Chi-Square Test of Independence

James M. Murray, Ph.D.
University of Wisconsin - La Crosse

Updated: September 24, 2017

PDF file location: http://www.murraylax.org/rtutorials/chisqindep.pdf

HTML file location: http://www.murraylax.org/rtutorials/chisqindep.html

---

Note on required packages: The following code required the package `gmodels` to perform statistics and data visualizations relating to the Chi-Square test of independence. If you have not already done so, install and load the package with the following code:

```
install.packages("gmodels")# This only needs to be executed once for your machine
```

```
library("gmodels") # This needs to be executed every time you load R
```

The **Chi-Square Test of Independence** tests for a relationship between two *categorical* variables.

1. Download the Data

The data set `empdata.RData` includes a small subset of data from on employment and race identification from 2014 for more than 21,000 people. The data comes from the American Community Survey, an ongoing survey by the U.S. Census Bureau that measures variables relating to population, housing, workforce, and demographics.

The code below downloads the data and opens it, creating a data frame object, `empdata`.

```
load(url("http://murraylax.org/datasets/empdata.RData"))
```

The data set includes three variables. The first, `Race`, is a string for the racial identify of the individual. The second, `Status`, is a string for the employment status of the individual, including whether they are employed, unemployed, or not in the labor force (not employed, but not looking for work). The final variable, `Class`, is the classification for employment. For employed people, this includes working for wages or self employed. For individuals not working, `Class` has a missing value.

In the example below, we will focus on employed people, and examine whether there is a relationship between **Race** and **Class** (self-employed vs wage income).

2. Contingency Table

For two categorical variables measures from the same sampling units (the sampling unit is one individual from the survey in our example), a contingency table breaks both variables into its categories and reports how many fall into each group and subgroup.

The code below reports a contingency table for **Race** and **Class**.

```
CrossTable(empdata$Race, empdata$Class)
```

```
##
##
```

```
##    Cell Contents
## |-------------------------|
## |                       N |
## | Chi-square contribution |
## |           N / Row Total |
## |           N / Col Total |
## |         N / Table Total |
## |-------------------------|
##
##
## Total Observations in Table:  9662
##
##
##                            | empdata$Class
##               empdata$Race |   Self Employed | Works for Wages |       Row Total |
## --------------------------|-----------------|-----------------|-----------------|
##                      Asian |              49 |             516 |             565 |
##                            |           2.441 |           0.297 |                 |
##                            |           0.087 |           0.913 |           0.058 |
##                            |           0.047 |           0.060 |                 |
##                            |           0.005 |           0.053 |                 |
## --------------------------|-----------------|-----------------|-----------------|
##                      Black |              47 |             811 |             858 |
##                            |          22.734 |           2.763 |                 |
##                            |           0.055 |           0.945 |           0.089 |
##                            |           0.045 |           0.094 |                 |
##                            |           0.005 |           0.084 |                 |
## --------------------------|-----------------|-----------------|-----------------|
## Native American / Alaskan |               6 |              86 |              92 |
##                            |           1.580 |           0.192 |                 |
##                            |           0.065 |           0.935 |           0.010 |
##                            |           0.006 |           0.010 |                 |
##                            |           0.001 |           0.009 |                 |
## --------------------------|-----------------|-----------------|-----------------|
##                      Other |              44 |             556 |             600 |
##                            |           6.794 |           0.826 |                 |
##                            |           0.073 |           0.927 |           0.062 |
##                            |           0.042 |           0.065 |                 |
##                            |           0.005 |           0.058 |                 |
## --------------------------|-----------------|-----------------|-----------------|
##                      White |             901 |            6646 |            7547 |
##                            |           8.462 |           1.028 |                 |
##                            |           0.119 |           0.881 |           0.781 |
##                            |           0.861 |           0.771 |                 |
##                            |           0.093 |           0.688 |                 |
## --------------------------|-----------------|-----------------|-----------------|
##               Column Total |            1047 |            8615 |            9662 |
##                            |           0.108 |           0.892 |                 |
## --------------------------|-----------------|-----------------|-----------------|
##
##
```

Let us look at the upper-left cell to understand the table contents. In this cell five numbers describing the frequency of employed people who identified themselves as **Asian** and **Self Employed**.

1. The first number indicates there were 49 people who identify themselves as self-employed and Asian.
2. The second number is an intermediate statistics used in the calculation of the Chi-square test of independence.
3. The third number is the proportion (between 0 and 1) of all those who identify as Asian (*row label*) who are self-employed. Therefore, 8.7% of employed Asian people are self employed; the remaining 91.3% work for wages.
4. The fourth number is the proportion (between 0 and 1) of all those who are self-employed (*column label*) who identify themselves as Asian. Of all self-employed people, 4.7% are Asian. The remaining 95.3% identify with other races.
5. The fifth number is the proportion (between 0 and 1) of all employed people in the sample who identify themselves as both Asian and self-employed. The percentage of all employed people who are Asian and self-employed is 0.5%.

The final column reports the row totals. Since the row labels are the race categories, the top cell in the final column reports that among all employed people in the sample, 565 of them, or 5.8% of them, are Asian.

The final row reports the column totals. Since the column labels are the employment classification categories, the first cell in the final column reports that among all employed people in the sample, 1047 of them, or 10.8% of them, are self-employed.

If race and employment classification are independent, than the proportions that we see in the column totals should be similar to the proportions reported in the third number in each cell.

For example, the lower-left cell indicates that 10.8% of all people in the sample are self-employed. If race and classification are independent, we should see similar percentages for each race. We see in the top-left cell that only 8.7% of Asian people are self employed. The next cell down shows that even fewer black people, 5.5% are self employed. In the `White / Self Employed` cell, we see a larger percentage, 11.9%, of white people are self employed.

Do these differences in self-employment rates by race indicate that there is a relationship between race and employment classification, or are these differences simply due to random sampling error? To answer that question, we conduct the **Chi-Square Test of Independence**.

2. Chi-Square Hypothesis Test for Independence

The null hypothesis for the Chi-square test of independence is that there is no relationship between the two categorical variables. The alternative hypothesis is that there is a relationship between the two categorical variables. For the example above, the hypotheses are given by the following:

Null: There is **no relationship** between race and employment classification

Alternative: there **is a relationship** between race and employment classification

The following code conducts the test and computes the p-value.

```
chisq.test(empdata$Race, empdata$Class)
```

```
##
##  Pearson's Chi-squared test
##
## data:  empdata$Race and empdata$Class
## X-squared = 47.117, df = 4, p-value = 1.442e-09
```

With a p-value equal to `1.44e-09` we reject the null hypothesis. We find sufficient statistical evidence that there is a relationship between race and employment classification.

3. Problems

1. Use the `CrossTable()` function to analyze the relationship between race and employment status. What percentage of each population Asian, Black, and White people are not in the labor force? (A person is not in the labor force when they are neither employed nor looking for employment. Common examples include college students, retired people, stay-at-home parents, etc.)
2. What percentage of all people in the sample are employed?
3. Is there evidence that there is a relationship between employment status and race?