Print

Associations between Categorical Variables

By examining distributions of categorical variables, you can determine the frequencies of data values and possible associations among variables. An association exists between two categorical variables if the distribution of one variable changes when the value of the other variable changes. You can also think of an association as the probability that one variable depends on the probability of the other. If there's no association, the distribution of the first variable is the same, regardless of the level of the other variable.

To look for associations, you examine the frequencies of values across the combinations of variables. For example, suppose you want to determine whether your mood is affected by the weather. The categorical response variable is Mood, and its values are either happy or grumpy. The categorical predictor variable is Weather, and its values are either sunny or cloudy. This table shows row frequency percentages for combinations of values of the two variables. On sunny days, you're happy 72% of the time and grumpy 28% of the time.

Now look at the frequency percentages for your mood on cloudy days. The row percentages are the same in each column, indicating that there's no change in your mood based on the weather. So, there's no association between these two variables.

What if these were your results? In this table, the row percentages are different in each column. On sunny days, you're happy 82% of the time and grumpy 18% of the time. On cloudy days, you're happy 60 % of the time and grumpy 40% of the time. Your mood changes based on the weather. It appears you're more likely to be happy if the weather is nicer, indicating a possible association between mood and weather.

Consider the Ames data. Which variables are associated with the response variable Bonus? Are homes with two fireplaces associated with bonus-eligible homes? Are homes with a basement area greater than 1000 square feet associated with bonus-eligible homes? In the next demonstration, we'll look for associations between home features and the response variable Bonus. Later, we'll test for statistically significant associations using hypothesis tests.

Close