

How do you calculate the correlation between categorical variables?

Authored by
stats writer

May 12, 2024

RECOMMENDED CITATION

stats writer (2024). *How do you calculate the correlation between categorical variables?*. PSYCHOLOGICAL SCALES. Retrieved from <https://scales.arabpsychology.com/?p=143875>

Calculating the correlation between categorical variables involves using statistical measures to determine the strength and direction of the relationship between two categorical variables. This is typically done by creating a contingency table and calculating the chi-square statistic or using other measures such as Cramer's V or Phi coefficient. These calculations can help to determine if there is a significant association between the two variables and the nature of that association. The resulting correlation coefficient can range from 0 to 1, with higher values indicating a stronger relationship. This method can be useful in understanding the relationship between categorical variables and can aid in making data-driven decisions.

Calculate Correlation Between Categorical Variables

Often we use the to calculate the correlation between continuous numerical variables.

However, we must use a different metric to calculate the correlation between categorical variables - that is, variables that take on names or labels such as:

Marital status (single, married, divorced)Smoking status (smoker, non-smoker)Eye color (blue, brown, green)

There are three metrics that are commonly used to calculate the correlation between categorical variables:

1. Tetrachoric Correlation: Used to calculate the correlation between binary categorical variables.

2. Polychoric Correlation: Used to calculate the correlation between ordinal categorical variables.

3. Cramer's V: Used to calculate the correlation between nominal categorical variables.

The following sections provide an example of how to calculate each of these three metrics.

Metric 1: Tetrachoric Correlation

Tetrachoric correlation is used to calculate the correlation between binary categorical variables. Recall that binary variables are variables that can only take on one of two possible values.

The value for tetrachoric correlation ranges from -1 to 1 where -1 indicates a strong negative correlation, 0 indicates no correlation, and 1 indicates a strong positive correlation.

For example, suppose we want to know whether or not gender is associated with political party preference so we take a simple random sample of 100 voters and survey them on their political party preference.

The following table shows the results of the survey:

		Political Party	
		Dem	Rep
Gender	Male	19	30
	Female	12	39

We would use tetrachoric correlation in this scenario because each categorical variable is binary - that is, each variable can only take on two possible values.

We can use the following code in R to calculate the tetrachoric correlation between the two variables:

```
library(psych)
```

```
#create 2x2 table
```

```
data = matrix(c(19, 12, 30, 39), nrow=2)
```

```
#view table
```

```
data
```

```
#calculate tetrachoric correlation
```

```
tetrachoric(data)
```

```
tetrachoric correlation
```

```
0.27
```

Metric 2: Polychoric Correlation

Polychoric correlation is used to calculate the correlation between ordinal categorical variables. Recall that are variables whose possible values have a natural order.

The value for polychoric correlation ranges from -1 to 1 where -1 indicates a strong negative correlation, 0 indicates no correlation, and 1 indicates a strong positive correlation.

For example, suppose want to know whether or not two different movie ratings agencies have a high correlation between their movie ratings.

We ask each agency to rate 20 different movies on a scale of 1 to 3 with 1 indicating "bad", 2 indicating "mediocre", and 3 indicating "good."

The following table shows the results:

Movie	Agency 1 Rating	Agency 2 Rating
#1	1	1
#2	1	1
#3	2	2
#4	2	1
#5	3	3
#6	2	3
#7	2	3
#8	3	2
#9	2	2
#10	3	3
#11	3	3
#12	2	3
#13	1	2
#14	2	2
#15	2	2
#16	1	1
#17	1	2
#18	1	1
#19	2	3
#20	2	3

We can use the following code in R to calculate the polychoric correlation between the ratings of the two agencies:

```
library(polycor)
```

```
#define movie ratings
```

```
x <- c(1, 1, 2, 2, 3, 2, 2, 3, 2, 3, 3, 2, 1, 2, 2, 1, 1, 1, 2, 2)
```

```
y <- c(1, 1, 2, 1, 3, 3, 3, 2, 2, 3, 3, 3, 2, 2, 2, 1, 2, 1, 3, 3)
```

```
#calculate polychoric correlation between ratings  
polychor(x, y)
```

0.7828328

The polychoric correlation turns out to be 0.78. This value is quite high, which indicates that there is a strong positive association between the ratings from each agency.

Metric 3: Cramer's V

Cramer's V is used to calculate the correlation between nominal categorical variables. Recall that nominal variables are ones that take on category labels but have no natural ordering.

The value for Cramer's V ranges from 0 to 1, with 0 indicating no association between the variables and 1 indicating a strong association between the variables.

For example, suppose we want to know if there is a correlation between eye color and gender so we survey 50 individuals and obtain the following results:

		Eye Color		
		Blue	Green	Brown
Gender	Male	6	8	12
	Female	9	5	10

We can use the following code in R to calculate Cramer's V for these two variables:

```
library(rcompanion)
```

```
#create table
```

```
data = matrix(c(6, 9, 8, 5, 12, 10), nrow=2)
```

```
#view table
```

```
data
```

```
6 8 12
```

```
9 5 10
```

```
#calculate Cramer's V
```

```
cramerV(data)
```

```
Cramer V
```

```
0.1671
```

Cramer's V turns out to be 0.1671. This value is quite

low, which indicates that there is a weak association between gender and eye color.

ARABPSYCHOLOGY.COM