

How can the correlation between continuous and categorical variables be calculated?

Authored by
stats writer

June 26, 2024

RECOMMENDED CITATION

stats writer (2024). *How can the correlation between continuous and categorical variables be calculated?*. PSYCHOLOGICAL SCALES. Retrieved from <https://scales.arabpsychology.com/?p=154413>

The correlation between continuous and categorical variables measures the strength of the relationship between two variables with different types of data. It is typically calculated using a statistical method called the Pearson correlation coefficient, which quantifies the linear association between two continuous variables. This method can also be used to determine the correlation between a continuous and a binary categorical variable. However, for categorical variables with more than two categories, alternative statistical measures such as the Spearman correlation coefficient or the Cramer's V statistic may be more appropriate. These methods take into account the categorical nature of the variable and provide a measure of the strength and direction of the relationship. Overall, the calculation of correlation between continuous and categorical variables is an important tool in statistical analysis, providing valuable insights into the relationship between different types of data.

Calculate Correlation Between Continuous & Categorical Variables

When we would like to calculate the correlation between two continuous variables, we typically use the .

However, when we would like to calculate the correlation between a continuous variable and a , we can use something known as point biserial correlation.

Point biserial correlation is used to calculate the correlation between a binary categorical variable (a variable that can only take on two values) and a continuous variable and has the following properties:

Point biserial correlation can range between -1 and 1. For each group created by the binary variable, it is

assumed that the continuous variable is normally distributed with equal variances. For each group created by the binary variable, it is assumed that there are no extreme outliers.

The following example shows how to calculate a point biserial correlation in practice.

Example: Calculating a Point Biserial Correlation

Suppose a college professor would like to determine if there is a correlation between gender and score on particular aptitude exam.

He collects the following data on 12 males and 12 females in his class:

Gender	Score
Female	77
Female	78
Female	79
Female	79
Female	82
Female	84
Female	85
Female	88
Female	89
Female	91
Female	91
Female	94
Male	84
Male	84
Male	84
Male	85
Male	85
Male	86
Male	86
Male	86
Male	89
Male	91
Male	94
Male	98

Since gender is a categorical variable and score is a continuous variable, it makes sense to calculate a point-biserial correlation between the two variables.

The professor can use any statistical software (including Excel, R, Python, SPSS, Stata) to calculate the point-biserial correlation between the two variables.

The following code shows how to calculate the point-biserial correlation in R, using the value 0 to represent females and 1 to represent males for the gender variable:

```
#define values for gender
```

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

```
#define values for score
```

```
score <- c(77, 78, 79, 79, 82, 84, 85, 88, 89, 91, 91, 94,  
84, 84, 84, 85, 85, 86, 86, 86, 89, 91, 94, 98)
```

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

```
cor.test(gender, score)
```

Pearson's product-moment correlation

data: gender and score

t = 1.3739, df = 22, p-value = 0.1833

alternative hypothesis: true correlation is not equal to 0

95 percent confidence interval:

-0.1379386 0.6147832

sample estimates:

cor

0.2810996

From the output we can see that the point biserial correlation coefficient is 0.281 and the corresponding p-value is 0.1833.

Since the correlation coefficient is positive, it tells us that there is a positive correlation between gender and score.

Since we coded the males as 1 and females as 0, this indicates that scores tend to be higher for males (i.e. scores tend to increase as gender "increases" from 0 to 1)

However, since the p-value is not less than .05, this correlation coefficient is not statistically significant.

The following tutorials explain how to calculate point biserial correlation using different statistical software: