BISERIAL CORRELATION

Authored by
mohammad looti

November 11, 2025

BISERIAL CORRELATION

Primary Disciplinary Field(s): Statistics, Psychometrics, Quantitative Research

1. Core Definition

The **biserial correlation coefficient** (r_b) is a specialized measure in statistics designed to quantify the linear relationship between two variables when those variables possess specific measurement properties. Specifically, it is utilized when one variable (often denoted X) is measured on a continuous scale (interval or ratio data), and the second variable (often denoted Y) is a genuine or natural **dichotomy**. A natural dichotomy implies that the variable inherently exists in only two true categories, such as male or female biological sex, presence or absence of a specific condition, or success or failure in an outcome. This crucial distinction separates the biserial correlation from the point-biserial correlation, which is used when the dichotomy is treated as purely nominal or arbitrary (e.g., assigning 0 and 1 scores without assuming an underlying continuity).

The theoretical foundation of the biserial coefficient rests on the assumption that while Variable Y is observed as binary, it actually reflects an underlying latent trait that is continuous and normally distributed. The primary purpose of calculating r_b is to estimate the correlation that would exist between the two variables if the latent continuous distribution of Y could be perfectly observed and measured. Consequently, the magnitude of the biserial correlation tends to be higher than that of the point-biserial correlation coefficient calculated on the same data, because it mathematically corrects for the information loss incurred when a continuous variable is artificially truncated into two categories.

2. Etymology and Historical Development

The need for specialized correlation methods arose in the early 20th century, driven largely by the burgeoning fields of **psychometrics** and educational testing. Early statistical pioneers, particularly those involved in developing standardized tests, faced the challenge of relating continuous scores (e.g., total test performance) to binary item responses (e.g., whether a test-taker answered a specific question correctly or incorrectly). These item responses, while dichotomous in observation, were assumed to reflect an underlying, continuous ability or trait possessed by the test-taker.

The conceptual framework for the biserial correlation was formalized to address this gap, allowing researchers to estimate the true correlation that existed prior to the dichotomization process. This mathematical technique became integral to classical test theory (CTT) for performing item analysis. By providing a tool that could estimate the relationship between an observed binary variable and a hypothetical continuous variable, the biserial correlation helped validate early assessment tools and ensure that individual test items were adequately measuring the intended continuous construct. Its historical development is intrinsically linked to the broader statistical efforts, spearheaded by figures like Karl Pearson, to handle variables that deviate from the perfect continuous, bivariate normal distribution required for standard Pearson's r.

3. Key Characteristics and Assumptions

The utility and validity of the biserial correlation depend heavily on meeting several critical statistical assumptions related to the nature of the data and its distribution. The most stringent and crucial assumption is that the underlying construct responsible for the observed dichotomy must be continuously distributed and must follow a **normal distribution**. If, for instance, a researcher is correlating test scores (continuous) with whether a student passed or failed a competency test (dichotomous), the assumption is that the underlying competency skill is normally distributed in the population being studied.

Failure to meet the normality assumption can lead to a severely biased estimate of the correlation coefficient, potentially overestimating or underestimating the true strength of the relationship. Furthermore, like the Pearson product-moment correlation, the biserial coefficient assumes that the relationship between the continuous variable X and the latent continuous variable Y is fundamentally **linear**. If the true underlying association is curvilinear (e.g., quadratic or exponential), the biserial correlation will provide an inadequate summary of the relationship strength.

Assumption of Latent Normality: The variable that is observed as dichotomous must be derived from an unobserved, underlying variable that is continuous and normally distributed.
Homoscedasticity: The variability (variance) of the continuous variable X should be similar across both categories of the dichotomous variable Y.
Linearity: The relationship between the continuous variable X and the latent continuous variable Y must follow a straight line.
Genuine Dichotomy: The variable Y must represent a natural division (e.g., a biological distinction) rather than an arbitrary split of a measured continuous score.

4. Calculation and Formula

The calculation of the biserial correlation coefficient (r_b) is mathematically derived from the concept of the standard normal curve, requiring specific inputs that reflect the separation of the continuous variable's distribution based on the dichotomy. Unlike simpler correlation measures, the formula incorporates a normalization factor to account for the assumed underlying continuity.

The core formula for the biserial correlation is typically expressed as:
$$r_b = frac{(bar{X}_1 - bar{X}_0)}{s_x} cdot frac{pq}{y}$$

Where:

$\bar{X}_1$ and $\bar{X}_0$ represent the mean scores of the continuous variable X for the individuals in category 1 and category 0 of the dichotomy, respectively.
s_x is the standard deviation of the continuous variable X for the entire sample.
p and q are the proportions of cases in category 1 and category 0, respectively ($p+q=1$).
y is the **ordinate** (the height) of the standard normal distribution curve at the point that divides the area under the curve into the proportions p and q. This value, y, is derived from the standard normal probability density function corresponding to the z-score cut point.

The term $\frac{pq}{y}$ acts as the correction factor that adjusts the observed difference in means (normalized by standard deviation) to estimate the correlation as if the variable were continuous. The necessity of calculating y (the ordinate) is what makes the computation of r_b more cumbersome and conceptually demanding than the point-biserial correlation, which simply uses p and q without the normal curve adjustment.

5. Relationship to Other Correlation Measures

The biserial correlation is one of a family of correlation coefficients designed to handle non-continuous data types. Its comparison with the **point-biserial correlation** (r_{pb}) is particularly important, as both methods address the correlation between one continuous and one dichotomous variable. The fundamental difference lies in their assumptions regarding the nature of the dichotomy. The point-biserial correlation treats the two categories as discrete nominal values (e.g., 0 and 1) and is mathematically equivalent to calculating the standard **Pearson product-moment correlation coefficient** on the scored data.

In contrast, the biserial correlation explicitly assumes the dichotomy stems from an underlying continuous variable. Because the biserial coefficient adjusts the correlation estimate based on this assumption of latent normality, r_b will generally yield a correlation value that is larger in magnitude than r_{pb} when calculated on the identical dataset. This magnitude difference reflects the biserial coefficient's attempt to estimate the maximum possible linear relationship if the dichotomous variable were fully continuous. The decision between using r_b and r_{pb} is therefore not a matter of computational preference but a theoretical choice based on the researcher's hypothesis about the origin of the binary data. If the dichotomy is artificial (a split applied to an inherently continuous measure), r_b is appropriate; if the dichotomy is naturally nominal (e.g., marital status: married/unmarried), r_{pb} is preferred.

6. Significance and Impact

The biserial correlation has historically played a significant and foundational role in classical psychometrics, primarily in the domain of **item analysis** for psychological and educational tests.

When constructing standardized assessments, psychometricians must verify that individual items contribute meaningfully to the overall measurement of the underlying construct (e.g., intelligence or anxiety). The biserial correlation provides the ideal tool for this purpose: correlating the score on a single item (dichotomous: correct/incorrect) with the total score on the entire test (continuous).

A high positive biserial correlation indicates strong item discrimination; that is, students who score high on the overall test are highly likely to answer that specific item correctly, and students who score low on the overall test are likely to answer it incorrectly. This capacity to assess **discriminatory power** is vital for optimizing test validity and reliability. Items that exhibit low or negative biserial correlations are flagged as poor items, suggesting they are either confusing, flawed, or measuring a construct different from the rest of the test, and must therefore be removed or revised.

7. Debates and Criticisms

Despite its theoretical elegance, the biserial correlation is often utilized with caution in modern statistical practice, largely due to its reliance on strict distributional assumptions that are frequently difficult to verify or maintain in real-world data. The central criticism revolves around the mandatory assumption that the underlying latent trait must be perfectly normally distributed. If this assumption is violated, the calculated coefficient is prone to bias, making its interpretation potentially misleading.

Furthermore, the statistical stability of the biserial coefficient is highly sensitive to the proportions of the dichotomy. When the split is highly skewed (e.g., 90% in one category and 10% in the other), the estimate of the ordinate (y) becomes unstable, leading to an inflated standard error and reduced statistical precision. In such cases, the resulting r_b might be a poor estimate of the true relationship. Consequently, many contemporary researchers in quantitative psychology and social science often prefer alternative techniques that make fewer distributional assumptions, such as robust nonparametric methods (e.g., Spearman's rho) or advanced modeling approaches like logistic regression or modern Item Response Theory (IRT) models, which can handle binary data without forcing the assumption of latent normality.

8. Further Reading

Biserial Correlation (Wikipedia)
Biserial Correlation Definition and Examples (Statistics How To)
The Point Biserial Coefficient and the Biserial Coefficient as Measures of Item Discrimination (Educational and Psychological Measurement)