NAME: Juan Villegas

Lab 4: Chi-Squared Significance Test

In this lab, you will conduct a chi-square analysis from data that has already been organized into tables. (Note: “chi” is pronounced “ki,” as in “Kyle.”). In Lab 4, you will begin preparing the data for next week’s lab. The data are from the Monitoring the Future Study.

1. The Federal Reserve has called remote deposit capture “the most important development the U.S. banking industry has seen in years.” This service allows users to scan checks and to transmit the scanned images to a bank for posting. In its annual survey of community banks, the American Bankers Association asked banks whether or not they offered this service. Here are the results classified by the asset size (in millions of dollars) of the bank:

|  |  |  |  |
| --- | --- | --- | --- |
|  |  | Offer | RDC |
|  |  | Yes | No |
| Asset | Under $100 | 57 | 64 |
| Size | $101 - $200 | 227 | 163 |
|  | $201 or more | 271 | 93 |

Table 1. Bank data

a. Set up the hypotheses to test whether there is an association between Asset size and Offer RDC.

 There is no association b/w Asset Size & Offer RDC

 There is an association b/w Asset Size & Offer RDC

b. Create a table of expected cell counts.

* Show how you would do at least two of the expected cell counts using only a calculator for calculations. (You can use R to calculate the remaining expected cell counts after completing the bullets below). After showing two of the calculations, complete the table below.

|  |  |  |  |
| --- | --- | --- | --- |
|  |  | Offer | RDC |
|  |  | Yes | No |
| Asset | Under $100 | (121)(555)/875 | (121)(320)/875 |
| Size | $101 - $200 | (390)(555)/875 | (390)(320)/875 |
|  | $201 or more | (364)(555)/875 | (364)(320)/875 |

Expected cell counts.

* Next, enter the observed data (from Table 1) into RStudio. Text

  Description automatically generated
* Use R to determine the value of the  statistic, the df, and the *p*-value. (**Do not apply Yates correction! Use correct = FALSE**)

Text

Description automatically generated

c. If the *p*-value is below 0.05, **reject the null hypothesis** in favor of the alternative. (If you reject the null hypothesis, you can say that the results are **significant** and conclude that there is a relationship between Asset size and Offer RDC.).

I have to reject the Null Hypothesis. The results are significant because the p-value I got is 7.033e-09 which is a lot less than .05. There is a relationship between Asset Size & Offer RDC.

d. To examine the nature of any association between the two variables, Asset size (**explanatory variable**) and Offer RDC (**response variable**), calculate either row or column percentages, whichever is more appropriate to the situation under study. What do your percentages reveal about this situation?

Text

Description automatically generated

The **Row percentages** is more appropriate for this study. These percentages reveal that the greater ones Assets Size is, the more likely they are to be offered an RDC.

e. Make a bar chart to display the conditional distributions of Offer RDC for each level of Asset size.

Chart, histogram

Description automatically generated

2. A random sample of registered voters was asked about their educational background and whether or not they voted in the November 2012 elections. (Voting in November 2020 is fast approaching!) Table 2 contains the results of the survey.

|  |  |  |  |
| --- | --- | --- | --- |
|  |  | Voted | Nov. 2012 |
|  |  | Yes | No |
| Highest | Not High School Graduate | 57 | 64 |
| Educational | High School Graduate/No College | 227 | 163 |
| Attainment | Some College or Associate's Degree | 271 | 93 |
|  | Bachelor's Degree or Higher | 303 | 51 |

Table 2. Survey results on voting and highest educational attainment.

a. In this situation, which is the explanatory variable, and which is the response variable? Justify your answer.

Highest Educational Attainment is the explanatory variable because this is the variable we are constantly changing to see if people come out to vote. This is the variable causing a change in the response variable. The Response Variable is “Voted in Nov. 2012” because we are measuring the response of the changes in Highest Educational Attainment.

b. Set up the hypotheses for testing whether educational attainment and voting in the 2012 presidential election are independent.

Ho = There is no association b/w voting in Nov. 2012 and highest educational attainment

Ha = There is an association b/w voting in Nov. 2012 and highest educational attainment

c. Use RStudio to calculate the chi-square test statistic, state the degrees of freedom, and determine the *p*-value. Are the results significant?

Text

Description automatically generated

The Degree’s of Freedom are: 3

The P-Value is: <2e-16

From the P-Value that I got; the results are significant because a P-Value of 2e-16 is a lot less than .05. This means that I have to reject the Null Hypothesis. This means that there is a relationship between Highest Educational Attained & Voting on Nov. 2012.

d. Make a bar chart that displays how voting patterns are related to highest educational attainment. (Your choice of which variable is the explanatory variable should be evident in your display.) Label the bars with the corresponding percentages. (At this point, you may just want to copy your display into Paint and then write the precents above each bar. We will learn how to do the labeling in R later.) Describe the nature of the relationship between the two variables.

![Chart, bar chart, histogram

Description automatically generated]()

From this bar plot, I can understand that there is an association between Education (The Explanatory Variable) and Voting on November 2012 (The Response Variable).

The graph shows that the people with higher education had higher percentages for voting in November 2012.

The graph also shows that the people that lower education had higher percentages for not going out to vote in November 2012.