

How to calculate Point-Biserial Correlation in R?

Authored by
stats writer

December 24, 2025

RECOMMENDED CITATION

stats writer (2025). *How to calculate Point-Biserial Correlation in R?*. PSYCHOLOGICAL SCALES. Retrieved from <https://scales.arabpsychology.com/?p=108603>

The Point-biserial correlation is a crucial statistical measure used to quantify the strength and direction of the association between two distinct types of variables: a binary variable (dichotomous) and a continuous variable. This technique is fundamentally rooted in the widely used Pearson product-moment correlation coefficient, adapted specifically for situations involving one categorical variable with exactly two levels (e.g., success/failure, yes/no, male/female) and one variable measured on an interval or ratio scale (e.g., test scores, height, time).

To accurately calculate the Point-biserial correlation coefficient (r_{pb}) in the statistical programming environment R, analysts can utilize the versatile built-in function, `cor.test()`. This function is exceptionally powerful as it not only computes the correlation coefficient but also provides vital supplementary information, including the associated p-value--which determines statistical significance--and the confidence interval for the estimated population correlation. When employing `cor.test()` for this purpose, it is essential to supply two vectors: the first representing the binary (dichotomous) data, and the second containing the continuous measurements.

Understanding Point-Biserial Correlation

The Point-biserial correlation (r_{pb}) is a specialized measure employed specifically to determine the relationship between a strictly binary variable, conventionally denoted as x , and a continuous variable, denoted as y . Unlike other correlation methods suitable for two continuous variables (like standard Pearson's R) or two ordinal variables (like Spearman's ρ), the point-biserial method is engineered to handle the unique numerical properties of a dichotomous predictor, where the two categories are typically coded as 0 and 1. This correlation is particularly valuable in psychometrics, educational testing, and medical research where outcomes often involve binary classifications (e.g., survived/died, passed/failed) matched against continuous scores.

It is crucial to differentiate the Point-biserial correlation from the Biserial correlation. The Point-biserial coefficient assumes that the underlying distribution of the dichotomous variable is truly discrete and binary (e.g., gender, coin flip result). In contrast, the Biserial correlation is used when the dichotomous variable is assumed to be an artificial dichotomy derived from an underlying continuous normal distribution (e.g., high vs. low anxiety, where anxiety is inherently continuous but categorized for analysis). Since the Point-biserial method requires only the simplest binary coding (0 and 1) and is directly calculable using standard correlation formulas applied to the coded data, it is far more common and straightforward to implement in statistical software like R.

Similar in structure and interpretation to the Pearson correlation coefficient, the Point-biserial correlation coefficient yields a value that is strictly bounded between -1 and 1. This range allows for a standardized interpretation of the strength and direction of the linear relationship being assessed. Understanding these bounds is foundational to accurately interpreting the statistical output

generated by R, providing immediate insight into how the continuous variable tends to change when the binary variable switches its state from 0 to 1.

Interpreting the Point-Biserial Coefficient Values

The resulting correlation coefficient, r_{pb} , provides a concise summary of the relationship observed in the sample data. Its magnitude (closeness to 1 or -1) indicates the strength of the relationship, while its sign (positive or negative) indicates the direction. A strong correlation suggests that the mean score of the continuous variable differs significantly between the two groups defined by the binary variable. A weaker correlation implies that the means of the continuous variable are closer together across the two groups.

The interpretation of the specific values follows a standard scale applicable to most linear correlation coefficients:

-1 indicates a perfectly negative correlation between the two variables. This means that every observation coded as '1' on the binary variable has a low score on the continuous variable, and every observation coded as '0' has a high score on the continuous variable.

0 indicates absolutely no linear correlation between the two variables. In this scenario, the mean of the continuous variable for the group coded '1' is statistically identical to the mean for the group coded '0'.

1 indicates a perfectly positive correlation between the two variables. Conversely, this signifies that all observations coded as '1' on the binary variable possess high scores on the continuous variable, while those coded '0' possess low scores.

For practical analyses, values typically fall between these extremes. For instance, a coefficient of +0.3 suggests a weak to moderate positive relationship, indicating that the mean of the continuous variable is moderately higher for the '1' group compared to the '0' group. Conversely, a coefficient of -0.6 suggests a strong negative relationship, implying the '1' group has a substantially lower mean on the continuous variable.

Setting Up Data for Analysis in R

Before executing any statistical test in R, it is paramount to ensure that the data is correctly structured and typed. The Point-biserial test requires two vectors of equal length: one strictly numeric vector containing only two unique values (usually 0 and 1), representing the binary variable, and one numeric vector containing the continuous measurements. While R often treats numeric vectors automatically, ensuring the binary variable is coded as numeric (0, 1) rather than a factor is generally best practice when utilizing functions like `cor.test()`, which defaults to Pearson's R computation, which is mathematically equivalent to the Point-biserial method when one variable is dichotomous.

Consider a practical scenario where we are examining the relationship between passing a certification exam (x : 1=Pass, 0=Fail) and the number of hours spent studying (y : continuous). We need to input these variables into R as distinct vectors. The following setup demonstrates how we define our hypothetical binary variable, x , and our continuous variable, y , representing a small dataset of 11 observations.

Suppose we have gathered the following hypothetical data:

```
x <- c(0, 1, 1, 0, 0, 0, 1, 0, 1, 1, 0)
```

```
y <- c(12, 14, 17, 17, 11, 22, 23, 11, 19, 8, 12)
```

In this example, x is our dichotomous outcome (Pass/Fail) and y is the corresponding continuous score (Study Hours). We can immediately observe that observations where x equals 1 tend to have scores on y such as 14, 17, 23, 19, and 8, while observations where x equals 0 have scores such as 12, 17, 11, 22, 11, and 12. The goal of the Point-biserial test is to statistically determine if the difference in the means of y between the two groups defined by x is substantial enough to suggest a meaningful relationship.

Executing the Point-Biserial Test Using `cor.test()`

The calculation of the Point-biserial correlation coefficient in R is remarkably straightforward, leveraging the standard `cor.test()` function. Since the Point-biserial formula is mathematically identical to applying the Pearson correlation formula to data where one variable is coded dichotomously (0/1), R's default settings are perfectly suited for this analysis. We instruct the function to test the correlation between the two defined vectors, x and y , without needing to specify a special method argument.

By calling `cor.test(x, y)`, we initiate a robust statistical procedure that not only computes the primary correlation statistic but also performs a hypothesis test to assess whether the true population correlation is significantly different from zero. This is an essential step for moving beyond mere descriptive statistics to inferential analysis, allowing us to generalize the findings from our sample to the broader population from which the data was drawn.

The execution of the test in the R console looks like this:

```
#calculate point-biserial correlation
```

```
cor.test(x, y)
```

Pearson's product-moment correlation

data: x and y

$t = 0.67064$, $df = 9$, $p\text{-value} = 0.5193$

alternative hypothesis: true correlation is not equal to 0

95 percent confidence interval:

-0.4391885 0.7233704

sample estimates:

cor

0.2181635

The output clearly identifies the procedure used as "Pearson's product-moment correlation," confirming the mathematical equivalence of the methods. The subsequent lines provide all the necessary statistical metrics for a complete analysis, including the test statistic (t), the degrees of freedom (df), the observed p-value, the null and alternative hypotheses, the confidence interval, and finally, the calculated correlation coefficient itself.

Detailed Interpretation of R Output Metrics

Analyzing the output generated by `cor.test(x, y)` requires careful examination of several key metrics beyond just the correlation coefficient. This comprehensive report allows us to determine the empirical relationship and assess its reliability. From the output, we can derive the following critical observations:

The calculated Point-biserial correlation coefficient, labeled 'cor' under 'sample estimates', is **0.2181635**. This positive value indicates that there is a weak positive linear relationship between the variables: when the binary variable x shifts from 0 to 1, the continuous variable y tends to increase slightly.

The corresponding t -statistic is **0.67064** with 9 degrees of freedom. This t -statistic is used to calculate the p-value, testing the null hypothesis that the true population correlation (ρ) is zero. A t -statistic close to zero, as seen here, suggests that the observed sample correlation (0.218) is not far from the hypothesized population correlation of zero.

The interpretation of the correlation coefficient **0.218** implies that the group coded '1' (Pass) has a higher average score on the continuous variable (Study Hours) compared to the group coded '0' (Fail). While this indicates a tendency, the magnitude of the coefficient itself suggests that this difference in means is small to moderate. To properly assess the importance of this observed relationship, we must turn to the statistical significance test.

Evaluating Statistical Significance and Confidence Intervals

The most crucial step in inferential statistics is determining whether the observed sample correlation is statistically significant, meaning it is unlikely to have occurred by random chance if the true population correlation were zero. This is determined by the p-value associated with the test.

From the output provided:

The calculated p-value is **0.5193**.

To declare a result statistically significant, the p-value must typically be less than a predetermined significance level (α), conventionally set at 0.05. Since 0.5193 is substantially greater than 0.05, we fail to reject the null hypothesis. Therefore, based on this sample, the correlation of 0.218 is **not statistically significant**. This indicates that the observed positive relationship is likely due to sampling variability, and we cannot confidently conclude that a relationship exists between passing the exam and study hours in the broader population.

Furthermore, the output also provides a 95% confidence interval (C.I.) for the true population correlation coefficient (ρ). This interval provides a range of plausible values for the correlation in the population, based on our sample data. The interval is reported as:

95% C.I. = (-0.4391885, 0.7233704)

The fact that this confidence interval spans from a strong negative correlation (-0.439) to a strong positive correlation (0.723) reinforces the conclusion regarding non-significance. Crucially, because the interval contains the value zero, it suggests that zero correlation is a plausible value for the population correlation, providing further evidence that the calculated sample correlation (0.218) is not statistically reliable.

Practical Applications and Considerations

The Point-biserial correlation is an indispensable tool in fields requiring item analysis and group comparisons. In psychometrics, it is often used to assess item quality by correlating performance on a single test item (a binary variable: correct/incorrect) with the total score on the entire exam (a continuous variable). A high positive correlation here suggests that the item effectively discriminates between high and low overall performers.

While the `cor.test()` function in R handles the calculation efficiently, users should be aware of limitations. The Point-biserial correlation, like Pearson's R, measures only linear relationships. If the relationship between the binary and continuous variables is nonlinear, this coefficient may severely underestimate the true association. Furthermore, correlation does not imply causation; a

significant r_{pb} only indicates that the means of the continuous variable are statistically different between the two groups.

Finally, for users seeking more detailed information on parameters, hypotheses, and alternative testing methods available within R, consulting the official documentation is always recommended. For instance, the complete documentation for the `cor.test()` function explains how to specify alternative hypotheses (e.g., one-sided tests) or change the confidence level from the default 95%.

The following tutorials explain how to calculate other correlation coefficients in R:

ARABPSYCHOLOGY.COM