

STAT 100 Lecture 35 and 36:
Analysis of Categorical Data
Part 3:
Contingency Table with Neither Margin Fixed.
Test of Independence

Nate Strawn

December 10 and 12

- 1 χ^2 test of homogeneity in a contingency table.
- 2 Z test to compare two proportions.

Today's Agenda

- 1 χ^2 test of independence in a contingency table.
- 2 Spurious dependence.

Developing a χ^2 test of independence

Example

400 persons were questioned regarding union membership and attitude toward decreased national spending on social welfare programs.

	Support	Indifferent	Opposed	Total
Union	112	36	28	176
Nonunion	84	68	72	224
Total	196	104	100	400

Proportion of observations in each cell

	Support	Indifferent	Opposed	Total
Union	.28	.09	.07	.44
Nonunion	.21	.17	.18	.56
Total	.49	.26	.25	1

Developing a χ^2 test of independence

Example

<i>Cell probabilities</i>				
	<i>Support</i>	<i>Indifferent</i>	<i>Opposed</i>	<i>Total</i>
<i>Union</i>	p_{U1}	p_{U2}	p_{U3}	p_U
<i>Nonunion</i>	p_{N1}	p_{N2}	p_{N3}	p_N
<i>Total</i>	p_1	p_2	p_3	1

The null hypothesis is the two classifications are independent.

The independence of the two classifications means that

$p_{U1} = p_U \times p_1$, $p_{U2} = p_U \times p_2$, and so on.

Therefore, the null hypothesis of independence can be formalized as
 H_0 : Each cell probability is product of the corresponding pair of marginal probabilities.

Developing a χ^2 test of independence

Example

To construct a χ^2 test, we need to determine the expected frequencies. Under H_0 , the expected cell frequencies are

$$\begin{array}{ccc} 400p_{UP_1} & 400p_{UP_2} & 400p_{UP_3} \\ 400p_{NP_1} & 400p_{NP_2} & 400p_{NP_3} \end{array}$$

The unknown **marginal probabilities** must be estimated from the data:

$$\hat{p}_U = \frac{176}{400} \quad \hat{p}_N = \frac{224}{400} \quad \hat{p}_1 = \frac{196}{400} \quad \hat{p}_2 = \frac{104}{400} \quad \hat{p}_3 = \frac{100}{400}$$

The expected frequency in the first cell is estimated as

$$400\hat{p}_U\hat{p}_1 = 400 \times \frac{176}{400} \times \frac{196}{400} = \frac{176 \times 196}{400} = 86.24$$

Developing a χ^2 test of independence

Example

Notice the expected frequency for each cell is

$$\frac{\text{Row total} \times \text{Column total}}{\text{Grand total}}$$

Observed and Expected Cell Frequencies for the Data.

	Support	Indifferent	Opposed
Union(O)	112	36	28
Union(E)	86.24	45.76	44
Nonunion (O)	84	68	72
Nonunion (E)	109.76	58.24	56

The Values of $(O - E)^2 / E$

	Support	Indifferent	Opposed	Total
Union	7.695	2.082	5.181	
Nonunion	6.046	1.636	4.571	
Total				27.847 = χ^2

The χ^2 test of independence

Example

d.f. of $\chi^2 = (\text{No. of cells}) - 1 - (\text{No. of parameters estimated})$

Since $p_U + p_N = 1$ and $p_1 + p_2 + p_3 = 1$ we really estimated $1 + 2 = 3$ parameters.

*d.f. of $\chi^2 = 6 - 1 - 3 = 2$ At the level of significance $\alpha = .05$ $\chi_{\alpha}^2 = 5.99$ with d.f. = 2 . Because the observed χ^2 is larger than the tabulated value, the null hypothesis of independence is **rejected** at $\alpha = .05$.*

Degree of Freedom of χ^2 Test Statistics in a General $r \times c$ Contingency Table

Fact

- We have rc cells.
- Initially we have $d.f. = rc - 1$.
- The number of estimated parameters is $(r - 1) + (c - 1)$ because there are $r - 1$ parameters among the row marginal probabilities and $c - 1$ parameters among the column marginal probabilities.
- Therefore, $d.f. = rc - 1 - (r - 1) - (c - 1) = rc - r - c + 1 =$

$$(r - 1)(c - 1) =$$

$$(No. of rows - 1) \times (No. of columns - 1)$$

The χ^2 Test of Independence in a General $r \times c$ Contingency Table

Rule

Null hypothesis

Each cell probability equals the product of the corresponding row and column marginal probabilities.

Test statistic

$$\chi^2 = \sum_{\text{cells}} \frac{(O - E)^2}{E}$$

$$d.f = (\text{No. of rows} - 1) \times (\text{No. of columns} - 1)$$

Rejection region

$$\chi^2 \geq \chi_{\alpha}^2.$$

Fact

- *A rejection of the null hypothesis of independence leads us to conclude that the data provide evidence of a **statistical association** between the two characteristics. However, we must refrain from making the hasty interpretation that these characteristics are directly related. A claim of casual relationship must draw from **common sense**, which statistical evidence must not be allowed to supersede.*
- *Two characteristics may appear to be strongly related due to the common influence of a third factor that is not included in the study. In such cases, the dependence is called a **spurious dependence**.*

Next time

- 1 MINITAB Project 05! Due Friday (Dec 12th).
- 2 Final Review: Thursday, December 11th, 5-7pm in ARM 0126
- 3 FINAL EXAM: Monday, December 15th, 1:30-3:30pm in SPH 1312
- 4 Group Problems:

Group	1	2	3	4	5
Problem	13.21	13.23	13.24	13.25	13.26
Group	6	7	8	9	10
Problem	13.21	13.23	13.24	13.25	13.26