Tidy Tuesday Data Project Step 3

Saakshi Shah

April 2021

Chi-square Test of Independence for Analysis of a Contingency Table

Research Question :

Is there an association between pursing a certain category of majors and the sex of the respondents?

Response Variable

Pursuing a certain category of majors

Type: Categorical

Explanatory Variable

Sex of the respondents

Type: Categorical

Creating Contingency Table From Data

To begin the analysis, the first thing done was manipulate the original dataset into a data frame that only concerns the categorical variables needed for this relationship analysis. In order to do this, we used the data frame and gather functions as seen below.

```r
my_old_data <- data.frame(Men, Women, Major_category)
data <- gather(my_old_data, gender, response, Women:Men)

attach(data)
```

With the new data frame, we know have a variable ‘gender’ and can proceed to create a contingency table. Contingency tables allow us to see the association between two variables. The following code allowed us to create a contingency table and add margins to the table so we can see the respective marginal distributions.

```
#Constructing the contingency table
contingency_table <- xtabs(response~gender+Major_category)

#Add margins to the table
addmargins(contingency_table)
```

```
##        Major_category
## gender  Agriculture & Natural Resources    Arts Biology & Life Science Business
##   Men                             40357  134390                 184919   667852
##   Women                           35263  222740                 268943   634524
##   Sum                             75620  357130                 453862  1302376
##        Major_category
## gender  Communications & Journalism Computers & Mathematics Education
##   Men                        131921                  208725    103526
##   Women                      260680                   90283    455603
##   Sum                        392601                  299008    559129
##        Major_category
## gender  Engineering  Health Humanities & Liberal Arts
##   Men        408307   75517                    272846
##   Women      129276  387713                    440622
##   Sum        537583  463230                    713468
##        Major_category
## gender  Industrial Arts & Consumer Services Interdisciplinary
##   Men                                103781              2817
##   Women                              126011              9479
##   Sum                                229792             12296
##        Major_category
## gender  Law & Public Policy Physical Sciences Psychology & Social Work
##   Men                 91129             95390                    98115
##   Women               87978             90089                   382892
##   Sum                179107            185479                   481007
##        Major_category
## gender  Social Science     Sum
##   Men           256834 2876426
##   Women         273132 3895228
##   Sum           529966 6771654
```

Step 1: Specify the Null and Alternative Hypotheses.

$\mathbf{H_0}$: Pursing a certain category of majors and the sex of the respondent are independent.

$\mathbf{H_A}$: Pursing a certain category of majors and the sex of the respondent are dependent.

Step 2: State and check whether the Assumptions about Statistical Model is met.

Since 6771654 graduates are randomly selected, the trials are independent and the probabilities are viewed as remaining constant from trial to trial. (Sourced from here and here)

Thus, the assumptions regarding a multinomial experiment is met.

Step 3: State the Value of the Observed Test-Statistic.

We can use the built-in chisq.test function of R to find out test statistics. The following code shows how to:

```
chisq.test(contingency_table)
```

```
##
##  Pearson's Chi-squared test
##
## data:  contingency_table
## X-squared = 783669, df = 15, p-value < 2.2e-16
```

```
format(2.2e-16, scientific = FALSE)
```

```
## [1] "0.00000000000000022"
```

The degrees of freedom is 15 (=(2-1)x(16-1)). The value of the observed test-statistic is $\chi^2 = 783669$

Step 4: State the p-value of the Observed Test-Statistic.

Using the built-in value of the format function we see that the p-value of the observed test-statistic is

p-value = 0.00000000000000022

Step 5: Make a Decision (e.g., reject Ho, fail to reject Ho) at the Significance-Level of = 0.05.

We see that the p-value (0.00000000000000022) < 0.05

Decision: We reject $\mathbf{H_0}$ at the significance-level of $\alpha = 0.05$

Step 6: In plain, Non-Statistical Language, give a Conclusion (if any, at all) from your Analysis.

Conclusion:

- We have strong evidence to indicate that pursing a certain category of majors and the sex of the respondents are dependent (associated).
- This result provides strong evidence against Ho.
- It seems likely that pursing a certain category of majors and the sex of the respondents are associated in the population.
- If the variables were independent, it would be highly unusual for a random sample to have a large χ^2 statistic.