squared for the t-test and represents the proportion of shared variance between the two variables. For example, for an r^2 value equal to .50, the researcher might conclude that the x and y variables share 50% of their variance. With r^2, it is reasonable to use Cohen's conventions for eta-squared.

For the **confidence interval** you must first convert r_{xy} (r_{obs}) to a z-score using Fisher's r-to-z transformation table. This gives you *r'*. You then build your CI around *r'*: $? \leq \text{pop } r' \leq ?$

$$r' \pm (z_{crit})\left(\sqrt{\frac{1}{n-3}}\right)$$

After you obtain your lower and upper bounds for *r'* you go back to Fisher's r-to-z transformation table and back-convert the *r'* values to *r* values. Your CI will then look like, $? \leq \rho \leq ?$. You judge the width (precision) of the interval based on a range of -1 to +1.

APA Style Example

As predicted depression and anxiety were linearly associated, $r(54) = .45$, $p < .01$ (95% CI: .18 to .66). The two variables shared 20% of their variance, which represents a large effect using Cohen's conventions.