The Pearson Chi-Square Test

Let's start with our null hypothesis, that there's no association between the variables Lot_Shape_2 and Bonus, meaning that the probability of a home sale being bonus eligible is the same regardless of lot shape. The alternative hypothesis is that there is an association between Lot_Shape_2 and Bonus, meaning the probability of a home sale being bonus eligible is not the same for irregular and regular lot shapes.

To formally test the association for statistical significance, we'll use the Pearson chi-square test, often referred to as simply the chi-square test. It measures the difference between the observed cell counts and the cell counts that are expected if there's no association between the variables, and the null hypothesis is in fact true.

The chi-square test calculates the expected counts for each cell by multiplying the row total (R) by the column total (C), and then dividing the result by the total sample size (T). The larger the difference between the observed and expected cell counts, the more evidence of a statistically significant association between the variables. A significant chi-square statistic provides evidence to reject the null hypothesis and conclude that an association exists. To calculate the chi-square test statistic, you square the difference between the observed and expected counts for each cell, and divide by the expected cell count to get the cell chi-square values.

$$\sum \frac{(\text{observed} - \text{expected})^2}{\text{expected}}$$

The overall chi-square, or the test statistic, is calculated by then adding the cell chi-square values over all cells.

$$\sum\sum \left(\frac{(\text{observed}_{rc} - \text{expected}_{rc})^2}{\text{expected}_{rc}} \right)$$

The degrees of freedom associated with the test statistic is the number of rows minus one, times the number of columns minus one. In this case, with a two-by-two table, we have one degree of freedom. Keep in mind that neither the chi-square statistic nor its p-value tells you the magnitude of an association. They indicate only how confident you can be that an association exists.

Chi-square statistics and their p-values depend on and reflect the sample size. A larger sample size yields a larger chi-square statistic and a smaller corresponding p-value, even though the association might not be strong.

For example, if you double the size of your sample by duplicating each observation, you double the value of the chi-square statistic, even though the strength of the association does not change.

Cramer's V statistic is one measure of the strength of an association between two categorical variables, and its value is derived from the chi-square statistic. For two-by-two tables, Cramer's V is in the range of -1 to 1, and for larger tables, its in the range of 0 to 1.Values farther away from 0 indicate a relatively strong association between the variables. The closer Cramer's V is to 0, the weaker the association is between the two variables.

Like other measures of strength of association, the Cramer's V statistic is not affected by sample size.

Close