Statistical Thinking for the 21st Century

Chapter 12

Modeling categorical relationships

12.3 Contingency tables and the chi-square test of independence

Another way that we often use the chi-square test is to ask whether two discrete measures are related to one another. When we use chi-square test in this way, we are conducting a chi-square test of independence.

As an example, let's take the question of whether a Black driver, compared to a white driver, is more likely to be searched when they are pulled over by a police officer.

The Stanford Open Policing Project (https://openpolicing.stanford.edu) has studied this question and provided data that we can use to analyze the question. We will use the data from the state of Connecticut because the numbers are easy enough for us to calculate ourselves.

The standard way to represent data for a chi-square test of independence is through a contingency table, which presents the frequency of observations that fall into each possible combination -- each contingency.

NOTE: Wikipedia (2020) tells us that "in statistics, a contingency table (also known as a cross tabulation or crosstab) is a type of table in a matrix format that displays the frequency distribution of more than one variable. Contingency tables are heavily used in survey research, business intelligence, engineering and scientific research. They provide a basic picture of the interrelation between two variables. The term contingency table was first used by Karl Pearson in ... 1904."

Table 12.2: Contingency table for police search data

Driver	Searched	Not Searched		Total		Driver	Searched	Not Searched		Total
Black	1,219	36,244		37,463		Black	0.004	0.130		0.134
White	3,108	239,241		242,349		White	0.011	0.855		0.866
Total	4,327	275,485		279,812		Total	0.115	0.985		1.000

Table 12.2 shows the contingency table for the police search data.

When looking at a contingency table, it's often easier to read the frequencies as proportions (Relative Frequency) rather than as integers (Absolute Frequency), because Relative Frequencies are easier to compare visually. Therefore, best practice in a contingency table is to include both Absolute and Relative Frequencies.

We can use our contingency table to calculate a Pearson chi-square test of independence, which will test whether the observed frequencies are independent.

As you remember from the chapter on probability, if X and Y are independent, then:

$$P(X \cap Y) = P(X) * P(Y)$$

That is, the joint probability under the null hypothesis of independence is simply the product of the *marginal* probabilities of each individual variable. We can compute those marginal probabilities, and then multiply them together to get the expected proportions under independence.

For the police search data, the chi-square value we calculate is 828.3. To compute a *p*-value, we compare our chi-square value to the null chi-squared distribution.

To calculate the degrees of freedom for a chi-square test of independence, we use the formula

df = (the number of Rows minus 1) * *(the number of Columns* minus 1)

In our police search data, we have two rows; therefore, the number of Rows minus 1 is 1.
In our police search data, we have two columns, therefore the number of Columns minus 1 is 1.
1 x 1 is 1, so the degrees of freedom for this chi-square test of independence is 1.

The *p*-value for the chi-squared statistic we calculated for the police search data is about as close to zero as possible. Therefore, we can reject the null hypothesis that a driver being Black or white is independent of whether the driver is searched when they are pulled over by a police officer.