Further Topics on Correlation and Simple Regression Analysis

Sunthud Pornprasertmanit Chulalongkorn University

Other Formula for Pearson's Correlation

You will learn that the correlation coefficient is the ratio of covariance of *X* and *Y* and the product of standard deviation of *X* and *Y*.

$$r = \frac{S_{XY}}{S_X S_Y}$$

If transforming raw data to standard score, the formula will be

$$r = \frac{\sum z_X z_Y}{n-1}$$

This formula states that the correlation coefficient is the covariance between standard scores of *X* and *Y*.

Another formula is

$$r = 1 - \left(\frac{\sum (z_X - z_Y)^2}{2(n-1)}\right)$$

The perfect positive relationship exists when all z_X and z_Y pairs of scores consist of two exactly equal values.

The degree of relationship will be a function of the departure from this "perfect" state.

Standardized Regression Coefficient

When transforming raw score to standard score and calculating regression coefficient, the regression equation will be

$$Y = \bar{Y} + b(X - \bar{X})$$

$$Y - \bar{Y} = b(X - \bar{X})$$

$$\frac{Y - \bar{Y}}{s_Y} = \frac{s_X}{s_Y} b \left(\frac{X - \bar{X}}{s_X}\right)$$

Author Note

This article was written in August 2007 for teaching in Introduction to Statistics in Psychology Class, Faculty of Psychology, Chulalongkorn University

Correspondence to Sunthud Pornprasertmanit. Email: psunthud@gmail.com

$$\hat{z}_Y = \beta z_X$$

The β in this equation is different from β in type II error.

This β tells about the effect of predictor variable. If *X* changes in one standard deviation, *Y* will change in β standard deviation.

In simple regression, the β is equal to correlation coefficient.

Hypothesis Testing for Correlation Coefficient

From lecture 3, the descriptive statistics for correlation coefficient is

$$r = \frac{S_{XY}}{S_X S_Y}$$

When estimating population correlation, the correlation coefficient will be

$$r = \frac{s_{XY}}{s_X s_Y}$$

Although the population correlation coefficient is equal to zero, the correlation statistics is not equal to zero by chance.

Therefore, the correlation statistics must be tested for confirming that this correlation is not stemmed from sampling error.

Null hypothesis	$H_0: \rho = 0$	
Alternative hypothesis	$H_1: \rho \neq 0$	(Two-tailed)
	$H_1: \rho > 0;\ \rho < 0$	(One-tailed)

From general form of *t* statistics,

$$t = \frac{\text{Statistics} - \text{Null hypothetical value}}{\text{Estimated Standard Error of Statistics}}$$

Statistics is correlation coefficient.

Null hypothetical value is 0.

Estimated standard error of statistics is

$$Var(r) = \frac{1 - r_{XY}^2}{n - 2}$$

Therefore,

$$t = \frac{r_{XY}\sqrt{n-2}}{\sqrt{1-r_{XY}^2}}$$

Degree of freedoms is equal to *n* - 2.

In statistics books, there are tables that show critical value of the Pearson r for desired alpha in both one-tailed and two-tailed test. See Table D6 in Kirk (2008)

Example

Assumption of this formula

1) Random sampling
2) The population distributions of *X* and *Y* are approximately normal
3) The relationship between *X* and *Y* is linear.
4) The distribution of *Y* for any value of *X* is normal with variance that does not depend on the *X* value selected. (homoscadasticity)
5) The null hypothetical value is equal to zero.

However, if null hypothetical value is not 0, this formula cannot be done because the sampling distribution is not *t* distribution and very skewed.

Fisher proposed the function transforming from *r* to *z* and sampling distribution of *z* is the same as normal distribution if *n* $\geq$ 10.

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

$$r = \frac{e^{2z\prime} - 1}{e^{2z\prime} + 1}$$

The Fisher *r*-to-*z*' transformation can be see in Table D7 in Kirk (2008)

Null hypothesis $H_0: \rho = \rho_0;\ Z = Z_0$

Alternative hypothesis $H_1: \rho \neq \rho_0; Z \neq Z_0$ (Two-tailed)

$H_1: \rho > \rho_0;\ \rho < \rho_0; Z > Z_0; Z < Z_0$ (One-tailed)

From general form of *z* statistic

$$z = \frac{\text{Statistics} - \text{Null hypothetical value}}{\text{Standard Error of Statistics}}$$

Statistics is transformed correlation coefficient.

Null hypothetical value is transformed null hypothetical value of correlation.

Estimated standard error of statistics is

$$Var(z') = \frac{1}{n-3}$$

The null hypothesis testing is

$$z = \sqrt{n-3}(z' - Z_0)$$

The confidence interval based on this null hypothesis testing is

A two-sided 100(1-α) % confidence interval for *Z* is given by

$$z' - z_{\alpha/2}\sqrt{\frac{1}{n-3}} < Z < z' + z_{\alpha/2}\sqrt{\frac{1}{n-3}}$$

Lower and upper one-sided 100(1-α) % confidence intervals for *Z* are given by

$$z' - z_{\alpha}\sqrt{\frac{1}{n-3}} < Z \qquad \text{and} \qquad Z < z' + z_{\alpha}\sqrt{\frac{1}{n-3}}$$

Example

Assumption of this null hypothesis testing and confidence interval

1) Random sampling
2) Null hypothetical value (ρ) is not too close to 1 or -1.
3) The population distributions of *X* and *Y* are approximately normal
4) The relationship between *X* and *Y* is linear.
5) The distribution of *Y* for any value of *X* is normal with variance that does not depend on the *X* value selected. (homoscadasticity)
6) The sample size *n* is greater than 10.

Cohen (1988) has suggested using *r*, a measure of the linear strength of association between two variables, to assess effect magnitude.

Small Strength of Association	*r* = .10
Medium Strength of Association	*r* = .30
Large Strength of Association	*r* = .50

Standard Error of Estimate in Regression Analysis

Standard error of estimate is the standard deviation of residuals.

$$S_{Y.X} = \sqrt{\frac{SS_{error}}{n}} = \sqrt{\frac{\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2}{n}}$$

$$S_{Y.X} = S_X\sqrt{1 - r_{YX}^2}$$

For the sake of parameter estimation, the standard error of estimate must be changed.

$$\hat{\sigma}_{Y.X} = \sqrt{\frac{SS_{error}}{n-2}} = \sqrt{\frac{\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2}{n-2}}$$

$$\hat{\sigma}_{Y.X} = \hat{\sigma}_Y\sqrt{1 - r_{YX}^2}$$

Hypothesis Testing for Simple Regression Coefficient

The simple regression equation is

$$\hat{Y} = b_0 + b_{YX}X$$

You must test $H_0: \rho = 0$ before run regression analysis. If the correlation is equal to 0, the predictor is not predicted more precise than $\bar{Y}$.

There are two null hypotheses testing in regression analysis.

$H_0: B_0 = B_0^*$ and $H_0: B_{YX} = B_{Y.X}^*$

The alternative hypotheses are

$H_0: B_0 \neq B_0^*$ and $H_0: B_{YX} \neq B_{Y.X}^*$ (Two-tailed)

$H_0: B_0 > B_0^*$ and $H_0: B_{YX} > B_{Y.X}^*$ (One-tailed)

$H_0: B_0 < B_0^*$ and $H_0: B_{YX} < B_{Y.X}^*$ (One-tailed)

These statistics are randomly distributed in *t*-distribution.

$$SE_{b_0} = \hat{\sigma}_{Y.X}\sqrt{\frac{1}{n} + \frac{\bar{X}^2}{(n-1)s_X^2}}$$

$$SE_{b_{YX}} = \frac{s_Y}{s_X}\sqrt{\frac{1 - r_{YX}^2}{n-2}}$$

Therefore, the null hypothesis testing for regression coefficient is

$$t = \frac{b_0 - B_0^*}{SE_{b_0}}$$

$$t = \frac{b_{Y.X} - B_{Y.X}^*}{SE_{b_{Y.X}}}$$

In both t statistics, *df* = *n* – 2.

The confidence intervals based on these null hypotheses testing are

A two-sided 100(1-α) % confidence intervals for b_0 and b_{YX} are given by

$$b_0 - t_{\alpha/2,df}SE_{b_0} < B_0 < b_0 + t_{\alpha/2,df}SE_{b_0}$$

$$b_{Y.X} - t_{\alpha/2,df}SE_{b_{Y.X}} < B_{Y.X} < b_{Y.X} + t_{\alpha/2,df}SE_{b_{Y.X}}$$

Lower and upper one-sided 100(1-α) % confidence intervals for b_0 and b_{YX} are given by

$$b_0 - t_{\alpha,df}SE_{b_0} < B_0 \qquad \text{and} \qquad B_0 < b_0 + t_{\alpha,df}SE_{b_0}$$

$$b_{Y.X} - t_{\alpha,df}SE_{b_{Y.X}} < B_{Y.X} \qquad \text{and} \qquad B_{Y.X} < b_{Y.X} + t_{\alpha,df}SE_{b_{Y.X}}$$

Example

Confidence Interval of Predicted Value

Sampling error was made by using the sample B_{YX} instead of the (unavailable) population regression coefficient. In addition to standard error of estimate, it will have more serious consequences for *X* values that are more distant from the *X* mean than for those near it.

The predicted value is distributed in t-distribution with *df* = *n* – 2.

$$SE_{\hat{Y}_i} = \hat{\sigma}_{Y.X}\sqrt{\frac{1}{n} + \frac{(X_i - \bar{X})^2}{(n-1)s_X^2}}$$

The null hypothesis testing may be useless. However, the confidence interval of predicted value is useful for estimation.

A two-sided 100(1-α) % confidence intervals for *Y* are given by

$$\hat{Y}_i - t_{\alpha/2,df}SE_{\hat{Y}_i} < Y < \hat{Y}_i + t_{\alpha/2,df}SE_{\hat{Y}_i}$$

Lower and upper one-sided 100(1-α) % confidence intervals for b_0 and b_{YX} are given by

$$\hat{Y}_i - t_{\alpha,df}SE_{\hat{Y}_i} < Y \qquad \text{and} \qquad Y < \hat{Y}_i + t_{\alpha,df}SE_{\hat{Y}_i}$$

Example