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 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 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(μ_1, σ) and the condition distribution of Y given X = 0 is N(μ_0, σ).

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: ρ = ρ_0 vs H_1: ρ > ρ_0 computed at ρ = ρ_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.

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 calculated from the values of the other parameters. Under most conditions, you would either select *Power* or *Sample Size*.

Select *Sample Size* when you want to determine the sample size needed to achieve a given power and alpha error level.

Select *Power* when you want to calculate the power.

Dichotomous Variable Type

Assume X's are

This option specifies whether the X's are *Random* or *Fixed*.

Select *Random* when X is the realization of a chance event.

Select *Fixed* when X is set for each row by the experimenter (no chance event).

Test Direction

Alternative Hypothesis

This option specifies the alternative hypothesis. This implicitly specifies the direction of the hypothesis test. The null hypothesis is H_0: ρ0 = ρ1.

Note that the alternative hypothesis enters into power calculations by specifying the rejection region of the hypothesis test. Its accuracy is critical.

Possible selections for H_1 are:

- **ρ1 ≠ ρ0**

  This is the most common selection. It yields the *two-tailed* test. Use this option when you are testing whether the correlation values are different, but you do not want to specify beforehand which correlation is larger.

- **ρ1< ρ0**

  This option yields a *one-tailed* test.

- **ρ1> ρ0**

  This option also yields a *one-tailed* test.

Power and Alpha

Power

This option specifies one or more values for power. Power is the probability of rejecting a false null hypothesis, and is equal to one minus Beta. Beta is the probability of a type-II error, which occurs when a false null hypothesis is not rejected. In this procedure, a type-II error occurs when you fail to reject the null hypothesis of equal correlations when in fact they are different.

Values must be between zero and one. Historically, the value of 0.80 (Beta = 0.20) was used for power. Now, 0.90 (Beta = 0.10) is also commonly used.

A single value may be entered here or a range of values such as *0.8 to 0.95 by 0.05* may be entered.

Alpha

This option specifies one or more values for the probability of a type-I error. A type-I error occurs when you reject the null hypothesis of equal correlations when in fact they are equal.

Values of alpha must be between zero and one. Historically, the value of 0.05 has been used for alpha. This means that about one test in twenty will falsely reject the null hypothesis. You should pick a value for alpha that represents the risk of a type-I error you are willing to take in your experimental situation.

You may enter a range of values such as *0.01 0.05 0.10* or *0.01 to 0.10 by 0.01*.

Sample Size

N (Sample Size)

This option specifies the number of observations in the sample. Each observation is made up of two values: an X value of 0 or 1 and a continuous Y value. The minimum value is 4.

Point Biserial Correlations

ρ0 (Correlation|H0)

Specify the value of ρ0. Note that the range of the correlation is between plus and minus one. This value is usually set to zero.

ρ1 (Correlation|H1)

Specify the value of ρ1, the population correlation under the alternative hypothesis. Note that the range of the correlation is between plus and minus one. The difference between ρ0 and ρ1 is being tested by this significance test.

You can enter a range of values separated by blanks or commas.

Dichotomous Variable (X)

P (Probability X = 1)

Specify the value of *p*, the probability that X = 1 when you have a random design. Since this is a probability, it must be between 0 and 1.

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

This section presents the values of each of the parameters needed to run this example. First, from the PASS Home window, load the **Point Biserial Correlation Tests** procedure window by expanding **Correlation**, then **Correlation**, then clicking on **Test (Inequality)**, and then clicking on **Point Biserial Correlation Tests**. 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	**Power**
Assume X's are	**Random**
Alternative Hypothesis	**ρ0 ≠ ρ1**
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**
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

Annotated Output

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

Numeric Results

Numeric Results Assuming the Dichotomous Variable is Random and H1: ρ0 ≠ ρ1

		Point	Point		
	Sample	Biserial	Biserial		Prob
	Size	Corr|H0	Corr|H1		X=1
Power	**N**	**ρ0**	**ρ1**	**Alpha**	**P**
0.0602	5	0.0000	0.2000	0.0500	0.5000
0.0843	10	0.0000	0.2000	0.0500	0.5000
0.1345	20	0.0000	0.2000	0.0500	0.5000
0.2373	40	0.0000	0.2000	0.0500	0.5000
0.4334	80	0.0000	0.2000	0.0500	0.5000
0.6663	140	0.0000	0.2000	0.0500	0.5000
0.8174	200	0.0000	0.2000	0.0500	0.5000
0.8941	250	0.0000	0.2000	0.0500	0.5000
.	.	.	.	.	.
.	.	.	.	.	.

Report Definitions

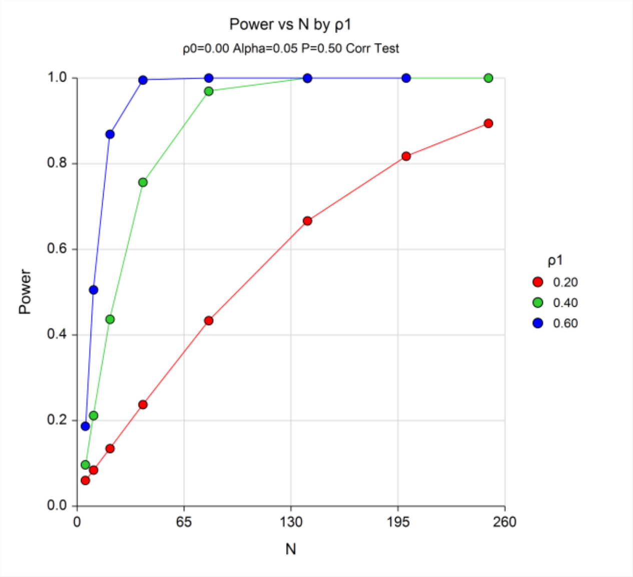
Power is the probability of rejecting a false null hypothesis.
N is the size of the sample drawn from the population.
ρ0 is the value of the point biserial correlation under the null hypothesis (H0).
ρ1 is the value of the point biserial correlation under the alternative hypothesis (H1).
Alpha is the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

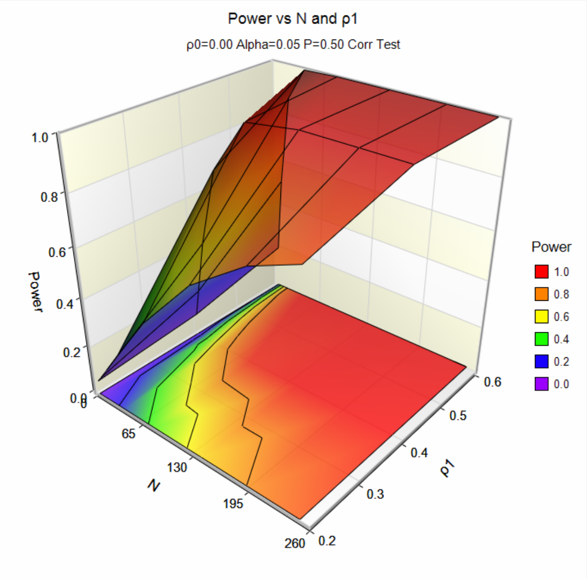
Summary Statements

A sample size of 5 achieves 6% power to detect the difference between the null hypothesis point biserial correlation of 0.0000 and the alternative hypothesis point biserial correlation of 0.2000 using a two-sided hypothesis test with a significance level of 0.0500. The probability that the dichotomous variable will be equal to 1 is assumed to be 0.5000.

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





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

Example 2 – Validation using Tate

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

This section presents the values of each of the parameters needed to run this example. First, from the PASS Home window, load the **Point Biserial Correlation Tests** procedure window by expanding **Correlation**, then **Correlation**, then clicking on **Test (Inequality)**, and then clicking on **Point Biserial Correlation Tests**. 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	**Power**
Assume X's are	**Random**
Alternative Hypothesis	**ρ0 ≠ ρ1**
Alpha	**0.1**
N (Sample Size)	**10**
ρ0 (Correlation|H0)	**0.0**
ρ1 (Correlation|H1)	**0.707**

Output

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

Numeric Results

```
Numeric Results Assuming the Dichotomous Variable is Random and H1: ρ0 ≠ ρ1

                        Point      Point
           Sample     Biserial   Biserial                 Prob
             Size     Corr|H0    Corr|H1                   X=1
Power           N          ρ0         ρ1       Alpha         P
0.8351         10      0.0000     0.7070      0.1000    0.3333
```

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.