Public Policy 529
Association Between Categorical Variables
Part 2

Jonathan Hanson

Gerald R. Ford School of Public Policy
University of Michigan

November 7, 2022

Outline

1. Recap: The χ^2 Test of Independence

2. Assessing the Relationship

3. Small Sample Tests

Outline

1. Recap: The χ^2 Test of Independence

2. Assessing the Relationship

3. Small Sample Tests

Key Points

- This test applies when our variables of interest are categorical and we have a (relatively) large sample.
- In a joint frequency distribution, the χ^2 statistic measures deviation from the scenario in which the variables are independent.
- The χ^2 statistic is non-negative, and 0 indicates the variables are independent.
- We reject the null hypothesis of independence when the χ^2 statistic is sufficiently large.
- This critical value of χ^2 is determined by α and degrees of freedom.

Formula for the Test Statistic

A deviation from the scenario of independence occurs when the observed frequency (f_o) differs from the expected frequency (f_e).

The χ^2 statistic sums up the deviations of the expected frequency (f_e) from f_o for all of the interior cells. The formula is:

$$\chi^2 = \sum \frac{(f_o - f_e)^2}{f_e}$$

For each cell, we square the deviation between f_o and f_e and divide the result by f_e. We then add up the resulting numbers.

Example: Satisfaction with ACA Health Plans

- The Affordable Care Act (i.e. ObamaCare) led to the creation of health insurance marketplaces in the states, where individuals could shop for health insurance plans.
- Suppose you are asked to perform an analysis that compares how satisfied people are with these plans compared to those who have other forms of health coverage.
- How could we do this?

Example: Satisfaction with ACA Health Plans

- The Health Reform Monitoring Survey (Oct 2016) provides data to compare satisfaction between ACA marketplace plans and other kinds of insurance.
- Dependent variable: level of satisfaction with health insurance (satisfied, neutral, and dissatisfied).
- Independent variable: health plan type (ACA health exchange or other health insurance).

Example: Satisfaction with ACA Health Plans

Health Plan Type

Satisfaction	ACA	Other	Total
Satisfied	71.2% (772)	79.9% (5,096)	78.6% (5,868)
Neutral	18.2% (197)	14.1% (899)	14.7% (1,096)
Dissatisfied	10.7% (116)	6.0% (382)	6.7% (498)
Total	100% (1,085)	100% (6,377)	100% (7,462)

What distribution should we see if Satisfaction and Health Plan Type are independent?

Expected Distribution Under H_O

	Health Plan Type		
Satisfaction	ACA	Other	Total
Satisfied	78.6% (853.2)	78.6% (5014.8)	78.6% (5,868)
Neutral	14.7% (159.4)	14.7% (936.6)	14.7% (1,034)
Dissatisfied	6.7% (72.4)	6.7% (425.6)	6.7% (498)
Total	100% (1,085)	100% (6,377)	100% (7,462)

Deviations from Expected Distribution

Each cell contains $f_o - f_e$

	Health Plan Type	
Satisfaction	ACA	Other
Satisfied	772 - 853.2	5096 - 5014.8
Neutral	197 - 159.4	899 - 936.6
Dissatisfied	116 - 72.4	382 - 425.6

Note: this is also how we calculate the cell "residuals," which will come up later.

Applying the χ^2 Formula

Each cell contains: $(f_o - f_e)^2/f_e$

	Health Plan Type	
Satisfaction	ACA	Other
Satisfied	$\frac{(772-853.2)^2}{853.2}$	$\frac{(5096-5014.8)^2}{5014.8}$
Neutral	$\frac{(197-159.4)^2}{159.4}$	$\frac{(899-936.6)^2}{936.6}$
Dissatisfied	$\frac{(116-72.4)^2}{72.4}$	$\frac{(382-425.6)^2}{425.6}$

Summing up the Results

Each cell contains: $(f_o - f_e)^2/f_e$

	Health Plan Type	
Satisfaction	ACA	Other
Satisfied	7.7	1.3
Neutral	8.9	1.5
Dissatisfied	26.3	4.7

$$\chi^2 = 7.7 + 1.3 + 8.9 + 1.5 + 26.3 + 4.7 = 50.4$$

Degrees of Freedom

d.f. $=$ ($\#$ rows - 1)($\#$ columns - 1)

	Health Plan Type	
Satisfaction	ACA	Other
Satisfied	7.7	1.3
Neutral	8.9	1.5
Dissatisfied	26.3	4.7

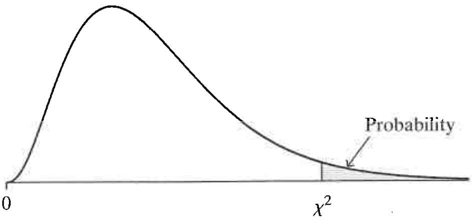
With 3 rows and 2 columns, degrees of freedom equals 2.

Finding the Critical Value of χ^2

- The χ^2 test is a one-sided test in which the entire rejection region is in the right-hand tail.
- If $\alpha = .05$, the critical value of χ^2 is the value that leaves 5% of the area in the right-hand tail.
- With 2 degrees of freedom, the critical value is 5.99 (see table).
- The χ^2 statistic from our test is 50.4, which exceeds the critical value.

$\Rightarrow$ Satisfaction with health insurance and health plan type are not independent.

TABLE C: Chi-Squared Distribution Values for Various Right-Tail Probabilities



df	Right-Tail Probability						
	0.250	0.100	0.050	0.025	0.010	0.005	0.001
1	1.32	2.71	3.84	5.02	6.63	7.88	10.83
2	2.77	4.61	5.99	7.38	9.21	10.60	13.82
3	4.11	6.25	7.81	9.35	11.34	12.84	16.27
4	5.39	7.78	9.49	11.14	13.28	14.86	18.47
5	6.63	9.24	11.07	12.83	15.09	16.75	20.52
6	7.84	10.64	12.59	14.45	16.81	18.55	22.46
7	9.04	12.02	14.07	16.01	18.48	20.28	24.32
8	10.22	13.36	15.51	17.53	20.09	21.96	26.12
9	11.39	14.68	16.92	19.02	21.67	23.59	27.88
10	12.55	15.99	18.31	20.48	23.21	25.19	29.59

Running the Test in Stata

```
. tab Satisfaction PlanType, chi
```

Satisfaction	Is your current coverage a health insurance plan through the marketplace? ACA	Other	Total
Satisfied	**772**	**5,096**	**5,868**
Neutral	**197**	**899**	**1,096**
Dissatisfied	**116**	**382**	**498**
Total	**1,085**	**6,377**	**7,462**

Pearson chi2(**2**) = **50.1539** Pr = **0.000**

Running the Test in R

```
> health.table <- table(health$Satisfaction, health$PlanType)
> addmargins(health.table)

                ACA Other  Sum
  Satisfied     772  5096 5868
  Neutral       197   899 1096
  Dissatisfied  116   382  498
  Sum          1085  6377 7462
> chisq.test(health.table, correct = F)

	Pearson's Chi-squared test

data:  health.table
X-squared = 50.154, df = 2, p-value = 1.286e-11
```

Outline

1. Recap: The χ^2 Test of Independence

2. Assessing the Relationship

3. Small Sample Tests

Assessing the Relationship

- The χ^2 statistic does not tell us anything about the nature of the relationship or the strength of association.
- It tells us whether we there is a statistically significant deviation from the scenario of independence between the variables.
- We can, however, examine the data and use other techniques to assess the nature and direction of the relationship.

  e.g. those with ACA insurance expressed lower levels of satisfaction.

Example: Satisfaction with ACA Health Plans

At the most basic level, we can just compare the percentages across the categories of the independent variable.

	Health Plan Type		
Satisfaction	ACA	Other	Total
Satisfied	71.2% (772)	79.9% (5,096)	78.6% (5,868)
Neutral	18.2% (197)	14.1% (899)	14.7% (1,096)
Dissatisfied	10.7% (116)	6.0% (382)	6.7% (498)
Total	100% (1,085)	100% (6,377)	100% (7,462)

Satisfaction is about 8.4 percentage points higher among those with other types of insurance.

Analysis of Residuals

- To be more systematic, we can analyze the residuals (i.e. the deviations).
- For which categories are the deviations positive or negative?
- Are the deviations large or small? We can measure this in context.

Calculating the Residuals

Each cell contains $f_o - f_e$

	Health Plan Type	
Satisfaction	ACA	Other
Satisfied	772 - 853.2	5096 - 5014.8
Neutral	197 - 159.4	899 - 936.6
Dissatisfied	116 - 72.4	382 - 425.6

Some residuals are positive and others are negative.

Calculating the Residuals

Note that the residuals cancel each other out mathematically, both horizontally and vertically.

	Health Plan Type		
Satisfaction	ACA	Other	Net Change
Satisfied	-81.2	81.2	0
Neutral	37.6	-37.6	0
Dissatisfied	43.6	-43.6	0
Net Change	0	0	0

We see that about 81 fewer people were satisfied with ACA plans than expected. These 81 people instead were Neutral or Dissatisfied.

Interpreting the Magnitude of the Residuals

- It is helpful to have a way to determine whether a residual is large or small.

  e.g. is 81 fewer people a lot or a little?

- To accomplish this goal, we can standardize the residuals (i.e. convert them into a z score).

$$z = \frac{f_o - f_e}{se} = \frac{f_o - f_e}{\sqrt{f_e(1 - \text{row proportion})(1 - \text{column proportion})}}$$

Example: Standardizing the Residual

The residual for the Satisfied/ACA cell was -81.2. What is z? Use the observed frequencies to find the proportions.

	Health Plan Type		
Satisfaction	ACA	Other	Total
Satisfied	772	5,096	5,868
Neutral	197	899	1,096
Dissatisfied	116	382	498
Total	1,085	6,377	7,462

The row proportion is $5{,}868/7{,}462 = .786$. The column proportion is $1{,}085/7{,}462 = .145$.

Example: Standardizing the Residual

Recalling that the expected frequency for the Satisfied/ACA cell (f_e) was 853.2:

$$
\begin{aligned}
z &= \frac{f_o - f_e}{se} \\
&= \frac{f_o - f_e}{\sqrt{f_e(1 - \text{row proportion})(1 - \text{column proportion})}} \\
&= \frac{-81.2}{\sqrt{853.2(1 - .786)(1 - .145)}} \\
&= -6.50
\end{aligned}
$$

A z-score of -6.50 is very far out in the left-hand tail of the standard normal distribution. This is a very large residual.

Example: All the Residuals

	Health Plan Type	
Satisfaction	ACA	Other
Satisfied	-6.51	6.51
Neutral	3.49	-3.49
Dissatisfied	5.74	-5.74

For all six cells, the standardized residuals are large. The sign also tells us the direction of the deviation from the null hypothesis.

Adjusted Residuals in Stata

- There is a supplementary Stata command, `tabchi`, that must first be installed. Type:

```
findit tabchi
```

- Identify the correct item and follow the directions to install. Then:

```
tabchi var1 var2, adjust
```

- In R, you need to install the `questionr` library. Then:

```
chisq.residuals(table, digits = 2, std = TRUE)
```

```
. tabchi Satisfaction PlanType, a

          observed frequency
          expected frequency
          adjusted residual
```

Satisfaction	Is your current coverage a health insurance plan through the marketplace? ACA	Other
Satisfied	**772**	**5096**
	853.227	**5014.773**
	-6.508	**6.508**
Neutral	**197**	**899**
	159.362	**936.638**
	3.492	**-3.492**
Dissatisfied	**116**	**382**
	72.411	**425.589**
	5.736	**-5.736**

```
         Pearson chi2(2) =  50.1539   Pr = 0.000
likelihood-ratio chi2(2) =  45.8674   Pr = 0.000
```

Working with Ordinal Variables

- When the variables are ordinal, there can be a positive or negative relationship between them.
- By “positive” we mean that when an observation is higher (lower) on one variable it tends to be higher (lower) on the other
- The χ^2 statistic does not measure this.
- A variety of other statistics do: gamma, Kendall’s tau-b, etc.
- These statistics range from -1 to $+1$, where 0 means there is no relationship between the variables.

```
. tab LegalPot RestrictGuns, col chi
```

Key
frequency
column percentage

POST: Should marijuana be legal?	PRE: Should fed govt make it more difficult to buy a gun 1. More d	2. Keep t	3. Easier	Total
1. Oppose	**509**	**462**	**81**	**1,052**
	26.16	**32.04**	**35.53**	**29.09**
2. Neither favor nor	**455**	**407**	**60**	**922**
	23.38	**28.22**	**26.32**	**25.50**
3. Favor	**982**	**573**	**87**	**1,642**
	50.46	**39.74**	**38.16**	**45.41**
Total	**1,946**	**1,442**	**228**	**3,616**
	100.00	**100.00**	**100.00**	**100.00**

Pearson chi2(**4**) = **44.8000** Pr = **0.000**

```
. tab LegalPot RestrictGuns, col taub gamma
```

Key
frequency
column percentage

POST: Should marijuana be legal?	PRE: Should fed govt make it more difficult to buy a gun 1. More d	2. Keep t	3. Easier	Total
1. Oppose	**509**	**462**	**81**	**1,052**
	26.16	**32.04**	**35.53**	**29.09**
2. Neither favor nor	**455**	**407**	**60**	**922**
	23.38	**28.22**	**26.32**	**25.50**
3. Favor	**982**	**573**	**87**	**1,642**
	50.46	**39.74**	**38.16**	**45.41**
Total	**1,946**	**1,442**	**228**	**3,616**
	100.00	**100.00**	**100.00**	**100.00**

gamma = **-0.1611** ASE = **0.025**
Kendall's tau-b = **-0.0967** ASE = **0.015**

Interpretation

- Both the gamma statistic and the tau-b statistic suggest a mild negative relationship between the two variables.
- People who are opposed to legalizing marijuana are more likely to favor making it easier to buy guns, and vice-versa.
- These measures allow us to be a bit more systematic when assessing the relationship between ordinal variables.

```
. tab SpendSchools SpendChildCare, col taub gamma nokey
```

PRE: Federal Budget Spending: public schools	PRE: Federal Budget Spending: Social Security 1. Decrea	2. Kept t	3. Increa	Total
1. Decreased	**180**	**101**	**43**	**324**
	31.86	**6.25**	**2.11**	**7.69**
2. Kept the Same	**151**	**571**	**230**	**952**
	26.73	**35.36**	**11.30**	**22.59**
3. Increased	**234**	**943**	**1,762**	**2,939**
	41.42	**58.39**	**86.58**	**69.73**
Total	**565**	**1,615**	**2,035**	**4,215**
	100.00	**100.00**	**100.00**	**100.00**

```
                 gamma =   0.6213  ASE = 0.018
       Kendall's tau-b =   0.3696  ASE = 0.013
```

These variables have a positive relationship.

```
. tab SpendWelfare DeficitImpt, col taub gamma nokey
```

PRE: Federal Budget Spending: welfare programs	POST: Importance of reducing deficit 1. Not at	2. A litt	3. Modera	4. Very I	5. Extrem	Total
1. Decreased	**15** **36.59**	**31** **23.13**	**198** **29.55**	**589** **44.96**	**862** **59.24**	**1,695** **46.95**
2. Kept the Same	**7** **17.07**	**45** **33.58**	**287** **42.84**	**495** **37.79**	**418** **28.73**	**1,252** **34.68**
3. Increased	**19** **46.34**	**58** **43.28**	**185** **27.61**	**226** **17.25**	**175** **12.03**	**663** **18.37**
Total	**41** **100.00**	**134** **100.00**	**670** **100.00**	**1,310** **100.00**	**1,455** **100.00**	**3,610** **100.00**

```
              gamma =  -0.3393  ASE = 0.021
    Kendall's tau-b =  -0.2254  ASE = 0.014
```

These variables have a negative relationship.

The gamma and Kendall tau-b Statistics in R

- First, one must install the `DescTools` library.
- Second, make your table object with the table command.
- For the gamma statistic: `GoodmanKruskalGamma(table)`
- For the tau-b statistic: `KendallTauB(table)`.

Outline

1. Recap: The χ^2 Test of Independence

2. Assessing the Relationship

3. Small Sample Tests

Fisher's Exact Test

- We have already seen the small-sample counterpart of the χ^2 test: the Fisher's Exact test.
- The Fisher's test continues to work for larger sample sizes, though χ^2 is computationally easier.
- When expected cell sizes become too small, however, we can no longer use χ^2.

```
. tab Satisfaction PlanType, chi exact

Enumerating sample-space combinations:
stage 3:  enumerations = 1
stage 2:  enumerations = 103
stage 1:  enumerations = 0
```

	Is your current coverage a health insurance plan through the marketplace?		
Satisfaction	ACA	Other	Total
Satisfied	**772**	**5,096**	**5,868**
Neutral	**197**	**899**	**1,096**
Dissatisfied	**116**	**382**	**498**
Total	**1,085**	**6,377**	**7,462**

Pearson chi2(**2**) = **50.1539** Pr = **0.000**
Fisher's exact = **0.000**