Chapter 11: Association Between Two Categorical Variables

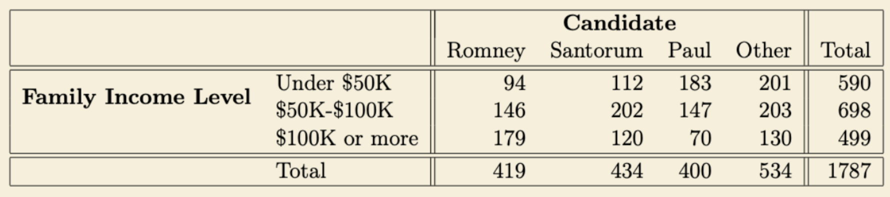
Yu Yang

School of Statistics
University of Minnesota

November 21, 2022

Example 11.1

“Entrance polls” from the 2012 Iowa Republican caucuses collected information from 1,787 caucus voters. This survey data is summarized in the following contingency table which breaks down the number of votes for each candidate by family income level:



		Candidate				
		Romney	Santorum	Paul	Other	Total
Family Income Level	Under $50K	94	112	183	201	590
	$50K-$100K	146	202	147	203	698
	$100K or more	179	120	70	130	499
	Total	419	434	400	534	1787

What are the two categorical variables in this study?

Family income level and candidate.

Is there an association between candidate and family income level among the entire population of Iowa Republican caucus voters?

Chi-Squared Test for Independence

Definition: Independent Variables

Two categorical variables are *independent* if the distribution of one of the variables is not influenced by the observed value of the other.

Goal

Use the **chi-squared test for independence** to test for a dependence or association between two categorical variables.

Main Idea

Compare observed cell counts to the cell counts we would expect to see if the 2 variables are independent. If there is a large enough discrepancy, we will conclude that the variables are dependent.

Expected Cell Counts:

In general, when two categorical variables are independent, we can calculate the expected value of each cell in the contingency table:

$$\text{expected cell count} = \frac{\text{Row total} \ \times \ \text{Column Total}}{\text{Total Sample Size}}$$

Chi-Squared Test for Independence

1. Assumptions
   - random sample
   - large enough sample size such that expected cell count ≥ 5 in all cells
2. Hypothesis
   H_0: The two variables are independent
   H_a: The two variables are dependent (associated)
3. Test Statistic
   Recall: A test statistic is a measure of how compatible the data is with H_0. In this case, a test statistic should measure the degree to which the observed contingency table agrees with the assumption of independence between the two variables.
   $O =$ observed cell count
   $E =$ expected cell count

$$X^2 = \sum \frac{(O-E)^2}{E}$$

Chi-Squared Test for Independence

3. Test Statistic

   When H_0 is true, do you expect X^2 to be a large or a small number?

   $$\begin{aligned} H_0 \text{ true} \quad &\Rightarrow O \text{ and } E \text{ should be similar} \\ &\Rightarrow X^2 \text{ should be small} \end{aligned}$$

   What is the distribution of X^2 when H_0 is true?

   $$X^2 \sim \chi^2_{(r-1)(c-1)}$$

   That is, X^2 has a χ^2 (chi-squared) distribution with df $= (r-1)(c-1)$, where $r =$number of rows and $c =$number of columns in the contingency table.

4. p-value

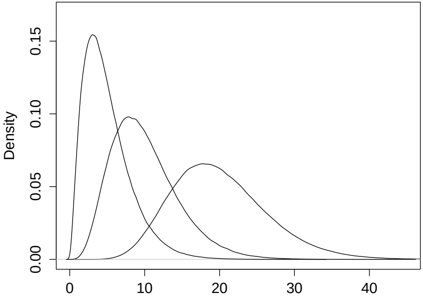
   $p\text{-value} = P(\chi^2_{(r-1)(c-1)} \geq X^2)$

5. Conclusion

   If p-value $< \alpha$, reject H_0

   If p-value $\geq \alpha$, fail to reject H_0

The Chi-Squared Distribution



Properties:

1. continuous
2. right skewed
3. only takes on non-negative values (≥ 0)
4. shape is specified by the df
5. the larger df, the more spread out

χ^2 calculations

Use the pchisq() function in R to compute an exact *p*-value for a given test statistic X^2.

1. Let df=20 and find $P(\chi^2 \geq 34.17)$.

   ```
   > pchisq(34.17, df = 20, lower.tail = FALSE)
   [1] 0.02499745
   ```

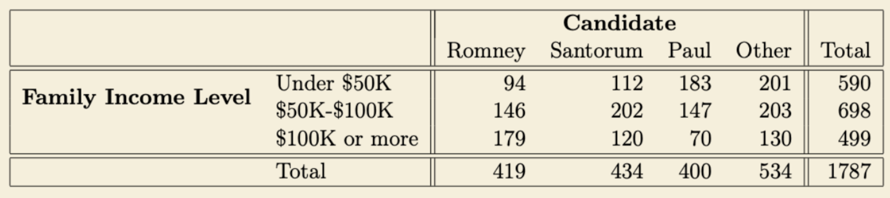
   $P(\chi^2 \geq 34.17) = .025$

2. Let df=13 and estimate $P(\chi^2 \geq 24)$.

   ```
   > pchisq(24, df = 13, lower.tail = FALSE)
   [1] 0.03113006
   ```

Example 11.1 (continued)

"Entrance polls" from the 2012 Iowa Republican caucuses collected information from 1,787 caucus voters. This survey data is summarized in the following contingency table which breaks down the number of votes for each candidate by family income level:



		Candidate				
		Romney	Santorum	Paul	Other	Total
Family Income Level	Under $50K	94	112	183	201	590
	$50K-$100K	146	202	147	203	698
	$100K or more	179	120	70	130	499
	Total	419	434	400	534	1787

Example 11.1 (continued)

(a) Calculate the expected cell counts under the assumption that one's family income level and candidate preference are independent.

	Romney	Santorum	Paul	Other	Total
Under $50K	94 **(138.3)**	112 **(143.3)**	183 **(132.1)**	201 **(176.3)**	590
$50K-$100K	146 **(163.7)**	202 **(169.5)**	147 **(156.2)**	203 **(208.6)**	698
$100K or more	179 **(117.0)**	120 **(121.2)**	70 **(111.7)**	130 **(149.1)**	499
Total	419	434	400	534	1787

Example 11.1 (continued)

(b) At the 0.05 level, test for an association between one's family income and their candidate preference.

Assumptions:

random samples, expected cell counts all > 5

Hypotheses:

H_0: Family income level and candidate are independent

H_a: Family income level and candidate are dependent

Test statistic:

$$\begin{aligned} X^2 &= \sum \frac{(\mathrm{O} - \mathrm{E})^2}{\mathrm{E}} \\ &= \frac{(94 - 138.3)^2}{138.3} + \cdots + \frac{(130 - 149.1)^2}{149.1} \\ &= 103.85 \end{aligned}$$

$\mathrm{df} = (r - 1)(c - 1) = (3 - 1)(4 - 1) = 6$

Example 11.1 (continued)

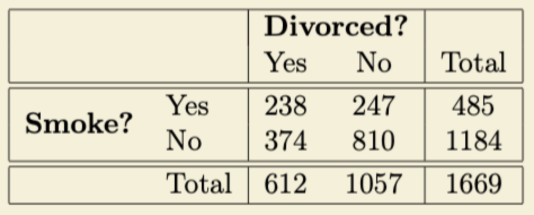
p-value:
$p\text{-val} = P(\chi_6^2 \geq 103.85) = 3.94 \times 10^{-20}$

Conclusion: Reject H_0. There is a significant association between family income and candidate at the 0.05 level.

Nov 21 Lecture Stopped Here

Example 11.2

Are smoking and divorce related? A random sample of 1669 adults were interviewed about their marriage and smoking statuses:



		Divorced?		
		Yes	No	Total
Smoke?	Yes	238	247	485
	No	374	810	1184
	Total	612	1057	1669

Use R to test for an association between smoking and divorce at the 0.01 level.

Example 11.2

- Step 1: Put the data in matrix form.

```
> dat <- matrix(c(238, 247, 374, 810), nrow = 2, byrow = TRUE)
> dat
     [,1] [,2]
[1,]  238  247
[2,]  374  810
```

  NOTE: ‘nrow=2’ tells R that there are 2 rows in the table and ‘byrow=T’ tells R that we are entering in the data for the first row followed by the data for the second row (as opposed to entering data for the first column followed by the second column).

- Step 2: Run the Chi-squared test.

```
> chisq.test(dat, correct = FALSE)

Pearson’s Chi-squared test


data:  dat
X-squared = 45.292, df = 1, p-value = 1.697e-11
```

Example 11.2

Assumptions:
random sample, expected cell counts > 5 (see note 1 below)

Hypotheses:
H_0: smoking and divorce are independent
H_a: smoking and divorce are dependent

Test statistic:
$X^2 = 45.29, \mathrm{df} = 1$

p-value:
$p\text{-val} = P(\chi^2_1 \geq 45.29) = 1.697 \times 10^{-11}$

Conclusion: Reject H_0. There is a significant association between smoking and divorce at the 0.01 level.

Measures of Association

The chi-squared test addresses:
whether or not an association between 2 categorical variables *exists*.

The chi-squared test does *not* address:
the strength of the association.

Measures of Association

Definition: risk

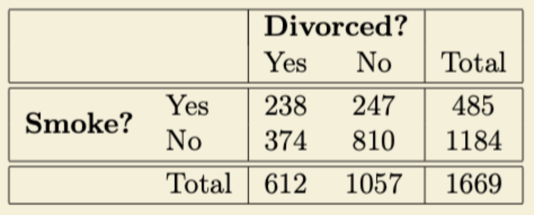
The *risk* of an outcome is the probability of its occurrence.

Definition: relative risk

The ratio of risks for two groups is called the *relative risk* and can be used to measure the strength of the association between two categorical variables.

Example 11.2 (continued)

We previously showed that there is a significant association between the incidence of smoking and divorce. We now want to *describe* this association.



		Divorced?		
		Yes	No	Total
Smoke?	Yes	238	247	485
	No	374	810	1184
	Total	612	1057	1669

Example 11.2 (continued)

(a) What is the estimated risk of divorce among smokers?
$$\frac{238}{485} = 0.491$$

(b) What is the estimated risk of divorce among non-smokers?
$$\frac{374}{1184} = 0.316$$

(c) Calculate and interpret the estimated relative risk of divorce among smokers and non-smokers.

$$\frac{0.491}{0.316} = 1.554$$

Based on this sample, we would estimate that those who smoke are about 1.55 times more likely to divorce than those that do not smoke.