22.3: Contingency Tables and the Two-way Test

Another way that we often use the chi-squared test is to ask whether two categorical variables are related to one another. As a more realistic example, let's take the question of whether a black driver is more likely to be searched when they are pulled over by a police officer, compared to a white driver. The Stanford Open Policing Project (https://openpolicing.stanford.edu/) has studied this, and provides data that we can use to analyze the question. We will use the data from the State of Connecticut since they are fairly small. These data were first cleaned up to remove all unnecessary data.

The standard way to represent data from a categorical analysis is through a *contingency table*, which presents the number or proportion of observations falling into each possible combination of values for each of the variables. The table below shows the contingency table for the police search data. It can also be useful to look at the contingency table using proportions rather than raw numbers, since they are easier to compare visually, so we include both absolute and relative numbers here.

Table 22.2: Contingency table for police search data

searched	Black	White	Black (relative)	White (relative)
FALSE	36244	239241	0.13	0.86
TRUE	1219	3108	0.00	0.01

The Pearson chi-squared test allows us to test whether observed frequencies are different from expected frequencies, so we need to determine what frequencies we would expect in each cell if searches and race were unrelated – which we can define as being *independent*. Remember from the chapter on probability that 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. The marginal probabilities are simply the probabilities of each event occuring regardless of other events. We can compute those marginal probabilities, and then multiply them together to get the expected proportions under independence.

	Black	White	
Not searched	P(NS)*P(B)	P(NS)*P(W)	P(NS)
Searched	P(S)*P(B)	P(S)*P(W)	P(S)
	P(B)	P(W)	

Table 22.3: Summary of the 2-way contingency table for police search data

searched	driver_race	n	expected	stdSqDiff
FALSE	Black	36244	36884	11.1
TRUE	Black	1219	579	706.3
FALSE	White	239241	238601	1.7
TRUE	White	3108	3748	109.2

We then compute the chi-squared statistic, which comes out to 828.3. To compute a p-value, we need to compare it to the null chi-squared distribution in order to determine how extreme our chi-squared value is compared to our expectation under the null hypothesis. The degrees of freedom for this distribution are $df = (nRows - 1) * (nColumns - 1)$ - thus, for a 2X2 table like the one here, $df = (2 - 1) * (2 - 1) =$. The intuition here is that computing the expected frequencies requires us to use three values: the total number of observations and the marginal probability for each of the two variables. Thus, once those values are computed, there is only one number that is free to vary, and thus there is one degree of freedom. Given this, we can compute the p-value for the chi-squared statistic, which is about as close to zero as one can get: $3.79e^{-182}$. This shows that the observed data would be highly unlikely if there was truly no relationship between race and police searches, and thus we should reject the null hypothesis of independence.

We can also perform this test easily using the `chisq.test()` function in R:

```
##
##  Pearson's Chi-squared test
##
## data:  summaryDf2wayTable
## X-squared = 828, df = 1, p-value <2e-16
```