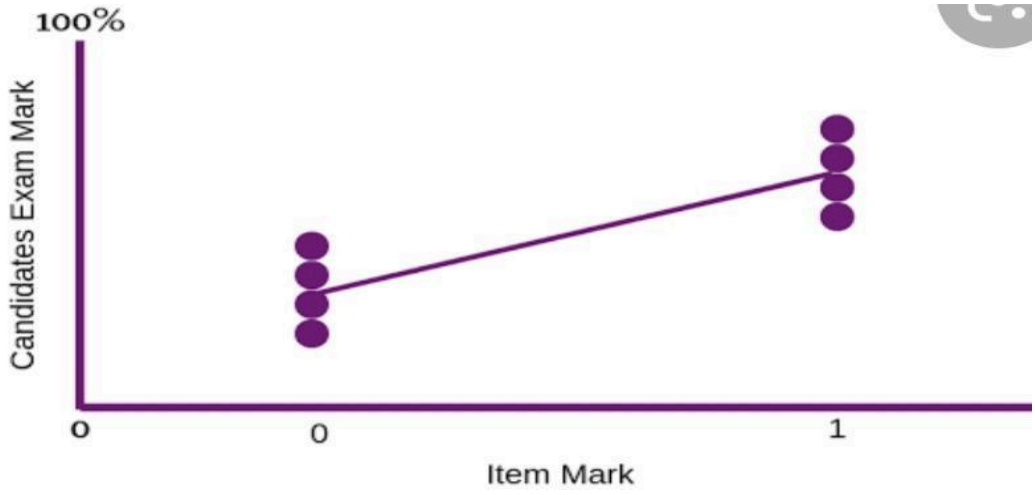
Point Biserial Correlation

A point- biserial correlation is used to measure the strength and direction of the association that exists between one continuous variable and one dichotomous variable. It is a special case of the Pearson's product-moment correlation, which is applied when you have two continuous variables, whereas in this case one of the variables is measured on a dichotomous scale.

For example, you could use a point-biserial correlation to determine whether there is an association between salaries, measured in US dollars, and gender (i.e., your continuous variable would be "salary" and your dichotomous variable would be "gender", which has two categories: "males" and "females"). Alternately, you could use a point-biserial correlation to determine whether there is an association between cholesterol concentration, measured in mmol/L, and smoking status (i.e., your continuous variable would be "cholesterol concentration", a marker of heart disease, and your dichotomous variable would be "smoking status", which has two categories: "smoker" and "non-smoker").



Assumptions of Point Biserial Correlation

Every statistical method has assumptions. Assumptions mean that your data must satisfy certain properties in order for statistical method results to be accurate.

The assumptions for Point-Biserial correlation include:

- Continuous and Binary
- Normally Distributed
- No Outliers
- Equal Variances

Let's dive in to each one of these separately.

- **Continuous and Binary**

For this test, you should have one continuous and one binary variable. Continuous means that the variable can take on any reasonable value. Some good examples of continuous variables include age, weight, height, test scores, survey scores, yearly salary, etc.

Binary means that your variable is a category with only two possible values. Some good examples of binary variables include smoker(yes/no), sex(male/female) or any True/False or 0/1 variable.

- **Normally Distributed**

The variable that you care about must be spread out in a normal way. In statistics, this is called being normally distributed (aka it must look like a bell curve when you graph the data). Only use Point-Biserial Correlation on your data if the variable you care about is normally distributed

- **No Outliers**

The variables that you care about must not contain outliers. Point-Biserial correlation is sensitive to outliers, or data points that have unusually large or small values. You can tell if your variables have outliers by plotting them and observing if any points are far from all other points.

- **Equal Variances**

One of the assumptions of Point-Biserial correlation is that there is similar spread between the two groups of the binary variable. You can check for this assumption by plotting your continuous variable in each of your two groups and visually identifying if the spread of the data is similar.

When we use Point Biserial Correlatio

1. You should use Point-Biserial Correlation in the following scenario:

2. You want to know the relationship between two variables

3. Your variables of interest include one continuous and one binary variable

4. You have only two variables

Let's clarify these to help you know when to use Point-Biserial Correlation

Relationship; you are looking for a statistical test to look at how two variables are related. Other types of analyses include testing for a difference between two variables or predicting one variable using another variable (prediction).

One Continuous and One Binary

For this test, you should have one continuous and one binary variable. Continuous means that the variable can take on any reasonable value. Some good examples of continuous variables include age, weight, height, test scores, survey scores, yearly salary, etc.

Binary means that your variable is a category with only two possible values. Some good examples of binary variables include smoker(yes/no), sex(male/female) or any True/False or 0/1 variable.

If you have two continuous variables, you should use Pearson Correlation. And if you have at least one ordinal variable, you should use Spearman's Rho or Kendall's Tau instead.

Point Biserial Correlation Example

Two Variables

- Variable 1: Height.
- Variable 2: Gender.

In this example, we are interested in the relationship between height and gender. To begin, we collect these data from a group of people.

Before running Point-Biserial Correlation, we check that our variables meet the assumptions of the method. After confirming that our continuous variable is normally distributed, has no outliers, and has equal variances in each gender, we move forward with the analysis.

The analysis will result in a correlation coefficient (called “r”) and a p-value. R values range from -1 to 1. A negative value of r indicates that the variables are inversely related, or when one variable increases, the other decreases. On the other hand, positive values indicate that when one variable increases, so does the other. In this example, whether r is positive or negative depends on which gender you represent with a value of 0 and which you represent with a value of 1.

Point Biserial correlation formula