

Chapter 302

Point-Biserial and Biserial Correlations

Introduction

This procedure calculates estimates, confidence intervals, and hypothesis tests for both the point-biserial and the biserial correlations.

The *point-biserial correlation* is a special case of the product-moment correlation in which one variable is continuous and the other variable is binary (dichotomous). The categories of the binary variable do not have a natural ordering. For example, the binary variable gender does not have a natural ordering. That is, it does not matter whether the males are coded as a zero or a one. Such variables are often referred to as nominal binary variables. It is assumed that the continuous data within each group created by the binary variable are normally distributed with equal variances and possibly different means.

The *biserial correlation* has a different interpretation which may be explained with an example. Suppose you have a set of bivariate data from the bivariate normal distribution. The two variables have a correlation sometimes called the product-moment correlation coefficient. Now suppose one of the variables is dichotomized by creating a binary variable that is zero if the original variable is less than a certain variable and one otherwise. The biserial correlation is an estimate of the original product-moment correlation constructed from the point-biserial correlation. For example, you may want to calculate the correlation between IQ and the score on a certain test, but the only measurement available is whether the test was passed or failed. You could then use the biserial correlation to estimate the more meaningful product-moment correlation.

The formulas used are found in Tate (1954, 1955), Sheskin (2011), and an article by Kraemer (2006).

Technical Details

Point-Biserial Correlation

Suppose you want to find the correlation between a continuous random variable Y and a binary random variable X which takes the values zero and one. Assume that n paired observations (Y_k, X_k) , $k = 1, 2, \dots, n$ are available. If the common product-moment correlation r is calculated from these data, the resulting correlation is called the *point-biserial correlation*.

Sheskin (2011) gives the formula for the point-biserial correlation coefficient as

$$r_{pb} = \left(\frac{\bar{Y}_1 - \bar{Y}_0}{s_Y} \right) \sqrt{\frac{np_0(1-p_0)}{n-1}}$$

Point-Biserial and Biserial Correlations

where

$$s_Y = \sqrt{\frac{\sum_{k=1}^n (Y_k - \bar{Y})^2}{n-1}}$$

$$\bar{Y} = \frac{\sum_{k=1}^n Y_k}{n}$$

$$p_1 = \frac{\sum_{k=1}^n X_k}{n}$$

$$p_0 = 1 - p_1$$

Tate (1954) shows that, for large samples, the distribution of r_{pb} is normal with mean ρ and variance

$$\sigma_r^2 = \frac{(1 - \rho^2)^2}{n} \left[1 + \rho^2 \left(\frac{1 - 6p_0(1 - p_0)}{4p_0(1 - p_0)} \right) \right]$$

This population variance can be estimated by substituting the sample value r_{pb} for ρ . An approximate confidence interval based on the normal distribution can be calculated from these quantities using

$$r_{pb} \pm z_{\alpha/2} \sqrt{\frac{(1 - r_{pb}^2)^2}{n} \left[1 + r_{pb}^2 \left(\frac{1 - 6p_0(1 - p_0)}{4p_0(1 - p_0)} \right) \right]}$$

The hypothesis that $\rho = 0$ can be tested using the following test which is equivalent to the two-sample t-test.

$$t_{pb} = \frac{r_{pb}\sqrt{n-2}}{\sqrt{1 - r_{pb}^2}}$$

This test statistic follows Student's t distribution with $n - 2$ degrees of freedom.

Biserial Correlation

Suppose you want to find the correlation between a pair of bivariate normal random variables when one has been dichotomized. Sheskin (2011) states that the biserial correlation can be calculated from the point-biserial correlation r_{pb} using the formula

$$r_b = \left(\frac{r_{pb}}{h} \right) \sqrt{p_0(1 - p_0)}$$

where

$$h = \frac{e^{-u^2/2}}{\sqrt{2\pi}}$$

$$Pr[Z \geq u | Z \sim N(0,1)] = p_1$$

Kraemer (2006) gives a method for constructing a large sample confidence interval for ρ_b which is described as follows. Let $g(x)$ be Fisher's z-transformation

$$g(x) = \frac{1}{2} \ln \left(\frac{1+x}{1-x} \right)$$

then

$$g \left(\frac{2r_b}{\sqrt{5}} \right) \sim N \left[g \left(\frac{2\rho_b}{\sqrt{5}} \right), \frac{5}{4n} \right]$$

Point-Biserial and Biserial Correlations

It follows that a $(1-\alpha)\%$ confidence interval for g , denote G_1 and G_2 , can be calculated using

$$G_1 = g\left(\frac{2r_b}{\sqrt{5}}\right) - |z_{\alpha/2}| \sqrt{\frac{5}{4n}}$$

and

$$G_2 = g\left(\frac{2r_b}{\sqrt{5}}\right) + |z_{\alpha/2}| \sqrt{\frac{5}{4n}}$$

These limits can then be inverted to obtain corresponding confidence limits for ρ_b . The result is

$$CL_1 = \frac{\sqrt{5}}{2} \left(\frac{e^{2G_1} - 1}{e^{2G_1} + 1} \right)$$

$$CL_2 = \frac{\sqrt{5}}{2} \left(\frac{e^{2G_2} - 1}{e^{2G_2} + 1} \right)$$

A large sample z-test of $\rho_b = 0$ based on $g(x)$ can be constructed as follows

$$z = \frac{g\left(\frac{2r_b}{\sqrt{5}}\right)}{\sqrt{\frac{5}{4n}}}$$

Procedure Options

This section describes the options available in this procedure.

Variables Tab

This panel specifies the variables used in the analysis.

Input Type

There are three ways to organize your data for use by this procedure. Select the type that reflects the way your data is presented on the spreadsheet.

One or More Continuous Variables and a Binary Variable

The continuous data is in one variable (column) and the binary group identification is in another variable. Each row contains the values for one subject. If multiple continuous variables are selected, a separate analysis is made for each. If the binary variable has more than two levels (unique values), a separate analysis is made for each pair.

Example Dataset

Continuous	Binary
12	0
35	0
24	0
26	1
44	1
33	1

Point-Biserial and Biserial Correlations

Two Continuous Variables, One for each Binary Group

The continuous values for each binary group are in separate variables (columns). Each cell of the spreadsheet gives the entry for a different subject.

Example Dataset

Binary0	Binary1
12	26
35	44
19	36
24	33

Two or More Continuous Variables used Two at a Time

The continuous values for each binary group are in separate variables (columns). Each cell of the spreadsheet gives the entry for a different subject. A separate analysis is conducted for each pair of variables.

Example Dataset

Binary1	Binary2	Binary3
12	26	25
35	44	29
19	36	41
24	33	28

Variables

(Input Type: One or More Continuous Variables and a Binary Variable)

Continuous Variable(s)

Specify one or more variables (columns) containing the continuous data values. (The binary group identification is given in another variable.) Each row contains the values for one subject.

If multiple continuous variables are selected, a separate analysis is made for each.

Example Dataset

Continuous	Binary
12	0
35	0
19	0
24	0
26	1
44	1
36	1
33	1

Binary Variable

Specify the variable that defines the binary grouping of the continuous data. The values in this variable may be text or numeric. If they are text, they will be assigned a numeric 0 or 1 alphabetically. Numeric values must be assigned because correlation is only defined for numeric values.

The binary identification is in this variable and the continuous values are in another variable.

Rows missing a binary value or a continuous value will be ignored.

If the binary variable has more than two levels, a separate analysis is made for each pair of categories.

Point-Biserial and Biserial Correlations

Example Dataset

Continuous	Binary
12	1
35	1
19	1
24	1
26	2
44	2
36	2
33	2

(Input Type: Two Continuous Variables, One for each Binary Group)

Binary 0 (or 1) Continuous Variable

Specify the variable that contains the continuous data values for the 0 (or 1) category of the binary variable. The number of values in each column need not be the same.

(Input Type: Two or More Continuous Variables used Two at a Time)

Continuous Variable(s)

Specify two or more variables containing the continuous data values. All continuous values for one binary category are placed in a single variable (column). The first variable will be assigned to the '0' category and the second variable will be assigned to the '1' category.

If more than two variables are specified, a separate analysis will be made for each pair. Each variable listed in the variable box will be paired with every other variable in the box.

Example Dataset

Group1	Group2	Group3
12	26	25
35	44	29
19	36	41
24	33	
32		

Reports Tab

The following options control which reports and plots are displayed.

Select Reports

Point-Biserial ... Tests of Normality and Equal Variance Assumptions

These options specify which numeric reports are displayed.

Confidence Level and Alphas

Confidence Level

This confidence level is used for confidence intervals that are displayed. Typical confidence levels are 90%, 95%, and 99%, with 95% being the most common.

Test Alpha

Alpha is the significance level used in the hypothesis tests. A value of 0.05 is most commonly used, but 0.1, 0.025, 0.01, and other values are sometimes used. Typical values range from 0.001 to 0.20.

Point-Biserial and Biserial Correlations

Assumptions Alpha

Assumptions Alpha is the significance level used in all the assumptions tests. A value of 0.05 is typically used for hypothesis tests in general, but values other than 0.05 are often used for the case of testing assumptions. Typical values range from 0.001 to 0.20.

Report Options Tab

The following options control the formatting of the reports.

Report Options

Variable Names

This option lets you select whether to display variable names, variable labels, or both.

Decimal Places

Correlations – Test Statistics

These options allow you to specify the number of decimal places directly or based on the significant digits. If one of the Auto options is used, the ending zero digits are not shown. For example, if ‘Auto (Up to 7)’ is chosen, 0.0500 is displayed as 0.05 and 1.314583689 is displayed as 1.314584.

The output formatting system is not always designed to accommodate ‘Auto (Up to 13)’, and if chosen, this will likely lead to lines that run on to a second line. This option is included, however, for the rare case when a very large number of decimals is needed.

Plots Tab

These options let you specify which plots are displayed.

Plot to Check Model

Y vs X

These options control whether the Y vs X scatter plot is displayed, its size, and its format.

Click the large plot format button to change the plot settings.

Plots to Check Assumptions

Histogram, Probability Plot, and Box Plot

These options control whether the corresponding plot is displayed, its size, and its format. Click the large plot format button to change the plot settings.

Example 1 – Correlating Test Result with IQ

This example correlates the IQ scores of 100 subjects with their result on a pass-fail test. The researcher will quantify the correlation using the point-biserial correlation coefficient. These data are contained on the *IQ Test* dataset.

You may follow along here by making the appropriate entries or load the completed template **Example 1** by clicking on Open Example Template from the File menu of the Point-Biserial and Biserial Correlation window.

1 Open the IQTest dataset.

- From the File menu of the NCSS Data window, select **Open Example Data**.
- Click on the file **IQTest.NCSS**.
- Click **Open**.

2 Open the Point-Biserial and Biserial Correlations window.

- Using the Analysis menu or the Procedure Navigator, find and select the **Point-Biserial and Biserial Correlations** procedure.
- On the menus, select **File**, then **New Template**. This will fill the procedure with the default template.

3 Specify the variables.

- On the procedure window, select the **Variables tab**.
- Select **One or More Continuous Variables and a Binary Variable** as the **Input Type**.
- Double-click in the **Continuous Variable(s)** box. This will bring up the variable selection window.
- Select **IQ** from the list of variables and then click **Ok**.
- Double-click in the **Binary Variable** box. This will bring up the variable selection window.
- Select **Test** from the list of variables and then click **Ok**.

4 Run the procedure.

- From the Run menu, select **Run Procedure**. Alternatively, just click the green Run button.

Point-Biserial and Biserial Correlations

Point-Biserial and Biserial Correlations									
Continuous Variable = IQ, Binary Variable = Test									
Type	Correlation r	Lower 95.0% C.L. of ρ	Upper 95.0% C.L. of ρ	Std Dev of ρ	r^2	Count N	N0/N P	Test for $\rho = 0$	Prob Level
Pt-Biserial	0.7435	0.6690	0.8181	0.0380	0.5529	100	0.5000	11.008	0.0000
Biserial	0.9319	0.8421	0.9943		0.8684	100	0.5000	10.729	0.0000

This report shows the point-biserial correlation and associated confidence interval and hypothesis test on the first row. It shows the biserial correlation and associated confidence interval and hypothesis test on the second row.

Type

The type of correlation coefficient shown on this row. Note that, although the names point-biserial and biserial sound similar, these are two different correlations that come from different models.

Correlation

The computed values of the point-biserial correlation and biserial correlation. Note that since the assignment of the zero and one to the two binary variable categories is arbitrary, the sign of the point-biserial correlation can be ignored. This is not true of the biserial correlation.

Point-Biserial and Biserial Correlations

Lower and Upper 95% C.L. of ρ

These are the lower and upper limits of a two-sided, 95% confidence interval for the corresponding correlation.

Std Dev of ρ

This is the standard deviation of the estimate of the point-biserial correlation. This value is not available for the biserial correlation.

r^2

This is the r-squared value for the correlation presented on this row. R-squared is a measure of the strength of the relationship.

Count N

This is the total sample size.

N0/N P

This is the proportion of the sample that is in the group defined by the binary variable being 0. It is the value of p_0 in the formulas presented earlier in the chapter.

Test for $\rho = 0$

This is value of the test statistic used to test the hypothesis that the correlation is zero. For the point-biserial correlation, this is the value of the t-test with $N - 2$ degrees of freedom. It is identical to the two-sample t-test for testing whether the means are different.

For the biserial correlation, this is the value of the z-test which is based on the standard normal distribution.

Prob Level

This is the p-value of the hypothesis test mentioned above. If it is less than 0.05 (or whatever value you choose), then the test is 'significant' and the null hypothesis that the correlation is zero is rejected.

Means, Standard Deviations, and Confidence Intervals of Means

Means, Standard Deviations, and Confidence Intervals of Means

Continuous Variable = IQ, Binary Variable = Test

Name	Count	Mean	Standard Deviation	Lower 95.0% C.L.	Upper 95.0% C.L.
Test=0	50	100.24	5.227713	98.7543	101.7257
Test=1	50	111.22	4.734976	109.8743	112.5657
Combined	100	105.73	7.420767	104.2576	107.2024
Difference	100	10.98	4.987433	9.000521	12.95948

This report shows the descriptive statistics of the two individual groups, the combination of both groups, and the difference between the two groups.

Tests of Normality and Equal Variance

Tests of Normality and Equal Variance

Continuous Variable = IQ, Binary Variable = Test

Assumption	Test Name	Test Value	Prob Level	Conclusion ($\alpha = 0.050$)
Normality of Test=0	Shapiro-Wilk	0.977	0.4359	Cannot reject normality
Normality of Test=1	Shapiro-Wilk	0.972	0.2894	Cannot reject normality
Equal Variances	Brown-Forsythe	0.182	0.6710	Cannot reject equal variances

This report presents the results of the Shapiro-Wilk normality test of each group as well as the Brown-Forsythe Equal Variance test (sometimes called the Modified-Levene test).

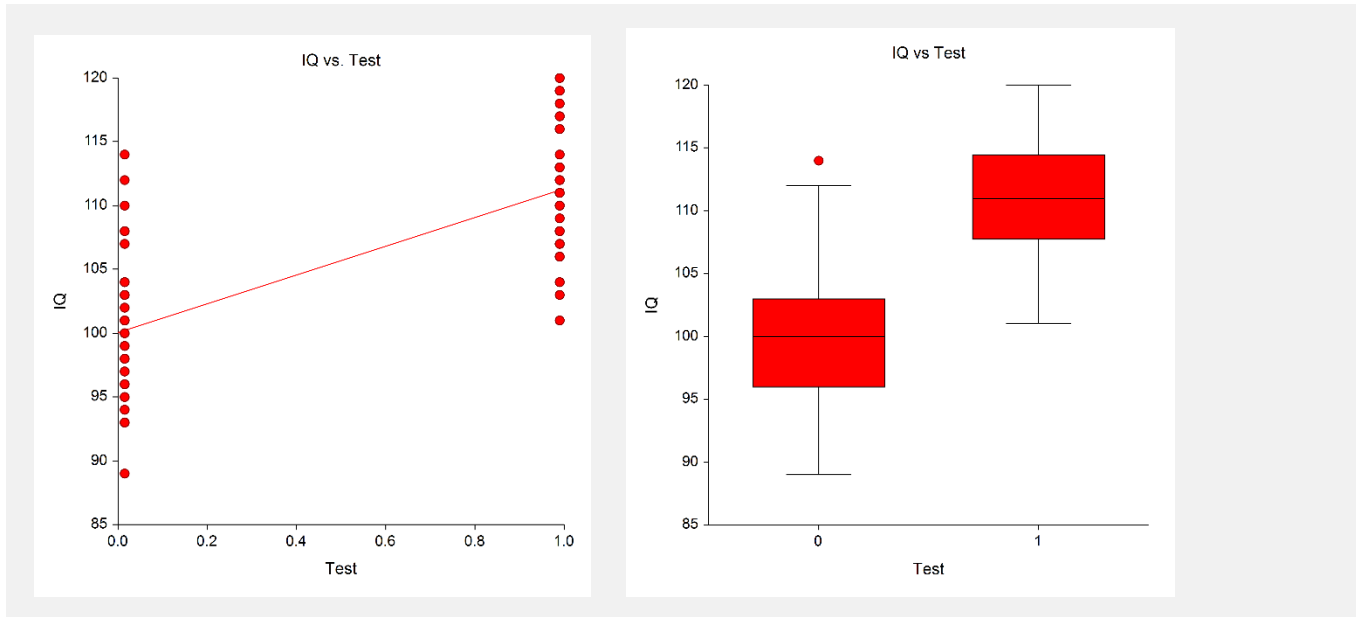
Point-Biserial and Biserial Correlations

Note that the point-biserial correlation demands that the variances are equal but is robust to mild non-normality. On the other hand, the biserial correlation is robust to unequal variances, but demands that the data are normal.

This report presents the usual descriptive statistics.

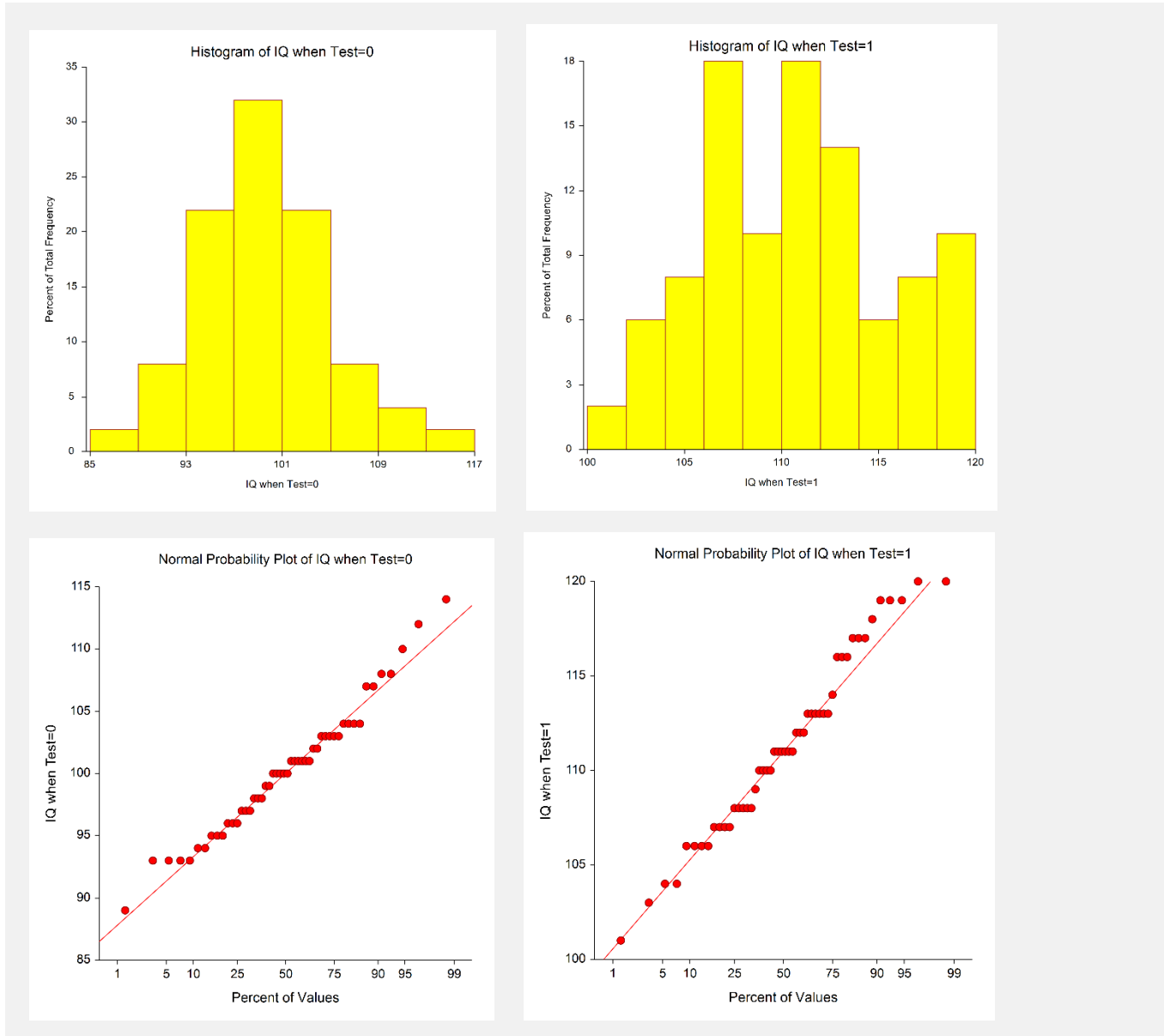
This report displays a brief summary of a linear regression of Y on X.

Plots to Evaluate Correlation



These plots let you investigate the relationship between the two variables more closely. The box plot is especially useful for comparing the variances of the two groups.

Plots to Evaluate Normality



The histograms and normal probability plots help you assess the viability of the assumption of normality within each group.