Experimental Humanities II (HUMB002) 2016
STATISTICAL ANALYSIS

NONPARAMETRIC TESTS

Lecture 7

Pavla Linhartová

The lectures and exercises are based on the lectures from the subject PSY117 – Statistical analysis by Stanislav Ježek and Jan Širůček from Department of Psychology, Faculty of Social Studies MU Brno

$\chi 2$ goodness of fit test

- Do empirically observed frequencies differ from theoretically expected frequencies?
  - e.g. Political parties preference in elections
  - = one-sample test
- We are testing the probability of the difference between observed (f_o) and expected (f_e) frequencies
- The difference is expressed by value of χ^2 with χ^2 distribution with df=k-1, where k is the number of categories and mean = df
- Excel: CHISQ.DIST(χ^2; df; 1); CHISQ.INV(*p*; df)
- The expected frequencies are theoretically inferred
- f_o and f_e always as relative frequenceis, never as percent

$$\chi^2 = \sum_{i=1}^{k} \frac{(fo_i - fe_i)^2}{fe_i}$$

In which city would you like to live?

Category	fo	p	fe	(fo-fe)^2/fe
Paris	28	0,2	28	0
New York	28	0,2	28	0
London	28	0,2	28	0
L.A.	28	0,2	28	0
Tokio	28	0,2	28	0
Total	140	1	140	0
Chi²				**0**

$$\chi^2 = \sum_{i=1}^{k} \frac{(fo_i - fe_i)^2}{fe_i}$$

In which city would you like to live?

Category	fo	p	fe	(fo-fe)^2/fe
Paris	38	0,2	28	3,57
New York	37	0,2	28	2,89
London	22	0,2	28	1,29
L.A.	25	0,2	28	0,32
Tokio	18	0,2	28	3,57
Total	140	1	140	11,64
Chi²				**11,64**

$$\chi^2 = \sum_{i=1}^{k} \frac{(fo_i - fe_i)^2}{fe_i}$$

P(c2 > 11,64 | c2 = 4)=1-CHISQ.DIST(11,64;4;1)=0,02

Relationship between two categorical variables

- What is the relationship between political parties preferrence and income level?
- Based on contingency table: rows x columns = i x j
- Marginal frequencies: e.g. N_{12} means number of people in the interception of the first row and the second column

Categories	**B_1**	**B_2**	**…**	**B_s**	**Row marginal frequencies**
A_1	n_{11}	n_{12}	…	n_{1s}	**$n_{1.}$**
A_2	n_{21}	n_{22}	…	n_{2s}	**$n_{2.}$**
…	…	…	…	…	**…**
A_r	n_{i1}	n_{i2}	…	n_{ij}	**$n_{i.}$**
Column marginal frequencies	**$n_{.1}$**	**$n_{.2}$**	**…**	**$n_{.j}$**	**n**

Relationship between two categorical variables

- Chi-square independence test
- Observed frequencies = n_{ij}, expected frequencies = m_{ij}
- df=(i-1)*(j-1)

$$f_e = m_{ij} = \frac{n_{i.} n_{.j}}{n} \qquad \chi^2 = \sum_{i=1}^{r} \sum_{j=1}^{s} \frac{(n_{ij} - m_{ij})^2}{m_{ij}} = \sum_{i=1}^{r} \sum_{j=1}^{s} \frac{(fo_{ij} - fe_{ij})^2}{fe_{ij}}$$

Categories	B_1	B_2	…	B_s	**Row marginal frequencies**
A_1	n_{11}	n_{12}	…	n_{1s}	$\mathbf{n_{1.}}$
A_2	n_{21}	n_{22}	…	n_{2s}	$\mathbf{n_{2.}}$
…	…	…	…	…	…
A_r	n_{i1}	n_{i2}	…	n_{ij}	$\mathbf{n_{i.}}$
Column marginal frequencies	$\mathbf{n_{.1}}$	$\mathbf{n_{.2}}$	…	$\mathbf{n_{.j}}$	**n**

Relationship between residence size and number of rubber boots

Observed frequencies Row %	0	1	>2	Row marginal frequencies
Big city	10 67%	1 7%	4 27%	**15**
Small town	15 43%	19 54%	1 3%	**35**
Village	15 30%	20 40%	15 30%	**50**
Column marginal frequencies	**40**	**40**	**20**	**100**

Expected frequencies / cell χ^2	0	1	>2	Row marginal frequencies
Big city	6 / 2,7	6 / 4,2	3 / 0,3	**15**
Small town	14 / 0,1	14 / 1,8	7 / 5,1	**35**
Village	20 / 1,3	20 / 0	10 / 2,5	**50**
Column marginal frequencies	**40**	**40**	**20**	**100**

$$\chi^2 = \sum_{i=1}^{r}\sum_{j=1}^{s}\frac{(n_{ij} - m_{ij})^2}{m_{ij}} = \sum_{i=1}^{r}\sum_{j=1}^{s}\frac{(fo_{ij} - fe_{ij})^2}{fe_{ij}}$$

χ^2=17,9   df=(3-1)*(3-1)=4   P(χ^2 > 17,9 | χ^2 = 4)=0,001

Association strength and assumptions

- Strength of association in contingency table
  - Indexes: Cramer V, phi
  - Standardized residuals: standardized difference between observed and expected frequencies for each contingency table cell
    - $R = (n_{ij} – m_{ij}) / \sqrt{m_{ij}}$
  - Standardized residuals have normal distribution, we consider as significant standard residuals higher than 1.96
- Assumptions
  - Expected frequency in each contingency table cell should be at least 5

Association strength in contingency table

Observed frequencies Row % Expected frequencies Standardized residuals	0	1	>2	Row marginal frequencies
Big city	10 67% 6 **1,6**	1 7% 6 **2,0**	4 27% 3 **0,6**	**15**
Small town	15 43% 14 **0,3**	19 54% 14 **1,3**	1 3% 7 **2,3**	**35**
Village	15 30% 20 **1,1**	20 40% 20 **0**	15 30% 10 **1,6**	**50**
Column marginal frequencies	**40**	**40**	**20**	**100**

Nonparametric ordinal tests

- Alternatives to t-tests
- Robust towards distribution shape
- Differences in medians (mean ranks):
  - One-sample: Wilcoxon test, sign test
  - Independent samples: Mann-Whitney U test (Median test)