

Statistics Notes – χ^2

The χ^2 statistic tests categorical data either for goodness of fit (one-way table) or for independence (two-way table).

Goodness of Fit

Tests the null hypothesis that a sample distribution follows a claimed distribution.

k = number of categories degrees of freedom = k – 1

n = total sample size n_i = count of a cell

expected count of a cell $E(n_i) = (n)(\text{expected or claimed probability})$ **requires $E(n_i) \geq 5$**

H_0 : the sample follows the claimed distribution

H_a : at least one cell probability is significantly different from the cell probability of the null
(does not tell you which one(s)!))

$$\chi^2 = \sum_{i=1}^k \frac{[n_i - E(n_i)]^2}{E(n_i)} \quad \text{if } \chi^2 > \chi^2_{\alpha}, \text{ reject the null}$$

confidence interval: observed $p_i \pm z_{\alpha/2} \sqrt{\frac{p_i(1-p_i)}{n}}$

Example

An experiment produces the following data: Cell 1 = 78, Cell 2 = 60, Cell 3 = 182

Null: $p_1 = .25, p_2 = .25, p_3 = .50$

$$n = 320 \quad k = 3 \quad df = k - 1 = 2 \quad \alpha = .05 \quad \chi^2_{\alpha} = 5.99147$$

$$E_1 = .25(320) = 80 \quad E_2 = .25(320) = 80 \quad E_3 = .50(320) = 160$$

$$\chi^2 = \frac{(78-80)^2}{80} + \frac{(60-80)^2}{80} + \frac{(182-160)^2}{160} = 11.10 > \chi^2_{\alpha} \rightarrow \text{reject the null}$$

confidence interval for Cell 2: observed $p_2 = 60/320 = 0.1875$

$$0.1875 \pm 1.96 \sqrt{\frac{(.1875)(.8125)}{320}} = 0.1875 \pm 0.04277$$

we are 95% confident that the actual probability for Cell 2 falls between 0.1447 and 0.2303

Two-Way (Contingency) Tables – Test for Independence

Chi-squared can be used to test for association between a row variable and a column variable in a two-way, or contingency, table. The null hypothesis is that the variables are independent. If they are independent, then any cell probability $P(A \cap B) = P(\text{row A})P(\text{column B})$.

$$\text{Expected cell count } E_{ij} = \frac{(\text{row total})(\text{column total})}{\text{total } n} \quad \text{requires } E_{ij} \geq 5$$

$$\text{degrees of freedom} = (\# \text{ of rows} - 1)(\# \text{ of columns} - 1)$$

$$\chi^2 = \sum \frac{[n_{ij} - E_{ij}]^2}{E_{ij}}$$

confidence interval for difference between two observed proportions:

$$(p_1 - p_2) \pm z_{\alpha/2} \sqrt{\frac{p_1 q_1}{n_1} + \frac{p_2 q_2}{n_2}} \quad \text{where } n_1 \text{ and } n_2 \text{ are column (or row) totals for the cells}$$

Create bar chart of cell frequencies by row vs columns; if independent, should see same pattern across the rows. Or, use conditional frequencies by column and compare rows.

If test χ^2 does not exceed χ^2_{α} , do not reject the null, but *do not accept* the hypothesis of independence – risk of a Type II error, no way to calculate β . Cannot conclude that any two particular classifications are independent.

If test rejects the null, do not infer any causal relationships.

Example

Observed:

	Col 1	Col 2	Col 3	Total
Row 1	9	34	53	96
Row 2	16	30	25	71
Total	25	64	78	167

Expected:

	Col 1	Col 2	Col 3	Total
Row 1	14.4	36.8	44.8	
Row 2	10.6	27.2	33.2	
Total				167

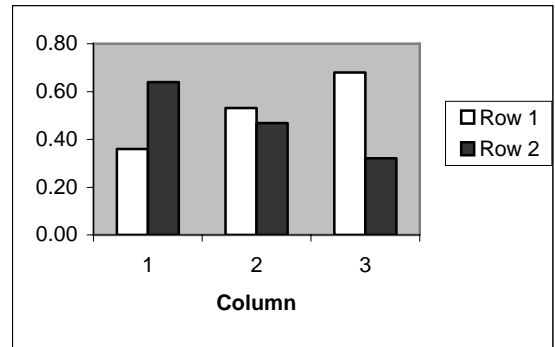
$$E_{ij} = \frac{R_i C_j}{n}$$

$$df = (1)(2) = 2 \quad \alpha = .01 \quad \chi^2_{\alpha} = 9.21034$$

$$\chi^2 = 8.71 \rightarrow \text{do not reject the null that the row and column categories are independent}$$

Construct a bar chart by converting cell frequencies to conditional percentages by column:

	Col 1	Col 2	Col 3
Row 1	0.36	0.53	0.68
Row 2	0.64	0.47	0.32



Do the rows and columns look independent? No!
Of course, if we had tested at the .05 level, we would have rejected the null, so it becomes a matter of desired confidence level (or further investigation).

Computer programs calculate estimated p-values in chi-squared tests. For 2 X 2 tables, you can compute an exact p-value using:

Fisher's Exact Test

Given a table:

	Col 1	Col 2	Total
Row 1	W	X	R ₁
Row 2	Y	Z	R ₂
Total	C ₁	C ₂	T

The probability of this exact configuration, given the observed marginal frequencies, is:

$$p = \frac{\binom{R_1}{W} \binom{R_2}{Y}}{\binom{T}{C_1}}$$

In the exact test, also compute the probabilities of all the more extreme configurations in the same direction (one-tail) or in both directions (two-tail) and add p-values for a total p score. The more extreme cases have p-values less than that of the original matrix.

Example Given:

An Excel chi-squared test gives a p-value of .03570.

	Col 1	Col 2	Total
Row 1	22	9	31
Row 2	2	5	7
Total	24	14	38

For Fisher's exact p-value:

1) Compute the value for this specific table:

$$p = \frac{\binom{31}{22} \binom{7}{2}}{\binom{38}{24}} = .04378$$

2) Looking at column 1, more extreme configurations in the same direction (*keep same marginal frequencies*) would be:

	Col 1	Col 2	Total
Row 1	23	8	31
Row 2	1	6	7
Total	24	14	38

p = .00571

and

	Col 1	Col 2	Total
Row 1	24	7	31
Row 2	0	7	7
Total	24	14	38

p = .00027

Exact p for the test is $.04378 + .00571 + .00027 = .04976$.