Chi-square Tests and Table Data

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

Overview

- Next we will look at approaches when we have two or more variables - a step further than difference of means or proportions
- We will start with contingency table data – two variables that are cross-tabulated with each other
- The variables are usually categorical, although they could be ordinal
- We will introduce the chi-square goodness of fit test to our previous notion of a "model of independence"
- Along with some measures of association

We looked this data earlier as a proportion problem

- Geneticists have identified E2F1 transcription factor as an important component of cell proliferation control. The researchers induced DNA synthesis in two batches of serum-starved cells. In one group of 92 cells (treatment), cells were micro-injected with the E2F1 gene. A control group of 158 cells was not exposed to E2F1. After 30 hours, researchers determined the number of altered growth cells in each batch. The data are given below.

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

What would our data look like if the two variables were independent?

- By now you might realize that one strategy in statistics is to **propose a hypothesized value** and then **compare** what we **observe to** what is **expected** under the Null hypothesis
- We could propose a **model of independence**.
  - If our variables were independent of each other, then the data would be based only on the marginal distributions
  - We have already done this in previous lectures - a model of independence
- If there are substantial differences between what we observe and what we expect, it would cast doubt on the expectations under the Null Hypothesis

Observed versus Expected

Observed Frequencies from our Experiment			
	Altered	**Not Altered**	Row Total
E2F1	**41**	**51**	92
Control	**15**	**143**	158
Column Total	56	194	**250**

Expected Frequencies from Model of Independence			
	Altered	**Not Altered**	Row Total
E2F1	**20.608**	**71.392**	92
Control	**35.392**	**122.608**	158
Column Total	56	194	**250**

How do we solves for the Expected Frequencies?

- Remember, I wanted a model of independence, which means
  - $P(B|A) = P(A \cap B)/P(A) = P(B)$
  - $P(A|B) = P(A \cap B)/P(B) = P(A)$
- A simple way to make this happen is make the expected frequencies a function of the row and column marginals

Solving for Expected Frequencies

- **Altered, E2F1** = (56*92)/250 = 5,152/250 = **20.608**
- **Altered, Control** = (56*158)/250 = 8,848/250 = **35.392**
- **Not Altered, E2F1** = (194*92)/250 = 17,848/250 = **71.392**
- **Not Altered, Contro** = (194*158)/250 = 30,652/250 = **122.608**

Expected Frequencies from Model of Independence			
	Altered	**Not Altered**	Row Total
E2F1	**20.608**	**71.392**	92
Control	**35.392**	**122.608**	158
Column Total	56	194	**250**

The value of our model

- Generating expected frequencies under a model can be very useful
- We can compare our model to the data to see how well the data fits the expected frequencies – how we do this will come later!
- Depending upon our model, we may or may not want to see a good fit.
  - With a Model of Independence, we often don't want a good fit!
  - Because a bad fit means there is a relationship between the two variables

If two variables are not independent, they are related to each other!

Chi-Square Test for Independence

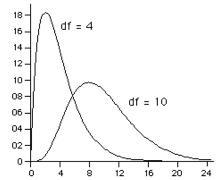
- We are now ready to make an inference
- In order to do this we need:
  - Data from a random sample which gives us Observed Frequencies O_{ij}
  - Expected frequencies based on a model of Independence E_{ij}
  - Knowledge of the form of the Sampling Distribution: Chi-square, denoted as χ^2
  - A Hypothesis Test – The test for a Model of Independence

i for row position
j for column position
For our data, i =2 and j = 2
It is a 2x2 Table

χ^2 Test for Independence

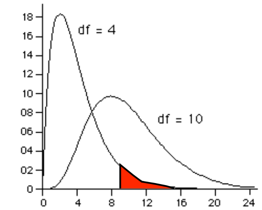
- The Chi-square test for independence seeks to determine if a relationship exists between two categorical variables
- This test is done by setting up a model of independence
- And seeing if the observed data depart from this model sufficiently to rule out independence
- The alternative hypothesis is that the variables are associated or related to each other
- This test is also known as the **Pearson Chi-square Test** or the **Chi-square Goodness of Fit Test**

χ^2 Distribution



- The Chi-square distribution is a positive skewed distribution defined by degrees of freedom.
  - The mean of a chi square distribution is the d.f., and the variance is 2*d.f.
  - The Chi-square distribution is positively skewed (right skewed), but less so as the d.f. increase. As the d.f. increase the chi-square distribution approximates a normal distribution
- The **degrees of freedom** for the contingency table test is:
  - **(Rows-1)*(Columns-1)**
- For our data, the degrees of freedom is:
  - **(2-1)*(2-1) = (1)*(1) = 1 d.f.**
- **The Chi-square distribution is involved in the t-distribution, the F-distribution, and also can be used to test hypotheses about variances and other, very general tests.**

Chi-Square Table



- When we look at the Probability Density Function (PDF) of the Chi-square distribution, we will look at probabilities of alpha, the probability of a Type I error.
- We focus on the probability in the right tail.
- Look at this partial table
  - The Chi-Square table is organized by degrees of freedom as the rows
  - And the level of alpha as the columns

For an α level of .05 and 1 d.f., the critical value of χ^2 is 3.841.

Our Tests Statistics needs to be larger than this to reject the Null Hypothesis

Chi Square Distribution Table

Area to the Right of the Critical Value

DF	0.995	0.990	0.975	0.950	0.900	0.100	0.050	0.025	0.010	0.005
1	0.000	0.000	0.001	0.004	0.016	2.706	3.841	5.024	6.635	7.879
2	0.010	0.020	0.051	0.103	0.211	4.605	5.991	7.378	9.210	10.597
3	0.072	0.115	0.216	0.352	0.584	6.251	7.815	9.348	11.345	12.838
4	0.207	0.297	0.484	0.711	1.064	7.779	9.488	11.143	13.277	14.860
5	0.412	0.554	0.831	1.145	1.610	9.236	11.070	12.833	15.086	16.750
6	0.676	0.872	1.237	1.635	2.204	10.645	12.592	14.449	16.812	18.548
7	0.989	1.239	1.690	2.167	2.833	12.017	14.067	16.013	18.475	20.278
8	1.344	1.646	2.180	2.733	3.490	13.362	15.507	17.535	20.090	21.955
9	1.735	2.088	2.700	3.325	4.168	14.684	16.919	19.023	21.666	23.589
10	2.156	2.558	3.247	3.940	4.865	15.987	18.307	20.483	23.209	25.188

χ^2 Test for Independence

- We will use the following test statistic for the Chi-square test for independence, χ^2*
- Where:
  - O_{ij} is the Observed frequency for row i and column j
  - E_{ij} is the Expected frequency for row i and column j
  - d.f. for the test is (r-1)*(c-1)

$$\chi^2* = \sum_{ij} \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$$

Computational Formula

$$\chi^2* = \sum_{ij} \frac{(O_{ij} - E_{ij})^2}{E_{ij}} =$$

$$\chi^2* = \sum_{ij} \frac{(O_{ij} - E_{ij})^2}{E_{ij}} = \sum_{ij} \frac{O_{ij}^2 - 2O_{ij}E_{ij} - E_{ij}^2}{E_{ij}}$$

$$\chi^2* = \sum_{ij} \frac{(O_{ij} - E_{ij})^2}{E_{ij}} = \sum_{ij} \frac{O_{ij}^2 - 2O_{ij}E_{ij} - E_{ij}^2}{E_{ij}} = \sum_{ij} \frac{O_{ij}^2}{E_{ij}} - \sum_{ij} 2O_{ij} + \sum_{ij} E_{ij}$$

$$\chi^2* = \sum_{ij} \frac{O_{ij}^2}{E_{ij}} - \sum_{ij} O_{ij}$$

- There is a computational formula
- It results in the same test statistic
- But is more simple and has less rounding error

1. Take each observed cell frequency
2. Square it
3. Divide by the expected cell frequency
4. Add them all together
5. Subtract n

χ^2 Test for Independence

- **Ho: Independence:** the row and column variables are independent
- **Ha: Association:** There is a relationship between the two variables
- Assumptions:
  - Random samples
  - **All expected frequencies are greater than or equal to one**
  - **No more than 20% of expected frequencies are less than 5**
- Test Statistic: $\chi^2* = \sum_{ij} \frac{O_{ij}^2}{E_{ij}} - \sum_{ij} O_{ij}$
- Rejection Region: $\chi^2_{\alpha, (r-1)(c-1) \text{ d.f}}$
- Decision: If $\chi^2* > \chi^2_{\alpha, (r-1)(c-1) \text{ d.f}}$ then reject Ho

Calculating χ^2*

$$\chi^2* = \sum_{ij} \frac{(O_{ij} - E_{ij})^2}{E_{ij}} = \qquad \chi^2* = \sum_{ij} \frac{O_{ij}^2}{E_{ij}} - \sum_{ij} O_{ij}$$

rc	Observed	Expected	Cell contribution of χ^2*	O²	O²/E
11	41	20.608	20.178	1,681	81.570
12	51	71.392	5.825	2,601	36.433
21	15	35.392	11.749	225	6.357
22	143	122.608	3.392	20,449	166.784
TOTAL	250	250	41.144		291.144

- $\chi^2* = 291.144 - 250 = 41.144$

Hypothesis Test for Altered Cell data

- Null Hypothesis
- Alternative Hypothesis
- Assumptions
- Test Statistic
- Rejection Region
- Conclusion

- **Ho: Independence**
- **Ha: Association**
- **All expected cells > 1; few to none < 5**
- **$\chi^2_* = 41.144$**
- **$\chi^2_{.05,\ 1\ d.f.} = 3.841$**
- **$\chi^2_* > \chi^2_{.05,\ 1\ d.f.}$**
- **41.144 > 3.841**
- **Reject Ho: Independence**

Hypothesis Test for Altered Cell data

- We found there was a significant difference between the sample data that we observed and the expected data under a model of independence.
- The Chi-Square test results implies that there is a relationship in the data, and that it is not likely that the relationship happened by chance.
- **Note: The $\chi 2$ test should agree with the difference in proportion test for a 2x2 table.**
- **The Chi-square test is a very general test.** Once we established a relationship, we should move to explore it more deeply
  - Look at conditional probabilities
  - A cell's contribution to chi-squared
  - Odds and odds ratios
  - Other Measures of Association

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?
  - What are the underlying assumptions?
  - How do I interpret it – at the extremes and in the middle

Examples: test statistic; odds ratio, conditional probability, correlation coefficient, R2, chi-square

Measures of Association for Table Data

- For Table Data, what Measures of Association depends upon
  - the complexity of the table (how many rows and columns)
  - Whether the levels are ordered or not
  - and whether you are able to specify one variable at dependent (or the response) variable.
- Measures of Association we will discuss
  - χ^2 very weak measure of association
  - Kramer's **V**
  - Phi **ϕ**
  - Contingency Coefficient **P**
  - Rho **ρ**
  - Odds Ratio
  - Yules **Q**

Entering the data into JMP

- The data can be in the classic form – each row is a subject and the columns represent each variable
- Or, most programs allow you to enter in data in summary form. For example, for our 2x2 table:
- r1 c1 count
- r1 c2 count
- r2 c1 count
- r2 c2 count

Subject	Treatment	Cell Result
1	E2F1	Not Altered
2	E2F1	Altered
3	Control	Not Altered
4	Control	Not Altered
5	Control	Not Altered
6	E2F1	Altered
7	Control	Altered
8	E2F1	Altered
250	Control	Altered

Treatment	Cells	FREQ
E2F1	Altered	41
E2F1	Not Altered	51
Control	Altered	15
Control	Not Altered	143

JMP Output

Contingency Analysis of Cell By Treatment

Freq: FREQ

Contingency Table

Treatment	Count / Row % / Expected / Cell Chi^2	Cell: Altered	Cell: Not Altered	
	E2F1	41 44.57 20.608 20.1783	51 55.43 71.392 5.8247	92
	Control	15 9.49 35.392 11.7494	143 90.51 122.608 3.3916	158
		56	194	250

Tests

N	DF	-LogLike	RSquare (U)
250	1	20.173585	0.1517

Test	ChiSquare	Prob>ChiSq
Likelihood Ratio	40.347	<.0001*
Pearson	41.144	<.0001*

Odds Ratio

Odds Ratio	Lower 95%	Upper 95%
7.664052	3.91278	15.01175

- The chi-square test reveals there is a significant relationship between the treatment and the response, p-value < .0001.
- Looking at the row percentages, the cells that received E2F1 showed a much higher percentage that were altered (44.57% vs 9.49% for the control)
- The contributions to chi-square show that almost 78% of $\chi^2{}_*$ is due to two cells where the expected frequencies are much different from the observed frequencies
  - for cell 1,1 (20.178) where observed for Altered E2F1 was higher than expected
  - cell 2,1 (11.749) where observed for Altered Control was lower than expected
- The odds ratio for E2F1 being altered vs the control being altered is 7.66 – E2F1 was nearly 7.7 times more likely to be altered.

Likelihood Ratio Chi-Square

- Most programs will give the Likelihood Ratio Chi-Square, sometimes referred to as G
- The Likelihood Ratio Chi-Square is very similar to Pearson's Chi-square in its results and its interpretation
- It is also based on observed and expected frequencies
- It is believed to have better asymptotic properties, especially in more complex modeling
- It would be rare that this result would not agree with the Pearson Chi-square
  - G = 40.347, p < .0001
  - $\chi^2{}_*$ = 41.144, p < .0001

$$G = 2\sum_{ij} O_{ij} * \ln\left(\frac{O_{ij}}{E_{ij}}\right)$$

Various ways to analyze the same data

- For a 2x2 table, we can:
  - A difference of proportion test
  - Conduct a χ^2 test of Independence
  - Conduct a test of the Odds Ratio

- **z* = 6.414, p < .001**
- **$\chi^2{}_*$ = 41.144, p < .001**
- **This test involves taking the natural log of the odds**
  - **Ho: ln(Odds) = 0**
  - **Standard error is a function of cell n**
  - **z* = [ln(7.664)-0]/.343**
  - **z* = 5.937, p < .001**

$$S.E._{\log odds} = \sqrt{\frac{1}{n_{11}} + \frac{1}{n_{12}} + \frac{1}{n_{21}} + \frac{1}{n_{22}}}$$

$$S.E._{\log odds} = \sqrt{\frac{1}{41} + \frac{1}{51} + \frac{1}{15} + \frac{1}{143}} = .34301$$

All these tests agree in their result - there is a difference between the treatment (E2F1) and the control group

Summary

- We established a way to test for a relationship in categorical (or ordinal) data in tables
- It is based on the difference of observed frequencies compared to expected frequencies under a specific model
- The model we looked at in a model of independence – as if there is no relationship between the two variables
- To test this, we used the chi-square distribution
- This is still based on the notion of a sampling distribution and that the relationship we observe could be by chance – we want to rule out the notion of chance
- Once we establish a relationship, we can move to explore the exact nature of that relationship with various measures of association