Chapter 816

Confidence Intervals for Point Biserial Correlation

Introduction

This routine calculates the sample size needed to obtain a specified width of a point biserial correlation coefficient confidence interval at a stated confidence level.

The **point biserial correlation** coefficient (ρ in this chapter) is the product-moment correlation calculated between a continuous random variable (Y) and a binary random variable (X). This correlation is related to, but different from, the **biserial correlation** proposed by Karl Pearson. In psychology, the point biserial correlation is often used as a measure of the degree of association between a trait or attribute and a measureable characteristic such as an ability to accomplish something.

Since it is a correlation, ρ ranges between plus and minus one. However, because of the discrete variable, the actual upper limit may be far less than one.

When ρ is used as a descriptive statistic, no special distributional assumptions need to be made about the variables (Y and X). When confidence intervals are calculated, it is assumed that the observation pairs are independent and that the values of Y are distributed normally conditional on the value of X. The distribution of Y when X = 1 is normal with mean μ_1 and variance σ^2, while the distribution of Y when X = 0 is normal with mean μ_0 and variance also σ^2.

If X is the result of a Bernoulli trial with probability of success (X = 1) *p*, then the design is said to be **random**. If X is set in advance, then the design is said to be **fixed**. This routine only calculates sample size for the random design.

Technical Details

Tate (1954, 1955) presents results that give the distribution of sample point biserial correlation *r* (assuming the continuous variables is conditional normal and n > 25) as approximately normal with mean ρ (population point biserial correlation) and variance

$$\sigma_r^2 = \frac{\rho^2 + 2P(1-P)(2-3\rho^2)}{4nP(1-P)}(1-\rho^2)^2$$

where *n* is the sample size and *P* is the probability that X = 1.

Confidence limits r_L and r_U are obtained using the usual formulas

$$r_L = r - z_{\alpha/2}\sigma_r$$

and

$$r_U = r + z_{\alpha/2}\sigma_r$$

One-sided limits may be obtained by replacing $\alpha/2$ by α.

Confidence Level

The confidence level, 1 – α, has the following interpretation. If thousands of samples of *n* items are drawn from a population using simple random sampling and a confidence interval is calculated for each sample, the proportion of those intervals that will include the true population correlation is 1 – α.

Procedure Options

This section describes the options that are specific to this procedure. These are located on the Design tab. For more information about the options of other tabs, go to the Procedure Window chapter.

Design Tab

The Design tab contains most of the parameters and options that you will be concerned with.

Solve For

Solve For

This option specifies the parameter to be solved for from the other parameters.

One-Sided or Two-Sided Interval

Interval Type

Specify whether the confidence interval for the population correlation is two-sided or one-sided. A one-sided interval is often called a **confidence bound** rather than a confidence interval because it only has one limit.

Two-Sided

The two-sided confidence interval is defined by two limits: an upper confidence limit (UCL) and a lower confidence limit (LCL).

These limits are constructed so that the designated proportion (confidence level) of such intervals will include the true population value.

Upper One-Sided

The upper confidence interval (or bound) is defined by a limit above the estimated parameter value. The limit is constructed so that the designated proportion (confidence level) of such limits has the true population value below them.

Lower One-Sided

The lower confidence interval (or bound) is defined by a limit below the estimated parameter value. The limit is constructed so that the designated proportion (confidence level) of such limits has the true population value above them.

Confidence

Confidence Level

The confidence level, 1 – α, has the following interpretation. If thousands of samples of n items are drawn from a population using simple random sampling and a confidence interval is calculated for each sample, the proportion of those intervals that will include the true population correlation is 1 – α.

Often, the values 0.95 or 0.99 are used. You can enter single values or a range of values such as *0.90, 0.95* or *0.90 to 0.99 by 0.01*.

Sample Size

N (Sample Size)

Enter one or more values for the sample size. This is the number of individuals selected at random from the population to be in the study.

You can enter a single value or a range of values.

Precision

Confidence Interval Width (Two-Sided)

This is the distance from the lower confidence limit to the upper confidence limit.

You can enter a single value or a list of values. The value(s) must be between 0 and 2.

Distance from r to Limit (One-Sided)

This is the distance from the sample correlation to the lower or upper limit of the confidence interval, depending on the Interval Type.

You can enter a single value or a list of values. The value(s) must be between 0 and 2.

Estimated Sample Values

r (Sample Point Biserial Correlation)

Enter the planning estimate of the sample point biserial correlation. This value can be obtained from prior studies, expert opinion, or as a reasonable guess. The sample size and width calculations assume that the value entered here is the actual correlation estimate obtained from the sample. The accuracy of your results will depend on the accuracy of this estimate.

The range of the values of the sample correlation that can be entered is -1 to 1.

You can enter a range of values such as 0.1 0.3 0.5 or 0.1 to 0.5 by .2.

P (Probability Dichotomous X = 1)

Specify the value of P, the probability that the dichotomous variable X = 1. Since this is a probability, it must be between 0 and 1.

You may enter a single value, a range, or a list.

Example 1 – Calculating Sample Size

Suppose a study is planned in which the researcher wishes to construct a two-sided 95% confidence interval for the point biserial correlation such that the width of the interval is no wider than 0.08. The researcher would like to examine a large range of *r* values to determine the effect of the correlation estimate on necessary sample size. Also, the researcher would like a report showing various values of *P*.

The goal is to determine the necessary sample size.

Setup

This section presents the values of each of the parameters needed to run this example. First, from the PASS Home window, load the **Confidence Intervals for Point Biserial Correlation** procedure window by expanding **Correlation**, then **Correlation**, then clicking on **Confidence Interval**, and then clicking on **Confidence Intervals for Point Biserial Correlation**. You may then make the appropriate entries as listed below, or open **Example 1** by going to the **File** menu and choosing **Open Example Template**.

Option	Value
Design Tab	
Solve For	**Sample Size**
Interval Type	**Two-Sided**
Confidence Level (1 – Alpha)	**0.95**
Confidence Interval Width (Two-Sided)	**0.08**
r (Sample Kendall’s Tau Correlation)	**0 0.1 0.3 0.5 0.7 0.9 0.95**
P (Probability Dichotomous X = 1)	**0.2 0.5 0.8**

Annotated Output

Click the Calculate button to perform the calculations and generate the following output.

Numeric Results

```
                                                       Sample
                                                        Point
               Sample                                Biserial       Prob       C.I.        C.I.
Conf             Size        Target        Actual        Corr      X = 1      Lower       Upper
Level               N         Width         Width           r          P      Limit       Limit
0.950            2401         0.080         0.080       0.000       0.20     -0.040       0.040
0.950            2401         0.080         0.080       0.000       0.50     -0.040       0.040
0.950            2401         0.080         0.080       0.000       0.80     -0.040       0.040
0.950            2355         0.080         0.080       0.100       0.20      0.060       0.140
0.950            2342         0.080         0.080       0.100       0.50      0.060       0.140
0.950            2355         0.080         0.080       0.100       0.80      0.060       0.140
0.950            2000         0.080         0.080       0.300       0.20      0.260       0.340
0.950            1899         0.080         0.080       0.300       0.50      0.260       0.340
.                   .             .             .           .          .          .           .
.                   .             .             .           .          .          .           .

Report Definitions
Confidence level is the proportion of confidence intervals (constructed with this same confidence level,
   sample size, etc.) that would contain the true correlation.
Sample Size N is the size of the sample drawn from the population.
Width is the distance from the lower limit to the upper limit.
Target Width is the value of the width that is entered into the procedure.
Actual Width is the value of the width that is obtained from the procedure.
r is the estimate of point biserial correlation.
Lower and Upper Limit are the lower and upper limits of the confidence interval.
P is the anticipated value of the probability that the dichotomous variable X = 1.
```

References

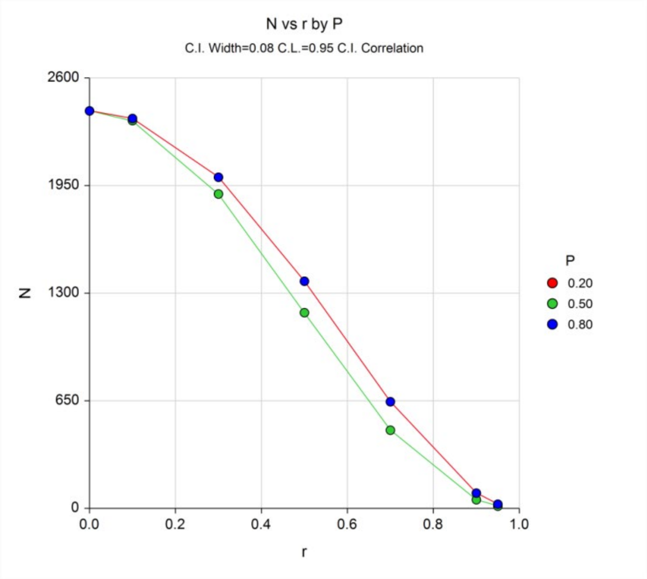
Tate, R. F. 1954. 'Correlation Between a Discrete and Continuous Variable. Point-Biserial Correlation.' Annals of Mathematical Statistics. Vol. 25, No. 3, pages 603-607.
Tate, R. F. 1955. 'Applications of ACorrelation Models for Biserial Data.' Journal of the American Statistical Association. Vol. 50, No. 272, pages 1078-1095.
Bonett, D. G. and Wright, T. A. 2000. 'Sample Size Requirements for Estimating Pearson, Kendall and Spearman Correlations.' Psychometrika, Vol 65, No 1 (March), 23-28.
Kraemer, H.C. 1980. 'Robustness of the Distribution Theory of the Product Moment Correlation Coefficient.', Journal of Educational Statistics, Volume 5, Number 2, pages 115-128.
Fisher, R. A. 1921. 'On the probable error of a coefficient of correlation deduced from a small sample.' Metron, i (4), 1-32.

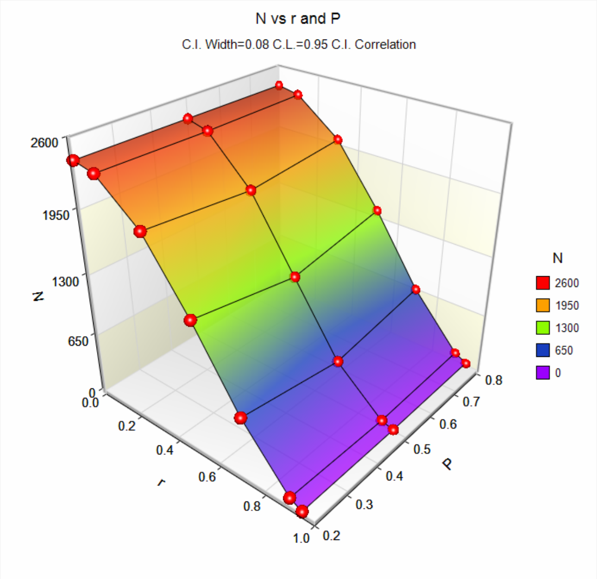
Summary Statements

A sample size of 2401 produces a two-sided 95% confidence interval with a width equal to 0.080 when the estimate of point biserial correlation is 0.000. The anticipated probability that the dichotomous variable X = 1 is 0.20.

This report shows the calculated sample size for each of the scenarios.

Plots Section





These plots show the sample size versus the sample correlation for the three values of P. It appears that the value of *r* contributes the most to the sample size.

Example 2 – Validation using Tate

Tate (1955), page 1085, gives example calculations of the limits of a two-sided confidence interval for the point biserial correlation when the confidence level is 99%, the sample point biserial correlation is 0.40, P is 0.65, and the interval is 0.19 to 0.61 for a width of 0.42. Their sample size is 100.

Setup

This section presents the values of each of the parameters needed to run this example. First, from the PASS Home window, load the **Confidence Intervals for Point Biserial Correlation** procedure window by expanding **Correlation**, then **Correlation**, then clicking on **Confidence Interval**, and then clicking on **Confidence Intervals for Point Biserial Correlation**. You may then make the appropriate entries as listed below, or open **Example 2** by going to the **File** menu and choosing **Open Example Template**.

Option	Value
Design Tab	
Solve For	**Sample Size**
Interval Type	**Two-Sided**
Confidence Level (1 – Alpha)	**0.99**
Confidence Interval Width (Two-Sided)	**0.42**
r (Sample Kendall's Tau Correlation)	**0.40**
P (Probability Dichotomous X = 1)	**0.65**

Annotated Output

Click the Calculate button to perform the calculations and generate the following output.

Numeric Results

				Sample			
				Point			
	Sample			Biserial	Prob	C.I.	C.I.
Conf	Size	Target	Actual	Corr	X = 1	Lower	Upper
Level	N	Width	Width	r	P	Limit	Limit
0.990	100	0.420	0.419	0.400	0.65	0.191	0.609

PASS also calculates the sample size to be 100.