Statistics & Explanatory Data Analysis

Tests for nominal data

dr Marcin Chlebus, dr Ewa Cukrowska - Torzewska

Uniwersytet Warszawski
Wydział Nauk Ekonomicznych

Tests for Nominal variables

- The tests for nominal variables:
  - is there an association between two nominal variables?
  - counts of observations for a nominal variable match a theoretical set of proportions for a variable?
- Nominal data is often presented in contingency tables

Hometown / Place of living	Rural area	Small Cities	Big Cities
Rural area	12	5	2
Small Cities	14	18	8
Big Cities	24	27	40

X_2 / X_1	$X_2 = 0$	$X_2 = 1$	Row totals
$X_1 = 0$	n_{11}	n_{12}	$n_{1.}$
$X_1 = 1$	n_{21}	n_{22}	$n_{2.}$
Col totals	$n_{.1}$	$n_{.2}$	n

Goodness-of-Fit Tests – for nominal data

DATA:
- A nominal variable with two or more levels
- Theoretical, typical, expected, or neutral values for the proportions for this variable are needed for comparison
- G-test and chi-square test may not be appropriate if there are cells with low counts (more then 5 observations for each cel)

HYPOTHESIS:
- H0: The proportions for the levels for the nominal variable are not different from the expected proportions
- H1 (2-sided): The proportions for the levels for the nominal variable are different from the expected proportions

Exact binomial test

$$P(x = X; n, p) = \frac{n!}{(n-X)!\,X!} p^X (1-p)^X$$

Exact multinomial test

$$P(x_1 = X_1, \dots, x_k = X_k; n, p_1, \dots, p_k) = n! \prod_{i=1}^{k} \frac{\pi_i^{x_i}}{x_i!}$$

$$p - value = \sum_{y: P(x=y) \le P(x=X)} P(y)$$

Pearson Chi^2 test

$$X^2 = \sum_{i=1}^{n} \frac{(O_i - E_i)^2}{E_i} \sim \chi^2(n - p(df))$$

G test

$$G = 2 \sum_{i=1}^{n} O_i ln\left(\frac{O_i}{E_i}\right) \sim \chi^2(n - p(df))$$

p is a number of parameters to be fitted + 1

Association Tests – for nominal data

DATA:
- Two nominal variables with two or more levels each.
- Experimental units aren't paired.
- There are no structural zeros in the contingency table. (Example: Gender vs. Pregnancy)
- G-test and chi-square test may not be appropriate if there are cells with low counts (n>5)

HYPOTHESIS:
- H0: There is no association between the two variables (they are independent).
- H1 (2-sided): There is an association between the two variables

Fischer Exact test for 2 x 2 table

$$p = \frac{(n_{11}+n_{12})(n_{11}+n_{21})(n_{12}+n_{22})(n_{21}+n_{22})}{n_{11}!\,n_{12}!\,n_{21}!\,n_{22}!} \qquad p-value = \sum_{y:P(y)\leq p} P(y)$$

Pearson Chi^2 independence test

$$X^2 = \sum_{i=1}^{r}\sum_{j=1}^{c}\frac{(O_{ij}-E_{ij})^2}{E_{ij}} \sim \chi^2((r-1)(c-1))$$

G independence test

$$G = 2\sum_{i=1}^{r}\sum_{j=1}^{c} O_{ij} ln\left(\frac{O_{ij}}{E_{ij}}\right) \sim \chi^2((r-1)(c-1))$$

Expected values depends on marginal probabilities

$$E_{ij} = Np_{i.}p_{.j}$$

Tests for Paired Nominal Data

DATA:
- Two nominal variables with two or more levels each, and each with the same levels.
- Observations are paired or matched between the two variables.
- McNemar and McNemar–Bowker tests may not be appropriate if discordant cells have low counts.

HYPOTHESIS:
- H0: The contingency table is symmetric. That is, the probability of cell [*i*, *j*] is equal to the probability of cell [*j*, *i*].
- H1 (two-sided): The contingency table is not symmetric
- In practice: A become more popular than B

For multiple times or groups, Cochran's Q test can be used.

McNemar's symmetry test

$$X^2=\frac{(n_{12}-n_{21})^2}{n_{12}+n_{21}}\sim\chi^2(1)$$

Exact symmetry test is also av ailable

There are a few alternatives:
1. Exact symmetry test
2. Stuart-Maxwell
3. ...

McNemar–Bowker' symmetry test

$$X^2=\sum_{i<j}\frac{(n_{ij}-n_{ji})^2}{n_{ij}+n_{ji}}\sim\chi^2\left(\frac{n(n-1)}{2}-R(off-diagonal\ nonzero\ cells)\right)$$

Difference between tests of symmetry & association for not paired data

Tea / $Coffee$	Yes	No	Row totals
Yes	37	17	54
No	9	25	34
Col totals	46	42	88

Pearson Chi^2 independence test

HYPOTHESIS:
- H0: There is no association between the two variables (they are independent)

P-value: p-value = 0.0002878
Conclusion: Drinking coffee is positively associated with drinking tea
Intuition: if n11+n22 is significantly bigger than n12+n21 - positive association (otherwise negative)

McNemar's symmetry test

HYPOTHESIS:
- H0: The contingency table is symmetric. That is, the probability of cell [*i*, *j*] is equal to the probability of cell [*j*, *i*].

P-value: p-value = 0.1698
Conclusion: Coffee is as popular as Tea
Intuition: If n12 is significantly higher than n21 – Coffee is more popular than Tea

Cochran–Mantel–Haenszel Test for 3-Dimensional Tables

DATA:
- Three nominal variables with two or more levels each.
- Data can be stratified as *n* x *n* tables with the third time or grouping variable

HYPOTHESIS:
- H0: There is no association between the two inner variables (OR=1).
- H1: There is an association between the two inner variables (OR<>1).

"There was a significant association between variable A and variable B [across groups/times]."

$$OR = \frac{\sum_{i=1}^{K} \frac{n_{11,i} n_{22,i}}{n_i}}{\sum_{i=1}^{K} \frac{n_{12,i} n_{21,i}}{n_i}} \qquad OR = \frac{P(Y=1|X=1) \Big/ P(Y=0|X=1)}{P(Y=1|X=0) \Big/ P(Y=0|X=0)}$$

K – number of groups/repeats

Cochran–Mantel–Haenszel Test

$$\xi_{CMH} = \frac{\left[\sum_{i=1}^{K} \left(n_{11,i} - \frac{n_{1.,i} n_{.1,i}}{n_i}\right)\right]^2}{\sum_{i=1}^{K} \frac{n_{1.,i} n_{2.,i} n_{.1,i} n_{.2,i}}{{n_i}^2 (n_i - 1)}} \sim \chi^2(1)$$

The CMH test supposes that the effect of the treatment is homogeneous in all groups/times.

Tests for homogenous association:
1. Breslow-Day test
2. Wolff test

Cochran's Q Test for Paired Nominal Data

DATA:
- The response variable is a dichotomous nominal variable.
- The responses are paired by experimental unit.
- The responses are measured across two or more times or factors.
- The data follow an unreplicated complete block design, with each experimental unit treated as the block.

HYPOTHESIS:
- H0: The marginal probability of a [positive] response is unchanged across the times or factors. That is, a positive response is equally likely across times or factors.
- H1: Positive responses are not equally likely across times or factors.

There was a significant difference in the proportion of positive responses across times or factors.

Cochran's Q Test

$$T = K(K-1)\frac{\sum_{i=1}^{K}\left(n_{.j} - \frac{N}{K}\right)^2}{\sum_{i=1}^{b} n_{i.}(K - n_{i.})} \sim \chi^2(K-1)$$

K – number of times or factors
b – numer of experimetal units
$n_{.j}$ - column total for jth factor or time
$n_{i.}$ - row total for ith experimetal units

	F/TP 1	…	F/TP K	
EU 1	1/0	1/0	1/0	$n_{1.}$
EU 2	1/0	1/0	1/0	$n_{2.}$
…	1/0	1/0	1/0	…
EU b	1/0	1/0	1/0	$n_{b.}$
	$n_{.1}$	…	$n_{.K}$	N

For 2 factors – equivalent for McNemar test