Chi-square Tests and Measures of Association

Dr Tom Ilvento
Department of Food and Resource Economics
University of Delaware

Overview

- This lecture will continue the discussion of Chi-squares tests for table data
- We will discuss some Measures of Association relevant to table data
- We will look at more complex tables

Measures of Association

- **Measures of Association** – summary measures that tell us the presence, direction, and strength of a relationship between two or more variables
- Key criteria of a measure of association
  - What is the range?
  - Is it bounded or either or both ends?
  - Does it show direction?
  - Is it symmetrical?
  - **Is it invariant to scale?**
  - What are the underlying assumptions?
  - How do I interpret it – at the extremes and in the middle

Chi-square as a Measure of Association

- By now you might have realized that the size of the test statistic gives an immediate sense of a finding – something greater than 2 is significant; something very large is very significant!
- If we think of the numerator of the test statistics as the **"effect"** (the difference from the null value, or the difference between means), the effect has the most impact on the size of the test statistic.
- The sample size also has an effect by reducing the denominator of the test statistic, thereby making is smaller
- With the chi-square test, the sample size has a very large impact on the size of chi-square
- Thus we don't want to use $\chi^2{}_*$ as a measure of association

Look what happens to $\chi^2{*}$ when n doubles

- $\chi 2{*}$ doubles when n doubles, but conditional probabilities and odds ratios would not change

DOUBLED Odds Ratio

Odds Ratio	Lower 95%	Upper 95%
7.664052	4.764336	12.32862

Table Measures of Association

- **General Measures**
  - $\chi^2{*}$ and cell contributions
  - Conditional probabilities
- **Specific to a 2x2 Table**
  - Odds Ratio
  - Yule's Q
  - Rho
- **Adjust to $\chi^2{*}$**
  - Cramer's V
  - Phi ϕ
  - Contingency Coefficient P

	Altered	Not Altered	Row Total
E2F1	41	51	92
Control	15	143	158
Column Total	56	194	**250**

Measures for a 2x2 Table: Odds Ratio

- **Odds Ratio θ** - the ratio of two odds. It is always positive and has no upper bound.
- The formula to the right is a shorthand way and is algebraically identical to the ratio of the two odds
- A value greater than one means the probability for the first group is larger than the probability for the second group
- θ is not a symmetric measure of association - it matters what order.
- But, in a 2x2 Table there is really only one unique odds ratio, but the result is different depending on how the table is organized
- In a general rxc table there may be many odds ratios

$$\theta = \frac{(c_{11} * c_{22})}{(c_{21} * c_{12})}$$

$$\theta = \frac{(41 * 143)}{(15 * 51)} = 7.664$$

The E2F1 group is 7.7 times more likely to altered cells compared to the Control group

Measures for a 2x2 Table: Yules' Q

- **Yule's Q -** variation of the odds ratio for a 2x2 table. It shows direction and strength of a relationship.
- It is like a correlation coefficient, a positive Q means more of variable 1 is associated with more of variable 2.
- Yule's Q bounds the odds ratio to -1 to 1.
  - A value close to 1 indicates a strong positive relationship between the two variables;
  - a value of -1 show a strong negative relationship.
  - A value of zero means no relationship
- Another way to express Q = (θ-1)/(θ+1)
- It is a symmetric measure of association - but the sign will change

$$Q = \frac{[(c_{11} * c_{22} - c_{12} * c_{21})]}{[(c_{11} * c_{22} + c_{12} * c_{21})]}$$

$$Q = \frac{[(41 * 143 - 15 * 51)]}{[(41 * 143 + 15 * 51)]} = .769$$

Moving from the E2F1 group to the Control group is strongly related to having more Un-Altered cells

Measures for a 2x2 Table: Rho, the Correlation Coefficient

- **Rho ρ** (correlation coefficient) - a measure of association that also shows strength and direction.
- Like Yule's Q, it ranges from -1 to 1.
- It assumes that the value in row 2 and column 2 are "more"
- The formula is for a 2x2 table. Notice the denominator is based on row and column marginals: c_{1*} is the row 1 marginal
- Thus the difference in the numerator is made relative to the square root of the product of the marginals in the table
- It is more complicated to calculate in the general rxc table and should only be used when the variables are ordinal.

$$\rho = \frac{\left[(c_{11} * c_{22} - c_{12} * c_{21})\right]}{\sqrt{(c_{1*} * c_{2*} * c_{*2} * c_{*1})}}$$

$$\rho = \frac{\left[(41 * 143 - 15 * 51)\right]}{\sqrt{(92 * 158 * 194 * 56)}} = .406$$

Moving from the E2F1 group to the Control group is positively related to having more Un-Altered cells

Table Measures of Association that adjust χ^2_*

- **Cramer's V** - a measure of association that ranges from 0 to 1. A value closer to 1 indicates stronger association between the two variables.

$$V = \sqrt{\frac{\chi^2}{n * \min(r-1, c-1)}}$$

V = .406

- **Phi** – another measure based on Chi-square. It ranges from zero to one, although its upper bound may not always be 1 (depending upon marginal distributions).

$$\phi = \sqrt{\frac{\chi^2}{n}}$$

ϕ = .406

- **Contingency Coefficient** - Denoted as **P**, the contingency coefficient is another measure based on chi-square, with a range of zero to one.

$$P = \sqrt{\frac{\chi^2}{(\chi^2 + n)}}$$

P = .376

New Example

- Smoking cessestion data
- Let's analyze it using chi-square and some measures of association

		Subject Still Smoking		
		YES	**NO**	**Row Margins**
Subject Treatment	**Nicotine Patch**	64	56	**120**
	Placebo	96	24	**120**
	Column Margins	**160**	**80**	**240**

Smoking Data

Notice that I expressed the table as Not Smoking to Smoking

Smoking Data Contingency Analysis

Freq: FREQ

Contingency Table

Treatment	Count Row % Expected Cell Chi^2	Outcome: Not Smoking	Outcome: Smoking	
	Nicotine	56 46.67 40 6.4000	64 53.33 80 3.2000	120
	Placebo	24 20.00 40 6.4000	96 80.00 80 3.2000	120
		80	160	240

Tests

N	DF	-LogLike	RSquare (U)
240	1	9.8043124	0.0642

Test	ChiSquare	Prob>ChiSq
Likelihood Ratio	19.609	<.0001*
Pearson	19.200	<.0001*

Fisher's Exact Test	Prob	Alternative Hypothesis
Left	1.0000	Prob(Outcome=Smoking)
Right	<.0001*	Prob(Outcome=Smoking)
2-Tail	<.0001*	Prob(Outcome=Smoking)

Odds Ratio

Odds Ratio	Lower 95%	Upper 95%
3.5	1.972778	6.209518

- **χ^2_* = 19.20, p < .0001**
  - **V = SQRT(19.20/240*1) = .283**
  - **ϕ = SQRT(19.20/240) = .283**
  - **P = SQRT(19.20/(19.20+240)) = .272**
- **Odds Ratio is 3.5: nicotine patch users 3.5 times more likely to not smoke after 8 weeks**
  - **Test for ln(θ); z* = 4.28, p < .001**
- **Yules Q = (3.5-1)/3.5+1 = .5565**
- **ρ = (56*96 - 24*64)/SQRT(80*120*160*120) = .283**
- **There is a significant moderate relationship between using a nicotine patch and not smoking after 8 weeks**

CHISQ.xls

- I have an Excel file that solves for 2x2 tables - Chisq.xls

Observed Frequencies

	Smoking	Not Smoking	Row Total
Nicotine	64	56	120
Placebo	96	24	120
Column Total	160	80	240

Expected Frequencies Under Independence

	Smoking	Not Smoking	Row Total	
Nicotine	80.000	40.000	120.0	Nicotine
Placebo	80.000	40.000	120.0	Placebo
Column Total	160.0	80.0	240.0	

Chi Square Test		19.200				
d.f.		1				
p-value		0.000	Conclusion: Reject Independence			
Critical Value		3.841				
G Likihood Ratio		19.609				
Odds of						Inverse
	Nicotine	Smoking	to	Not Smoking	1.143	0.875
	Placebo	Smoking	to	Not Smoking	4.000	0.250
Odds Ratio					0.286	3.500
Log Odds					-1.253	1.253
Yules Q		-0.556				
Rho		-0.283				
Phi		0.080	0.283			
Kramers V		0.0800	0.283			
Contingenecy Coef		0.0741	0.272			

What about a Larger Table?

- 3x3 table: Do you approve or disapprove of the way George W. Bush is handling his job as president? Broken down by party affiliation.
- How many degrees of freedom?

	Approve	Disapprove	Unsure	
Rebublican	202	66	17	**285**
Democrat	31	274	9	**314**
Independent	104	253	27	**384**
	337	**593**	**53**	**983**

The approach is the same:

1. **Generate expected frequencies under a Model of Independence**
2. **Calculate chi-square to test if the association is due to chance**
3. **If you can reject the Null Hypothesis, investigate further to understand the nature of the relationships**

JMP Output

- The degrees of freedom is (3-1)*(3-1) = 4 d.f.
- χ^2_* = 282.506, p < .0001
  - The Critical value of χ^2 for α = .01 and 4 d.f. is 13.277
  - Our value is certainly further than that in the tail of the distribution
- V = .38; ϕ=.54; P= .472
- What's going on?
  - 70.9% of Republicans Approve
  - 9.87% of Democrats Approve
  - 27.08% of Independents Approve

Contingency Analysis of APPROVAL By PARTY

Freq: FREQ

Contingency Table

APPROVAL

PARTY	Count Row % Expected Cell Chi^2	Approve	Disapprove	Not Sure	
	Democrat	31 9.87 107.648 54.5753	274 87.26 189.422 37.7644	9 2.87 16.9298 3.7143	314
	Independent	104 27.08 131.646 5.8057	253 65.89 231.65 1.9677	27 7.03 20.704 1.9146	384
	Rebublican	202 70.88 97.706 111.326	66 23.16 171.928 65.2640	17 5.96 15.3662 0.1737	285
		337	593	53	983

Tests

N	DF	-LogLike	RSquare (U)
983	4	147.06894	0.1804

Test	ChiSquare	Prob>ChiSq
Likelihood Ratio	294.138	<.0001*
Pearson	282.506	<.0001*

Adding other variables to the analysis

- It is possible to break down our table by a third or even fourth variable
- Example: who does the housework, men or women?
  - We start with a breakdown of how much housework by men and women
  - Then we break this table down by a third variable, whether they are married or not
- This type of analysis can be sensitive to the sample size
  - Let n = 400
  - For a 4x2 table, we would have an average of 400/8 = 50.0 per cell (though the actual distribution might be different)
  - If we add a third variable, which has 2 levels, we now need to fill 16 cells, 400/16 = 25.0

What about empty cells?

- Here's some data from a survey of students in the College of Agriculture and Natural Resources
- The focus was on student evaluations of their advisors, which tend to be positive
- JMP warns that 20% of cells have expected counts less than 5
- We can see that few of the students Strongly disagreed that their advisor "knows me," and few rated their advisor as "poor."
- One strategy is to collapse the data across some categories

Contingency Analysis of Overall By Knows Me

Contingency Table

Knows Me	Count / Row % / Expected	Overall: Excellent	Good	Neutral	Fair	Poor	
	Strongly Agree	94 64.38 63.61	39 26.71 44.2241	6 4.11 16.3568	7 4.79 12.722	0 0.00 9.08714	146
	Agree	10 18.87 23.0913	24 45.28 16.0539	11 20.75 5.93776	6 11.32 4.61826	2 3.77 3.29876	53
	Neutral	1 5.56 7.84232	6 33.33 5.45228	5 27.78 2.0166	3 16.67 1.56846	3 16.67 1.12033	18
	Disagree	0 0.00 6.09959	4 28.57 4.24066	3 21.43 1.56846	4 28.57 1.21992	3 21.43 0.87137	14
	Strongly Disagree	0 0.00 4.35685	0 0.00 3.02905	2 20.00 1.12033	1 10.00 0.87137	7 70.00 0.62241	10
		105	73	27	21	15	241

Tests

N	DF	-LogLike	RSquare (U)
241	16	66.604480	0.2040

Test	ChiSquare	Prob>ChiSq
Likelihood Ratio	133.209	<.0001*
Pearson	157.249	<.0001*

Warning: 20% of cells have expected count less than 5, ChiSquare suspect.

Summary

- We established a way to test for a relationship in categorical (or ordinal) data in tables using the chi-square goodness of fitness test
- It is based on a model of independence – as if there is no relationship between the two variables
- Once we establish a relationship, we can move to explore the exact nature of that relationship with various measures of association
- The chi-square test is a very general test used in many ways in statistics
- There are many other ways in which modeling can be done in contingency tables