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 measurable 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.

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

If the procedure window is not already open, use the PASS Home window to open it. The parameters for this example are listed below and are stored in the **Example 1** settings file. To load these settings to the procedure window, click **Open Example Settings File** in the Help Center or File menu.

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**

Output

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

Numeric Reports

Numeric Results

Solve For: Sample Size
Interval Type: Two-Sided

Confidence Level	Sample Size N	Confidence Interval Width		Sample Point Biserial Correlation	Probability Dichotomous X = 1	Confidence Interval Limits	
		Target	Actual	r	P	Lower	Upper
0.95	2401	0.08	0.080	0.00	0.2	-0.040	0.040
0.95	2401	0.08	0.080	0.00	0.5	-0.040	0.040
0.95	2401	0.08	0.080	0.00	0.8	-0.040	0.040
0.95	2355	0.08	0.080	0.10	0.2	0.060	0.140
0.95	2342	0.08	0.080	0.10	0.5	0.060	0.140
0.95	2355	0.08	0.080	0.10	0.8	0.060	0.140
0.95	2000	0.08	0.080	0.30	0.2	0.260	0.340
0.95	1899	0.08	0.080	0.30	0.5	0.260	0.340
0.95	2000	0.08	0.080	0.30	0.8	0.260	0.340
0.95	1372	0.08	0.080	0.50	0.2	0.460	0.540
0.95	1182	0.08	0.080	0.50	0.5	0.460	0.540
0.95	1372	0.08	0.080	0.50	0.8	0.460	0.540
0.95	644	0.08	0.080	0.70	0.2	0.660	0.740
0.95	472	0.08	0.080	0.70	0.5	0.660	0.740
0.95	644	0.08	0.080	0.70	0.8	0.660	0.740
0.95	92	0.08	0.080	0.90	0.2	0.860	0.940
0.95	52	0.08	0.080	0.90	0.5	0.860	0.940
0.95	92	0.08	0.080	0.90	0.8	0.860	0.940
0.95	25	0.08	0.079	0.95	0.2	0.911	0.989
0.95	13	0.08	0.079	0.95	0.5	0.911	0.989
0.95	25	0.08	0.079	0.95	0.8	0.911	0.989

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

Summary Statements

A single-group design will be used to obtain a two-sided 95% confidence interval for a single point biserial correlation coefficient. The anticipated probability that the dichotomous variable X equals 1 is 0.2. The sample point biserial correlation is assumed to be 0. To produce a confidence interval with a width of no more than 0.08, 2401 subjects will be needed.

Dropout-Inflated Sample Size

Dropout Rate	Sample Size N	Dropout-Inflated Enrollment Sample Size N'	Expected Number of Dropouts D
20%	2401	3002	601
20%	2401	3002	601
20%	2401	3002	601
20%	2355	2944	589
20%	2342	2928	586
20%	2355	2944	589
20%	2000	2500	500
20%	1899	2374	475
20%	2000	2500	500
20%	1372	1715	343
20%	1182	1478	296
20%	1372	1715	343
20%	644	805	161
20%	472	590	118
20%	644	805	161
20%	92	115	23
20%	52	65	13
20%	92	115	23
20%	25	32	7
20%	13	17	4
20%	25	32	7

Dropout Rate	The percentage of subjects (or items) that are expected to be lost at random during the course of the study and for whom no response data will be collected (i.e., will be treated as "missing"). Abbreviated as DR.
N	The evaluable sample size at which the confidence interval is computed. If N subjects are evaluated out of the N' subjects that are enrolled in the study, the design will achieve the stated confidence interval.
N'	The total number of subjects that should be enrolled in the study in order to obtain N evaluable subjects, based on the assumed dropout rate. After solving for N, N' is calculated by inflating N using the formula N' = N / (1 - DR), with N' always rounded up. (See Julious, S.A. (2010) pages 52-53, or Chow, S.C., Shao, J., Wang, H., and Lokhnygina, Y. (2018) pages 32-33.)
D	The expected number of dropouts. D = N' - N.

Dropout Summary Statements

Anticipating a 20% dropout rate, 3002 subjects should be enrolled to obtain a final sample size of 2401 subjects.

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.

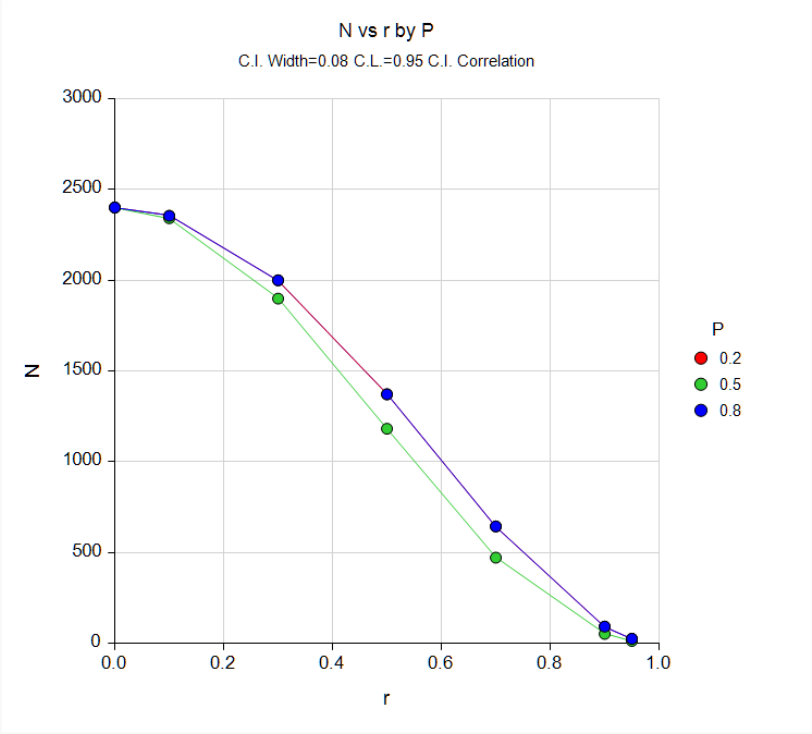
Kraemer, H.C. 1980. 'Robustness of the Distribution Theory of the Product Moment Correlation Coefficient.', Journal of Educational Statistics, Volume 5, Number 2, pages 115-128.

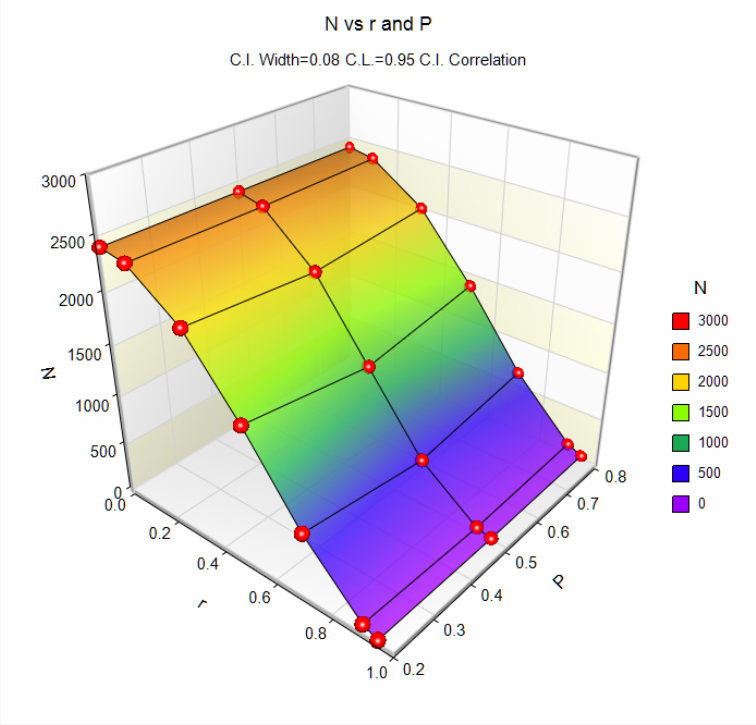
Fisher, R. A. 1921. 'On the probable error of a coefficient of correlation deduced from a small sample.' Metron, i (4), 1-32.

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

Plots Section

Plots





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 (1955)

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

If the procedure window is not already open, use the PASS Home window to open it. The parameters for this example are listed below and are stored in the **Example 2** settings file. To load these settings to the procedure window, click **Open Example Settings File** in the Help Center or File menu.

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**

Output

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

Numeric Results

Solve For: Sample Size
Interval Type: Two-Sided

Confidence Level	Sample Size N	Confidence Interval Width		Sample Point Biserial Correlation	Probability Dichotomous X = 1	Confidence Interval Limits	
		Target	Actual	r	P	Lower	Upper
0.99	100	0.42	0.419	0.4	0.65	0.191	0.609

PASS also calculates the sample size to be 100.