CATEGORICAL DATA AND CHI SQUARE TESTS

TOOLBOX

MACQUARIE University
SYDNEY·AUSTRALIA

What is a categorical variable?

Observations are classified into categories which are descriptive, not numerical

Frequency Table

A Single Categorical Variable

For example: Outcomes of 60 rolls of a six-sided die.

Category	1	2	3	4	5	6	Total
Frequency	8	7	12	13	5	15	60

Two Categorical Variables

For example: Netflix survey

	Number of children in the household				
Netflix?	0	1	2	3 or more	Total
Yes	48	37	86	36	207
No	72	53	54	14	193
Total	120	90	140	50	400

Expected Frequencies

The expected frequencies for independence are:

$$E_{IJ} = E(A = I, B = J) = \frac{ROW\_I\_TOTAL \times COL\_J\_TOTAL}{GRANDTOTAL}$$

Chi Squares Goodness of Fit Test

Hypothesis - Goodness of Fit

Null Hypothesis (H_0):
The observed data distribution is consistent with the expected distribution.

Alternative Hypothesis (H_a):
The observed data distribution is not consistent with the expected distribution.

Test Statistic

$$T = CHI^2 = \sum_{i=1}^{N} \frac{(OF_i - EF_i)^2}{EF_i}$$

Where:
N is the number of categories
OFi is the observed frequency of category i
EFi is the “expected” frequency of category i (meaning the frequencies predicted by our hypothesis).

The test statistic follows Chi-Square distribution with K = N – 1 degree of freedom. CV is defined so that Pr(T>CV) = α

Then compare the numerical value of this test statistic T with the CV from the CHISQ distribution.

The hypothesis used to generate the expected frequencies is called the **“null hypothesis”**

The probability value $\alpha\alpha$used to compute the critical value is called the **“level of significance”**

The critical value is computed as the 100 × (1 – α)% percentile of the CHISQ distribution with k=N-1 df

If T>CV we take this as evidence that the hypothesis used to generate the expected frequencies is doubtful and the probability that we would get this value of the test statistic from the data using that hypothesis is low, eg. less than $\alpha\alpha$ = 5%. - Reject the hypothesis

Chi Squares Test for Independence

Hypothesis - Test for Independence

Null Hypothesis (H_0):
The two categorical variables are not related or are independent.

Alternative Hypothesis (H_a):
The two categorical variables are related or dependent.

Contingency tables for testing the hypothesis of independence

Generalise the previous Chi-square technique to the case where two variables are involved

		Variable B		
		Level 1	Level 2	Row Total
Variable A	Level 1	F1,1	F1,2	F1,1 + F1,2
	Level 2	F2,1	F2,2	F2,1 + F2,2
	Column Total	F1,1 + F2,1	F1,2 + F2,2	Grandtotal
	Grandtotal = F1,1 + F1,2, F2,1 + F2,2			

Test Statistics for Test for Independence

The test statistic for testing independence is:

$$T = CHI^2 = \sum_{ALL\ I,J} \frac{(O_{I,J} - E_{I,J})^2}{E_{I,J}}$$

OI,J are observed frequencies and EI,J are expected frequencies.

- If the observed frequencies match the expected ones, then TS = 0.
- If TS is “large”, this is the evidence against the hypothesis of independence
- The TS has the CHISQ distribution with DF = (R-1)(C-1)

CHISQ.DIST.RT(TS,DF)

p-value = 5.78%

If α = 5%, there is insufficient evidence to conclude that sales performance depends on grad status.

We do not reject the Null Hypothesis.

Examples

OBSERVED FREQUENCIES

		TYPE OF EMPOYEE			
		graduate	non graduate	row total	PROPORTION
sales performance	above target	180	320	500	55.55%
	below target	120	280	400	44.44%
	column total	300	600	900	
	PROPORTION	33.33%	66.67%		

EXPECTED FREQUENCIES

		TYPE OF EMPOYEE		
		graduate	non graduate	row total
sales performance	above target	166.67	333.33	500
	below target	133.33	266.67	400
	column total	300	600	900

$$TS = \sum (F_{IJ} - E_{IJ})^2 / E_{IJ} = \frac{(180-166.67)^2}{166.67} + \frac{(120-133.33)^2}{133.33} + \frac{(320-333.33)^2}{333.33} + \frac{(280-266.67)^2}{266.67} = 3.6$$

OBSERVED FREQUENCIES

		TYPE OF EMPOYEE			
		graduate	non graduate	row total	PROPORTION
sales performance	above target	180	320	500	55.55%
	below target	120	280	400	44.44%
	column total	300	600	900	
	PROPORTION	33.33%	66.67%		

EXPECTED FREQUENCIES

		TYPE OF EMPOYEE		
		graduate	non graduate	row total
sales performance	above target	166.67	333.33	500
	below target	133.33	266.67	400
	column total	300	600	900

$$TS = \sum (F_{IJ} - E_{IJ})^2 / E_{IJ} = 3.6, DF = (2-1) \times (2-1) = 1$$

This material is provided to you as a Macquarie University student for your individual research and study purposes only. You cannot share this material publicly online without permission. Macquarie University is the copyright owner of (or has a license to use) the intellectual property in this material. Legal and /or disciplinary actions may be taken if this material is shared without the university's written permission.