12 The relationship between two variables on the nominal scale and the ordinal scale

At the Nominal scale one uses codes assigned to objects as labels; for example, "method of payment" can be generally categorized as (1) payment on account, (2) payment in cash, (3) payment by cheque and it has three categories (levels). For this scale or category some valid operations are equivalence and set membership. This is also called a categorical variable. The first part of this chapter is aimed to perform tests which determine whether two categorical variables are independent or not; consequently we may be interested in the degree of association between considered two variables which can be assessed by some coefficients.

The relationship between ordinal variables may also be represented in contingency tables, though this is less often done since we have more efficient tests for ordinal variables, which are performed in the second part of this chapter. In ordinal scale type, the numbers assigned to objects represent the rank order (1st, 2nd, 3rd etc.) of the entities assessed. An example of ordinal measurement is the results of a horse race, which say only which horses arrived first, second, third, etc. but include no information about times. The central tendency of an ordinal variable can be represented by its mode or its median, but the mean cannot be defined. For this scale the linear order is the valid relation.

The relationship between two nominal variables and related tests

Definition 12.1
Let X, Y be nominal random variables, where X has r categories: $x_{[1]}, \dots, x_{[r]}$ and Y has s categories: $y_{[1]}, \dots, y_{[s]}$. Consider the random sample $\binom{X_1}{Y_1}, \dots, \binom{X_n}{Y_n}$ from the distribution of $\binom{X}{Y}$. Let us denote as n_{jk} the joint frequency of the pair of categories $(x_{[j]}, y_{[k]})$. The table of joint frequencies n_{jk}, $j = 1, \dots, r;\ k = 1, \dots, s$ is called *contingency table.*

	$y_{[k]}$	$y_{[1]}$	$\dots$	$y_{[s]}$	$n_{j\cdot}$
$x_{[j]}$	n_{jk}				
$x_{[1]}$		n_{11}	$\dots$	n_{1s}	$n_{1\cdot}$
$\vdots$		$\vdots$		$\vdots$	$\vdots$
$x_{[r]}$		n_{r1}	$\dots$	n_{rs}	$n_{r\cdot}$
$n_{\cdot k}$		$n_{\cdot 1}$	$\dots$	$n_{\cdot s}$	n

The frequencies $n_{j\cdot} = \sum\limits_{k=1}^{s} n_{jk}$, $n_{\cdot k} = \sum\limits_{j=1}^{r} n_{jk}$ are called marginal frequencies.

Theorem 12.2
Let us consider test: H_0 : Variables X, Y are independent against H_1 : Variables X, Y are not independent. If H_0 is true then the test statistic

$$K = \sum_{j=1}^{r} \sum_{k=1}^{s} \frac{\left(n_{jk} - \frac{n_{j\cdot} n_{\cdot k}}{n}\right)^2}{\frac{n_{j\cdot} n_{\cdot k}}{n}} \approx \chi^2((r-1)(s-1)).$$

[K is said to follow asymptotic χ^2 distribution with $(r-1)(s-1)$ degrees of freedom.]
At the asymptotic significance level α the null hypothesis of independence is rejected in favor of alternative hypothesis H_1, if the realization of the test statistic satisfies the condition $K > \chi^2_{1-\alpha}((r-1)(s-1))$. Thus the critical region $W = \langle \chi^2_{1-\alpha}((r-1)(s-1)), \infty)$.

Remark 12.3

The statistic K is distributed approximately as χ^2 with $(r-1)(s-1)$ degrees of freedom provided that for the expression $\frac{n_{j\cdot}n_{\cdot k}}{n}$ it holds:
at least in 80% of cases $\frac{n_{j\cdot}n_{\cdot k}}{n} \geq 5$
at most in 20% of cases $\frac{n_{j\cdot}n_{\cdot k}}{n} \geq 2$.
If not then the pooling of appropriate categories (to attain large expected frequencies $\frac{n_{j\cdot}n_{\cdot k}}{n}$) is recommended.

Remark 12.4

Let $p_{jk} = P(X = x_{[j]} \wedge Y = y_{[k]}) \quad p_{j\cdot} = \sum_{k=1}^{s} p_{jk} \quad p_{\cdot k} = \sum_{j=1}^{r} p_{jk}$

The variables X and Y are independent if and only if the multiplicative relationship $p_{jk} = p_{j\cdot} \cdot p_{\cdot k}$ is true. The test statistic of independence of X and Y follows the idea that the difference between the joint frequencies and the expected joint frequencies, if independence is really true, should be "very small". As the marginal distributions $p_{j\cdot}$, $p_{\cdot k}$ are usually unknown, we estimate them through marginal frequencies: $\hat{p}_{j\cdot} = \frac{n_{j\cdot}}{n}$ and $\hat{p}_{\cdot k} = \frac{n_{\cdot k}}{n}$. Thus the expected joint frequencies $n \cdot p_{j\cdot} \cdot p_{\cdot k}$ can be estimated by frequencies $n \cdot \frac{n_{j\cdot}}{n} \cdot \frac{n_{\cdot k}}{n} = \frac{n_{j\cdot} \cdot n_{\cdot k}}{n}$.
Great differences between joint frequencies and estimated frequencies, which are expected under independence, bring evidence against the null hypothesis. Thus the critical region is concentrated at upper tail of χ^2 distribution. (That is this is an upper tailed test only.)

To determine the degrees of freedom we must reduce the number $r \cdot s$ of summands in double sum $\sum_{j=1}^{r} \sum_{k=1}^{s}$ with respect to the conditions for both marginal distributions: $\sum_{j=1}^{r} p_{j\cdot} = 1$ and $\sum_{k=1}^{s} p_{\cdot k} = 1$. Thus the first sum has $r-1$ independent summands and the second sum has $s-1$ independent summands. Hence the double sum has $(r-1)\cdot(s-1)$ independent summands. □

Definition 12.5

The degree of association between two nominal random variables X, Y is measured by *Cramer's coefficient*
$V = \sqrt{\frac{K}{n(m-1)}}$, where $m = \min\{r, s\}$. □
Cramer's coefficient is a monotone function of the statistic K. It's values range from 0 (corresponding to no association between the variables) to 1 (complete association).

Example 12.6

A sociological survey processed data about 360 students: the social origin and the type of school were recorded. The results of the survey are as shown in the table below:

	Social origin	I	II	III	IV	$n_{j\cdot}$
Type of schoole	n_{jk}					
university		50	30	10	50	140
polytechnic		30	50	20	10	110
economic		10	20	30	50	110
$n_{\cdot k}$		90	100	60	110	360

At the asymptotic significance level 0,05 carry out the test that the variables type of school and social origin are independent. Then determine the degree of association.

Solution

seminar session

Definition 12.7

The special case of 2×2 contingency table is called *fourfold table*; thus $r = s = 2$ and the joint frequencies are commonly denoted as follows: $n_{11} = a; n_{12} = b; n_{21} = c; n_{22} = d$.

	$y_{[k]}$	$y_{[1]}$	$y_{[2]}$	$n_{j\cdot}$
$x_{[j]}$	n_{jk}			
$x_{[1]}$		a	b	$a+b$
$x_{[2]}$		c	d	$c+d$
$n_{\cdot k}$		$a+c$	$b+d$	n

□

There are three available independence tests of the fourfold tables:

1.) Asymptotic χ^2 test.
This test suffers from the disadvantage that if below stated conditions are not satisfied, then the pooling of categories is not possible and the test statistic is not distributed approximately as chi^2.

2.) Asymptotic odds ratio test.
This test is based on the test statistic OR (odds ratio) which can be instrumental to measure the degree of association. It can be used for sufficiently large frequencies.

3.) Fisher's exact test.
If the assumptions of previously mentioned tests do not hold, Fisher's test can be used. Character of this test is discrete.

Theorem 12.8
Testing independence between two nominal variables in fourfold tables the test statistic K from the theorem 12.2 can be rearranged into the form

$$K = \frac{n(ad - bc)^2}{(a+b)(c+d)(a+c)(b+d)}$$

If H_0 is true then $K \approx \chi^2(1)$.

Remark 12.9
The statistic K is distributed approximately as χ^2 with 1 degree of freedom provided that following conditions hold: $a + b > 5$; $c + d > \frac{a+c}{3}$.

Example 12.10
Consider 135 applicants for particular university education. Suppose one random variable is the impression upon entrance examination committee and the other random variable is the faculty entrance. At the asymptotic level 0.05 carry out the test that the entrance and the impression are not associated.

	impression	good	bad	$n_{j\cdot}$
entrance	n_{jk}			
yes		17	11	28
no		39	58	97
$n_{\cdot k}$		56	69	125

Solution
seminar session

Remark 12.11
The fourfold tabs can be treated in a different way based on following idea. Particular experiment, which has two outcomes, can be carried in two groups. Thus the appropriate scheme is 2 × 2 contingency table, where X has two categories: success and failure and Y has two categories: group I and group II. (These groups might be men and women, an experimental group and a control group, or any other dichotomous classification.)

	Group	I	II	$n_{j\cdot}$
Outcomes	n_{jk}			
success		a	b	$a+b$
failure		c	d	$c+d$
$n_{\cdot k}$		$a+c$	$b+d$	n

The odds for column I are $\frac{a}{c}$ and for column II are $\frac{b}{d}$. If the "group" does not have an impact on the outcome of an experiment then the ratio of the two odds $\frac{\frac{a}{c}}{\frac{b}{d}}$ is equal to one.

Definition 12.12
Considering the fourfold table the statistic $OR = \frac{\frac{a}{c}}{\frac{b}{d}} = \frac{ad}{bc}$ is called odds ratio.
The constant $o\rho = \frac{p_{11}p_{22}}{p_{12}p_{21}}$ is called theoretic odds ratio.

Remark 12.13
If the variables X, Y are independent then $p_{jk} = p_{j\cdot}p_{\cdot k}$ and theoretic odds ratio $o\rho = 1$. The further from unit $o\rho$ is, the greater the dependence is. Under the condition that values in tab are non-zero, the value of $o\rho$ is within the interval $(0, \infty)$. Thus $o\rho$ values are not distributed symmetrically with respect to one and log odds ratios $\ln o\rho$ and $\ln OR$ are used.

Theorem 12.14
Let us consider the fourfold table for two nominal random variables X, Y.
The statistic $U = \frac{\ln OR - \ln o\rho}{\sqrt{\frac{1}{a}+\frac{1}{b}+\frac{1}{c}+\frac{1}{d}}} \approx N(0,1)$.

At the asymptotic significance level α the null hypothesis
$H_0 :\ \ln o\rho = 0$ [which is equivalent with independence of X, Y] is rejected in favor of alternative hypothesis H_1, if the realization of the test statistic

$U = \frac{\ln OR}{\sqrt{\frac{1}{a}+\frac{1}{b}+\frac{1}{c}+\frac{1}{d}}}$ falls within the critical region W. According to the form of the alternative hypothesis the list of corresponding critical regions follows :

pro oboustr. alt. $H_1 : \ln o\rho \neq 0$ je $W = (-\infty,\ -u_{1-\alpha/2}\rangle \cup \langle u_{1-\alpha/2},\ \infty)$
pro levostr. alt. $H_1 : \ln o\rho < 0$ je $W = (-\infty,\ -u_{1-\alpha}\rangle$
pro pravostr. alt. $H_1 : \ln o\rho > 0$ je $W = \langle u_{1-\alpha},\ \infty)$ □

Notice that $\ln o\rho > 0$ implies that the event is more likely in the first group. $\ln o\rho < 0$ implies that the event is less likely in the first group.

Theorem 12.15
Let us consider the fourfold table for two nominal random variables X, Y.
The asymptotic $100(1-\alpha)\%$ confidence interval for the theoretic odds ratio has the limits :
$d = e^{\ln OR - \sqrt{\frac{1}{a}+\frac{1}{b}+\frac{1}{c}+\frac{1}{d}}\,u_{1-\alpha/2}}$
$h = e^{\ln OR + \sqrt{\frac{1}{a}+\frac{1}{b}+\frac{1}{c}+\frac{1}{d}}\,u_{1-\alpha/2}}$ □

Null hypothesis of independence between X, Y [which is equivalent with $o\rho = 1$] is rejected if the asymptotic confidence interval for the theoretic odds ratio does not cover the value 1.

Example 12.16
Using the data from 12.10 calculate and interpret the odds ratio, construct the asymptotic confidence interval for the theoretic odds ratio and test hypothesis that the faculty entrance and impression upon committee are non-associated.

Solution
seminar session

Remark 12.17
An elaborated description of Fisher's exact test exceeds this course. Just short remark: this test of independence is exact and it can therefore be used regardless of the sample characteristics; it is based on odds ratios and could be one-tailed as well as two-tailed.

The relationship between two ordinal variables and related tests

The version of correlation performed in the 11th chapter applies to those cases where the values of X and of Y are both measured on an equal- interval scale. It is also possible to apply the apparatus of linear correlation to cases where X and Y are measured on a merely ordinal scale. When applied to ordinal data, the measure of correlation is spoken of as the Spearman's rank correlation coefficient. It assesses how well an arbitrary monotonic function could describe the relationship between two variables, without making any other assumptions about the particular nature of the relationship between the variables.

Definition 12.18
Let X, Y be ordinal random variables. Consider the random sample $\binom{X_1}{Y_1}, \dots, \binom{X_n}{Y_n}$ from the continuous distribution of the vector $\binom{X}{Y}$. Let R_i stands for the rank of X_i and Q_i stands for the rank of Y_i; $i = 1, 2, \dots, n$. The statistic

$$r_S = 1 - \frac{6}{n(n^2-1)} \sum_{i=1}^{n} (R_i - Q_i)^2,$$

serves as a measure of the rank-order correlation between X and Y and is called *Spearman's rank correlation coefficient.*

Remark 12.19
The values of Spearman's rank correlation coefficient are from the interval $\langle -1, 1 \rangle$, where +1 corresponds to the perfect positive relationship; -1 to the perfect negative relationship and 0 to no relationship (monotonic). (We are speaking of a positive relationship in case "the more of X, the more of Y"; and in case "the more of X, the less of Y" we are speaking of a negative relationship between the two variables.) r_S is the classic correlation coefficient applied on ranks $R_i,\ Q_i$ instead of original variables $X_i,\ Y_i$, thus from 11.2 follows 12.18. This formula is derived under the assumption of the continuous distribution of the vector $\binom{X}{Y}$ that is only rankings without ties may occur. And finally if the assumption of bivariate normality is not met in tests from the 11th chapter, Spearman's rank correlation coefficient may be used.

Theorem 12.20
Let X, Y be ordinal random variables. Consider the random sample $\binom{X_1}{Y_1}, \dots, \binom{X_n}{Y_n}$ from the continuous distribution of the vector $\binom{X}{Y}$ At the significance level α the null hypothesis H_0 : "There is no relationship between X and Y " is rejected in favor of the alternative hypothesis H_1 if the realization

of the test statistic Spearman's rank correlation coefficient r_S falls within the critical region W. According to the form of the alternative hypothesis the list of corresponding critical regions follows:

two-tailed test	H_0 : There exist relationship betw. X and Y.	$W = \langle -1,\ -r_{S,1-\alpha}(n)\rangle \bigcup \langle r_{S,1-\alpha}(n),\ 1\rangle$
left-tailed test	H_1 : The relationship betwen X, Y is negative.	$W = \langle -1,\ -r_{S,1-2\alpha}(n)\rangle$
right-tailed test	H_1 : The relationship betwen X, Y is positive.	$W = \langle r_{S,1-2\alpha}(n),\ 1\rangle$

where $r_{S,1-\alpha}(n)$ is tabulated critical value for given α and usually $n = 5, 6, \ldots, 30$. (For larger size of random sample there are asymptotic statistics.)

Theorem 12.21

Let the assumptions and formulation of the null hypothesis from 12.20 hold. Further let $n > 30$ and H_0 is true. Then the test statistic U_0 follows the standard normal distribution

$$U_0 = r_S\sqrt{n-1} \approx N(0,1)$$

and the critical region has the form $W = (-\infty,\ -u_{1-\alpha/2}\rangle \bigcup \langle u_{1-\alpha/2},\ \infty)$. Hypothesis about no relationship between X, Y is rejected in two-tailed test if the realization of $U_0 \in W$.

Example 12.22

Conditions of seven patients after particular surgery were assessed by two physicians. The highest score obtained that patient, whose condition was most serious.

patient's index	1	2	3	4	5	6	7
The 1st physician's assessment	4	1	6	5	3	2	7
The 2nd physician's assessment	4	2	5	6	1	3	7

Calculate the Spearman's rank correlation coefficient r_S and at the confidence level 0.05 carry out the test that there is no relationship between considered assessments.

Solution

seminar session