Chapter 807

Point Biserial Correlation Tests

Introduction

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 hypothesis tests are made, 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**.

Difference Between Linear Regression and Correlation

The point biserial correlation coefficient discussed in this chapter assumes that both X and Y are random variables. In the linear regression context, no statement is made about the distribution of X. In fact, X is not even a random variable. Instead, the values of X are set as part of the design. For example, a design might call for 20 men and 20 women to be included. Even though the same formula is used in this case, the results follow a different distribution with different sample size requirements. The analysis would then be termed linear regression and that procedure should be used to determine sample size and power.

Technical Details

The following results are found in Lev(1949) and Tate (1954). A random sample of n subjects is measured for the presence or absence of the trait (X) and the level of an ability (Y). This gives rise to n pairs of observations: (X_i, Y_i), $i = 1, 2, \ldots, n$.

Sample Point Biserial Correlation Coefficient

The point biserial correlation coefficient, r, is calculated using the common product-moment correlation

$$r = \frac{\sum(X - \bar{X})(Y - \bar{Y})}{\sqrt{\sum(X - \bar{X})^2 \sum(Y - \bar{Y})^2}}$$

$$= \frac{(\bar{Y}_1 - \bar{Y}_0)\sqrt{\frac{n_1 n_0}{n}}}{\sum_{i=1}^{n}(Y_i - \bar{Y})^2}$$

Random Design

If it is assumed that

1. The binomial variable X takes on the value 1 with probability p and 0 with probability $q = 1 - p$.
2. The condition distribution of Y given X = 1 is $N(\mu_1, \sigma)$ and the condition distribution of Y given X = 0 is $N(\mu_0, \sigma)$.

The Tate (1954) provides results for the test statistic t calculated as

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

When ρ is 0, t follows Student's t distribution with $n - 2$ degrees of freedom. When ρ is not 0, the distribution of t is a weighted sum of non-central t distributions each with degrees of freedom $n - 2$ and noncentrality parameter δ_R given by

$$\delta_R = \frac{\rho}{\sqrt{1-\rho^2}}\sqrt{\frac{n_1 n_0}{npq}}$$

The weights are based on the binomial distribution of X.

Thus, the power of an upper, one-sided test of H_0: $\rho = \rho_0$ vs. H_1: $\rho > \rho_0$ computed at $\rho = \rho_1$ is

$$\varphi(p, \rho_1) = \sum_{n_1=0}^{n} \binom{n}{n_1} p^{n_1} q^{n_0} \int_{t_\alpha}^{\infty} h(t; n_1, n, p, \rho_1) dt$$

where $h(...)$ is the density of the non-central t distribution with $n - 2$ degrees of freedom and non-centrality δ_R, and t_α is chosen so that $\varphi(p, \rho_0) = \alpha$.

The sample size can be solved from the power function using a binary search algorithm.

Example 1 – Finding the Power

Suppose a study will be run to test whether the point biserial correlation between a random binary variable (X) and continuous variable (Y) is significantly different from zero. The researchers want to investigate what the power will be for a variety of sample sizes (5, 10, 20, 40, 80, 140, 200, 250) when alpha is 0.50. They want to calculate the power when ρ1 is actually 0.2, 0.4, and 0.6.

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	**Power**
Assume X's are	**Random**
Alternative Hypothesis	**ρ1 ≠ ρ0**
Alpha	**0.05**
N (Sample Size)	**5 10 20 40 80 140 200 250**
ρ0 (Correlation|H0)	**0.0**
ρ1 (Correlation|H1)	**0.2 0.4 0.6**
P (Probability X = 1)	**0.5**

Plots Tab – 2D Plots	
X-Y Plots	**Click the Plot Setup button** (Scatter Plot Format window appears)
Y Axis Tab	**Click on this tab. Y – Axis Vertical appears**
Axis: Min:	**Set to 0**
Axis: Max:	**Set to 1**
OK button	**Click to save the settings and close this window**

Plots Tab – 3D Plots	
X-Y-Z Plots	**Click the Plot Setup button** (3D Surface Plot Format window appears)
3D Surfact Plot Tab	**Click on this tab. 3D Surface Plot window appears**
Point Symbols	**Uncheck this option**
Y Axis Tab	**Click on this tab. Y – Axis Vertical appears**
Axis: Min:	**Set to 0**
Axis: Max:	**Set to 1**
OK button	**Click to save the settings and close this window**

Output

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

Numeric Reports

Numeric Results

Solve For:	Power
Dichotomous Variable Type:	Random
Hypotheses:	H0: ρ1 = ρ0 vs. H1: ρ1 ≠ ρ0

Power	Sample Size N	Point Biserial Correlation ρ0	Point Biserial Correlation ρ1	Alpha	Probability X = 1 P
0.0602	5	0	0.2	0.05	0.5
0.0843	10	0	0.2	0.05	0.5
0.1345	20	0	0.2	0.05	0.5
0.2373	40	0	0.2	0.05	0.5
0.4334	80	0	0.2	0.05	0.5
0.6663	140	0	0.2	0.05	0.5
0.8174	200	0	0.2	0.05	0.5
0.8941	250	0	0.2	0.05	0.5
0.0967	5	0	0.4	0.05	0.5
0.2117	10	0	0.4	0.05	0.5
0.4365	20	0	0.4	0.05	0.5
0.7564	40	0	0.4	0.05	0.5
0.9692	80	0	0.4	0.05	0.5
0.9992	140	0	0.4	0.05	0.5
1.0000	200	0	0.4	0.05	0.5
1.0000	250	0	0.4	0.05	0.5
0.1868	5	0	0.6	0.05	0.5
0.5052	10	0	0.6	0.05	0.5
0.8687	20	0	0.6	0.05	0.5
0.9952	40	0	0.6	0.05	0.5
1.0000	80	0	0.6	0.05	0.5
1.0000	140	0	0.6	0.05	0.5
1.0000	200	0	0.6	0.05	0.5
1.0000	250	0	0.6	0.05	0.5

Power	The probability of rejecting a false null hypothesis when the alternative hypothesis is true.
N	The size of the sample drawn from the population.
ρ0	The value of the point biserial correlation under the null hypothesis (H0).
ρ1	The value of the point biserial correlation under the alternative hypothesis (H1).
Alpha	The probability of rejecting a true null hypothesis.
P	The probability that the dichotomous X = 1.

Summary Statements

A point biserial correlation (single group, continuous Y versus binary X) design will be used to test whether the point biserial correlation is different from 0 (H0: ρ = 0 versus H1: ρ ≠ 0). The comparison will be made using a two-sided, one-sample point biserial correlation test, with a Type I error rate (α) of 0.05. The probability that the binary (dichotomous) variable will be equal to 1 is assumed to be 0.5. To detect a point biserial correlation of 0.2 with a sample size of 5, the power is 0.0602.

Dropout-Inflated Sample Size

Dropout Rate	Sample Size N	Dropout-Inflated Enrollment Sample Size N'	Expected Number of Dropouts D
20%	5	7	2
20%	10	13	3
20%	20	25	5
20%	40	50	10
20%	80	100	20
20%	140	175	35
20%	200	250	50
20%	250	313	63

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 power is computed (as entered by the user). If N subjects are evaluated out of the N' subjects that are enrolled in the study, the design will achieve the stated power.
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. 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, 7 subjects should be enrolled to obtain a final sample size of 5 subjects.

References

Lev, Joseph. 1949. 'The Point Biserial Coefficient of Correlation.' Annals of Mathematical Statistics. Vol. 20, No. 1, pages 125-126.

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 Correlation Models for Biserial Data.' Journal of the American Statistical Association. Vol. 50, No. 272, pages 1078-1095.

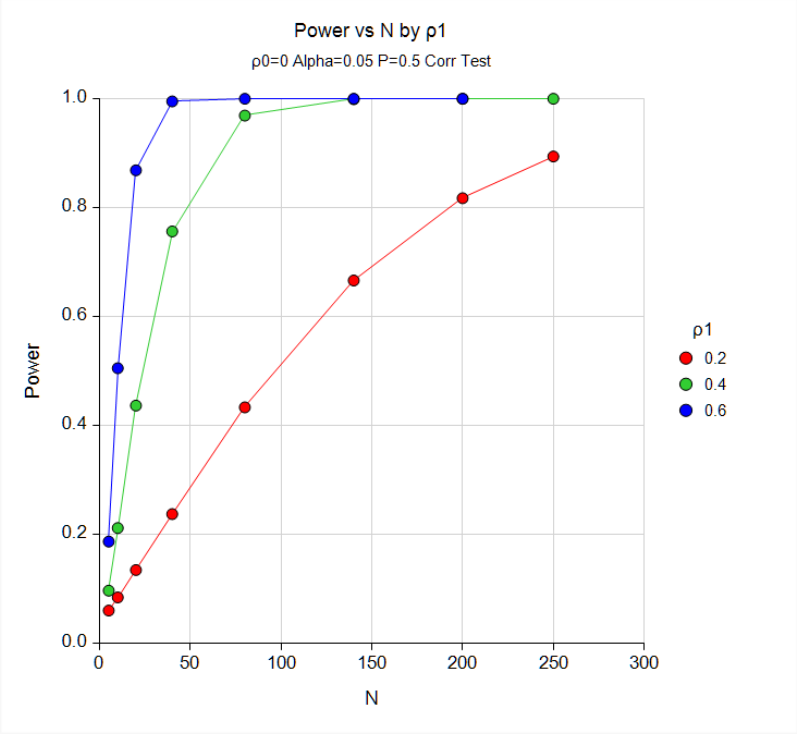
Kraemer, H.C. 1980. 'Robustness of the Distribution Theory of the Product Moment Correlation Coefficient.', Journal of Educational Statistics, Volume 5, Number 2, pages 115-128.

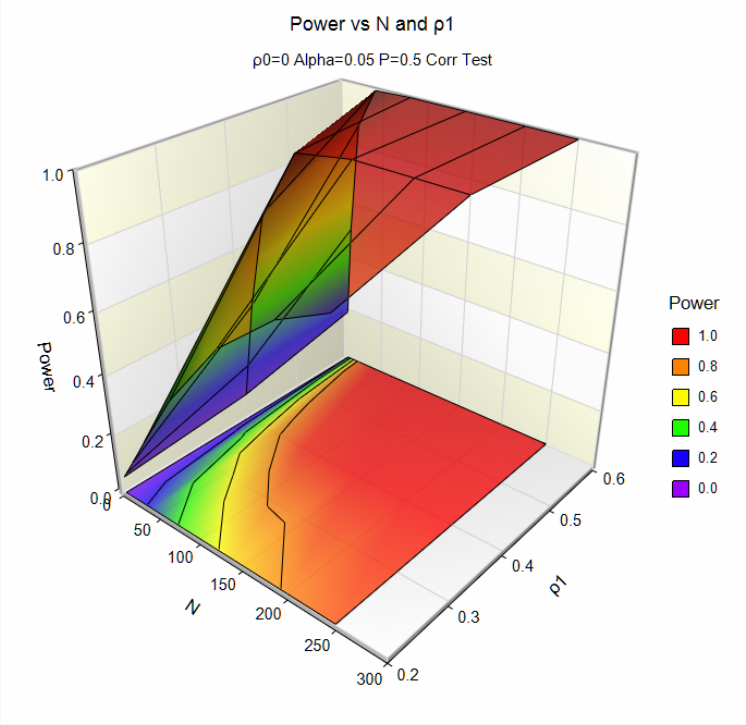
Gradstein, Mark. 1986. 'Maximal Correlation between Normal and Dichotomous Variables.', Journal of Educational Statistics, Volume 11, Number 4, pages 259-261.

This report shows the values of each of the parameters, one scenario per row. The values from this table are plotted in the charts below.

Plots Section

Plots





These plots show both a two-dimensional and a three-dimensional depiction of the relationship between power, sample size, and ρ1.

Example 2 – Validation using Tate (1955)

Tate (1955) page 1083 presents an example in which the power of a point biserial correlation coefficient is calculated. This example sets N = 10, alpha = 0.10, p = 1/3, ρ0 = 0, and ρ1 = 0.707. Tate calculates a power of 83.2% for a two-sided test.

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	**Power**
Assume X's are	**Random**
Alternative Hypothesis	**ρ1 ≠ ρ0**
Alpha	**0.1**
N (Sample Size)	**10**
ρ0 (Correlation|H0)	**0**
ρ1 (Correlation|H1)	**0.707**
P (Probability X = 1)	**0.3333333**

Output

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

Numeric Results

Solve For:	Power
Dichotomous Variable Type:	Random
Hypotheses:	H0: ρ1 = ρ0 vs. H1: ρ1 ≠ ρ0

	Sample Size	Point Biserial Correlation			Probability X = 1
Power	N	ρ0	ρ1	Alpha	P
0.83511	10	0	0.707	0.1	0.33333

The power of 0.8351 matches Tate's results to two decimal places. This is very good considering Tate explains in the article that they were using an approximation for the non-central t distribution.