

# Kontingenční tabulky, korelační koeficienty

## Ekonometrie

Jiří Neubauer

Katedra kvantitativních metod FVL UO Brno  
kancelář 69a, tel. 973 442029  
email: Jiri.Neubauer@unob.cz

## $\chi^2$ -test nezávislosti v kontingenční tabulce

Budeme předpokládat, že  $X$  a  $Y$  jsou kvalitativní náhodné veličiny, obor hodnot  $X$  obsahuje  $r$  hodnot (kategorií, které budou kódovány čísly  $1, 2, \dots, r$ ) a podobně obor hodnot  $Y$  obsahuje  $s$  hodnot (kategorií, které budou kódovány čísly  $1, 2, \dots, s$ ). Sdružená pravděpodobnostní funkce náhodných veličin  $X$  a  $Y$  je popsána vztahem

$$p(j, k) = P(X = j, Y = k), \quad j = 1, 2, \dots, r, \quad k = 1, 2, \dots, s.$$

Odpovídající marginální pravděpodobnostní funkce náhodné veličiny  $X$  je dána vztahem

$$p_X(j) = P(X = j) = \sum_{k=1}^s p(j, k), \quad j = 1, 2, \dots, r$$

a pravděpodobnostní funkce náhodné veličiny  $Y$  vztahem

$$p_Y(k) = P(Y = k) = \sum_{j=1}^r p(j, k), \quad k = 1, 2, \dots, s.$$

Hodnoty pravděpodobnostní funkce lze uspořádat do tabulky

$\chi^2$ -test nezávislosti

$X \setminus Y$	1	...	$k$	...	$s$	$\sum$	$X \setminus Y$	1	...	$k$	...	$s$	$\sum$
1	$p(1,1)$	...	$p(1,k)$	...	$p(1,s)$	$p_X(1)$	1	$n_{11}$	...	$n_{1k}$	...	$n_{1s}$	$n_{1.}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
$j$	$p(j,1)$	...	$p(j,k)$	...	$p(j,s)$	$p_X(j)$	$j$	$n_{j1}$	...	$n_{jk}$	...	$n_{js}$	$n_{j.}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
$r$	$p(r,1)$	...	$p(r,k)$	...	$p(r,s)$	$p_X(r)$	$r$	$n_{r1}$	...	$n_{rk}$	...	$n_{rs}$	$n_{r.}$
$\sum$	$p_Y(1)$	...	$p_Y(k)$	...	$p_Y(s)$	1	$\sum$	$n_{.1}$	...	$n_{.k}$	...	$n_{.s}$	$n$

**Tabulka:** Tabulka pravděpodobnostní funkce  $p(j, k)$  a kontingenční tabulka pro hodnoty náhodného výběru z diskrétního rozdělení

## $\chi^2$ -test nezávislosti

Pro nezávislé znaky  $X$  a  $Y$  platí

$$p(j, k) = p_X(j) \cdot p_Y(k).$$

Odhad  $p_X(j)$  lze určit jako  $\hat{p}_X(j) = \frac{n_{j.}}{n}$ , podobně odhad  $p_Y(k)$  je dán  $\hat{p}_Y(k) = \frac{n_{.k}}{n}$ , potom odhad  $p(j, k)$  je roven  $\hat{p}(j, k) = \hat{p}_X(j) \cdot \hat{p}_Y(k)$ . Četnosti, které lze očekávat v kontingenční tabulce při nezávislosti proměnných  $X$  a  $Y$ , jsou tvaru

$$o_{jk} = n \cdot \hat{p}(j, k) = n \cdot \hat{p}_X(j) \cdot \hat{p}_Y(k) = n \cdot \frac{n_{j.}}{n} \cdot \frac{n_{.k}}{n} = \frac{n_{j.} \cdot n_{.k}}{n} \quad (1)$$

a nazveme je **očekávané četnosti**. Jsou-li veličiny  $X$  a  $Y$  nezávislé, lze předpokládat, že empirické četnosti  $n_{jk}$  budou blízké očekávaným četnostem  $o_{jk}$ .

## $\chi^2$ -test nezávislosti

K testování nezávislosti náhodných veličin  $X$  a  $Y$  lze použít statistiku

$$\chi^2 = \sum_{j=1}^r \sum_{k=1}^s \frac{(n_{jk} - o_{jk})^2}{o_{jk}}, \quad (2)$$

kteřá má asymptoticky  $\chi^2$  rozdělení s  $\nu = (r - 1)(s - 1)$  stupni volnosti. Hypotézu nezávislosti proměnných  $X$  a  $Y$  pak zamítáme na hladině významnosti  $\alpha$ , když  $\chi^2 \geq \chi_{1-\alpha}^2(\nu)$ , kde  $\chi_{1-\alpha}^2(\nu)$  je 100(1 -  $\alpha$ )% kvantil Pearsonova  $\chi^2$  rozdělení s  $\nu$  stupni volnosti. Test lze použít při dostatečném obsazení všech buněk tabulky, tj. pokud všechny očekávané četnosti jsou dosti velké, obvykle se předpokládá, že  $o_{jk} \geq 5$ . Při nesplnění této podmínky se doporučuje spojování „sousedních“ obměn u jedné nebo druhé veličiny (sloučíme celé řádky nebo sloupce a při opakovaném testu s nimi zacházíme jako s jedinou třídou).

## Příklad

Při sociologickém průzkumu odpovídalo 100 náhodně vybraných osob na určitou otázku. Výsledky jsou v následující tabulce. Rozhodněte, zda odpověď závisí na pohlaví dotazovaných.

Pohlaví	Rozhodně ano	Spíše ano	Nevím	Spíše ne	Rozhodně ne	Celkem
Muž	2	20	10	15	8	55
Žena	4	15	15	8	3	45
Celkem	6	35	25	23	11	100

	$k$					
$j$	1	2	3	4	5	celkem
1	3,30	19,25	13,75	12,65	6,05	55,00
2	2,70	15,75	11,25	10,35	4,95	45,00
celkem	6,00	35,00	25,00	23,00	11,00	100,00

**Tabulka:** Tabulka teoretických četností

## Příklad

Alespoň 80 % těchto teoretických četností by mělo být větší než 5, což v našem případě není splněné (3 hodnoty z 10 jsou menší než 5). Proto je vhodné provést sloučení některých sloupců či řádků, slučování je však třeba provádět „rozumně“, zejména s ohledem na věcný význam spojovaných obměn. Pokud slučování není možné (např. u nás by to byly muži a ženy, nebo rozhodně ano a rozhodně ne), potom v krajním případě ponecháme původní sloupcové i řádkové obměny, ale s vědomím, že takovýto „prohřešek“ snižuje sílu testu. My sloučíme první dva sloupce v původní kontingenční tabulce, které odpovídají pozitivní reakci na danou otázku, a přepočteme příslušné teoretické četnosti.

## Příklad

Pohlaví	Pozitivní reakce	Nevím	Spíše ne	Rozhodně ne	Celkem
Muž	22	10	15	8	55
Žena	19	15	8	3	45
Celkem	41	25	23	11	100

	<i>k</i>				
<i>j</i>	1	2	3	4	celkem
1	22,55	13,75	12,65	6,05	55,00
2	18,45	11,25	10,35	4,95	45,00
celkem	41,00	25,00	23,00	11,00	100,00

**Tabulka:** Tabulka teoretických četností



## Příklad

	<i>k</i>				
<i>j</i>	1	2	3	4	celkem
1	0,013	1,023	0,437	0,629	2,101
2	0,016	1,250	0,534	0,768	2,568
celkem	0,03	2,273	0,97	1,397	4,669

Tabulka: Výpočet testové statistiky

Hodnota testové statistiky je tedy

$$\chi^2 = \sum_{j=1}^r \sum_{k=1}^s \frac{(n_{jk} - o_{jk})^2}{o_{jk}} = 4,66,$$

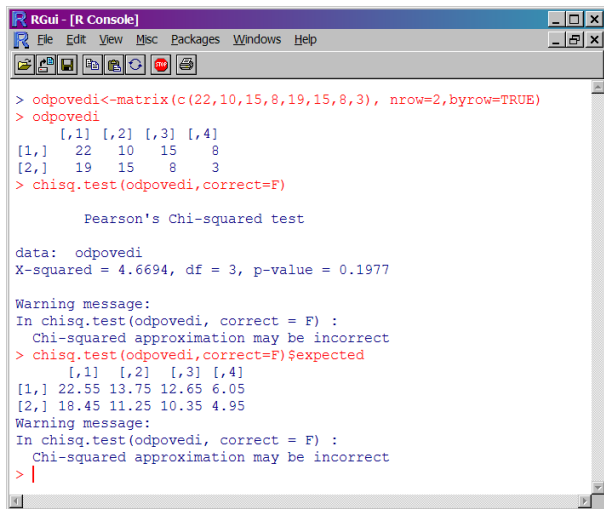
hladinu významnosti použijeme  $\alpha = 0,05$ , stupně volnosti jsou

$\nu = (r - 1)(s - 1) = 3 \cdot 1 = 3$ . Kritický obor je tvořen hodnotami většími než

$\chi_{1-\alpha}^2(3) = 7,815$ . Hodnota testového kritéria nepatří do kritického oboru, tedy se s 95% pravděpodobností neprokázalo, že odpověď na danou otázku závisí na pohlaví.

## $\chi^2$ -test nezávislosti v R

Test nezávislosti v kontingenční tabulce lze v programu R spočítat pomocí funkce `chisq.test`.



```
RGui - [R Console]
File Edit View Misc Packages Windows Help

> odpovedi<-matrix(c(22,10,15,8,19,15,8,3), nrow=2,byrow=TRUE)
> odpovedi
      [,1] [,2] [,3] [,4]
[1,]  22  10  15   8
[2,]  19  15   8   3
> chisq.test(odpovedi,correct=F)

      Pearson's Chi-squared test

data:  odpovedi
X-squared = 4.6694, df = 3, p-value = 0.1977

Warning message:
In chisq.test(odpovedi, correct = F) :
  Chi-squared approximation may be incorrect
> chisq.test(odpovedi,correct=F)$expected
      [,1] [,2] [,3] [,4]
[1,] 22.55 13.75 12.65 6.05
[2,] 18.45 11.25 10.35 4.95
Warning message:
In chisq.test(odpovedi, correct = F) :
  Chi-squared approximation may be incorrect
> |
```

## Koeficienty kontingence

Těsnotnost závislosti dvou nominálních znaků měříme pomocí tzv. **koeficientů kontingence**. Pro hodnocení intenzity závislosti mezi oběma ordinálními resp. nominálními proměnnými existují speciální charakteristiky:

- **Pearsonův koeficient**

$$K_1 = \sqrt{\frac{\chi^2}{n + \chi^2}},$$

- **Cramerův koeficient**

$$K_2 = \sqrt{\frac{\chi^2}{n \cdot \min(r - 1, s - 1)}},$$

- **Čuprovův koeficient**

$$K_3 = \sqrt{\frac{\chi^2}{n \cdot \sqrt{(r - 1)(s - 1)}}}.$$

Poznámka: 0  $\rightarrow$  nezávislost, 1  $\rightarrow$  závislost

## Pearsonův korelační koeficient

Mějme dvourozměrný datový soubor  $\begin{pmatrix} x_1 & y_1 \\ \vdots & \vdots \\ x_n & y_n \end{pmatrix}$ , označme  $\bar{x}$  a  $\bar{y}$  průměry hodnot znaků

a  $s_x$ ,  $s_y$  výběrové směrodatné odchytky hodnot znaků  $X$ ,  $Y$ . **Koeficient korelace (Pearsonův)** definujeme vztahem

$$r_{xy} = \frac{s_{xy}}{s_x s_y},$$

kde  $s_{xy} = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})$  je výběrová kovariance hodnot znaků  $X$  a  $Y$ ,

$s_x = \sqrt{\frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2}$  je výběrová směrodatná odchytky hodnot znaku  $X$  a

$s_y = \sqrt{\frac{1}{n-1} \sum_{i=1}^n (y_i - \bar{y})^2}$  je výběrová směrodatná odchytky hodnot znaku  $Y$ .

## Pearsonův korelační koeficient

Lze jej vyjádřit ve tvaru

$$\begin{aligned} r_{xy} &= \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2} \sqrt{\sum_{i=1}^n (y_i - \bar{y})^2}} = \\ &= \frac{n \sum_{i=1}^n x_i y_i - \sum_{i=1}^n x_i \sum_{i=1}^n y_i}{\sqrt{n \sum_{i=1}^n x_i^2 - (\sum_{i=1}^n x_i)^2} \sqrt{n \sum_{i=1}^n y_i^2 - (\sum_{i=1}^n y_i)^2}}. \end{aligned}$$

## Test významnosti korelačního koeficientu

$$H : \rho = 0 \rightarrow A : \rho \neq 0$$

Testové kritérium je statistika

$$t = \frac{r_{xy}}{\sqrt{1 - r^2}} \sqrt{n - 2} \sim t(n - 2).$$

Kritický obor je dán

$$W_\alpha : |t| \geq t_{1-\alpha/2}(n - 2).$$

Pokud hodnota testového kritéria padne do kritického oboru, podařila se prokázat lineární závislost mezi sledovanými proměnnými.

## Spearmanův korelační koeficient

Pearsonův korelační koeficient nachází uplatnění v případě, kdy sledované znaky jsou spojité a existuje mezi nimi lineární vazba. V případě, že tato vazba lineární není, tento koeficient jako míra těsnosti závislosti selhává. Mimo již zmíněného Pearsonova korelačního koeficientu lze závislost vyjádřit prostřednictvím **Spearmanova korelačního koeficientu**, který popisuje vazbu mezi dvěma spojitými znaky pomocí pořadí. Je vhodný v případě, kdy vztah mezi dvěma proměnnými může být popsán pomocí monotónní funkce.

Hodnotám  $x_i$ ,  $y_i$  přiřadíme čísla  $p_i$ ,  $q_i$  udávající pořadí jednotlivých hodnot při vzestupném uspořádání hodnot znaků podle velikosti. **Spearmanův koeficient pořadové korelace** je potom definován vztahem

$$r_{xy}^S = 1 - \frac{6 \sum_{i=1}^n (p_i - q_i)^2}{n(n^2 - 1)}.$$

## Spearmanův korelační koeficient

Pro náhodný výběr šesti států USA byly zjištěny spotřeby cigaret na hlavu a roční míra úmrtnosti na 100 000 lidí následkem rakoviny plic. Určete, zda existuje významná korelace mezi těmito znaky.

Stát USA	Spotřeba cigaret		Úmrtnost		$(p_i - q_i)^2$
	$x_i$	$p_i$	$y_i$	$q_i$	
Delaware	3400	6	24	5	1
Indiana	2600	4	21	4	0
Iowa	2200	2	17	1	1
Montana	2400	3	19	2	1
New Yersey	2900	5	26	6	1
Washington	2100	1	20	3	4

Suma kvadrátů v posledním sloupci je 8,

$$r_{xy}^S = 1 - \frac{6 \cdot 8}{6 \cdot (6^2 - 1)} = 0,77143.$$

Pozn. Kritická hodnota pro  $\alpha = 0,05$  je 0,829 ( $p$ -hodnota je 0,1028), korelace tedy nebyla prokázána.



## Kendallův korelační koeficient

Mějme dvourozměrný datový soubor. Řekneme, že dvojice  $(x_i, y_i)$  a  $(x_j, y_j)$  jsou ve shodě (concordant), pokud platí, že  $x_i > x_j$  a zároveň  $y_i > y_j$  nebo  $x_i < x_j$  a zároveň  $y_i < y_j$ . Řekneme, že nejsou ve shodě (discordant), pokud  $x_i < x_j$  a zároveň  $y_i > y_j$  nebo  $x_i > x_j$  a zároveň  $y_i < y_j$ . V případě, že  $x_i = x_j$  nebo  $y_i = y_j$  nemluvíme ani o shodě, ani o neshodě. Označme počet dvoji ve shodě  $n_c$  a počet dvojic, které ve shodě nejsou  $n_d$ . **Kendallův korelační koeficient** je definován vztahem

$$\tau = \frac{n_c - n_d}{\frac{1}{2}n(n-1)}.$$

Pro data z předchozího příkladu máme  $n_c = 12$ ,  $n_d = 3$ ,  $n = 6$ .

$$\tau = \frac{12 - 3}{\frac{1}{2} \cdot 6 \cdot (6 - 1)} = 0,6.$$

Pozn. Kritická hodnota pro  $\alpha = 0,05$  je 0,8 ( $p$ -hodnota je 0,1361), korelace tedy nebyla prokázána.

## Koeficient mnohonásobné korelace

Koeficient mnohonásobné korelace vyjadřuje sílu závislosti jedné proměnné na dvou a více jiných proměnných. Mějme  $k$  proměnných  $X_1, X_2, \dots, X_k$ , jejich korelační matice je rovna

$$\mathbf{R} = \begin{pmatrix} 1 & r_{12} & r_{13} & \dots & r_{1k} \\ r_{12} & 1 & r_{23} & \dots & r_{2k} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ r_{1k} & 1 & r_{2k} & \dots & 1 \end{pmatrix}$$

Koeficient mnohonásobné korelace popisující závislost  $X_1$  na  $X_2, \dots, X_k$  se určí ze vztahu

$$R_{1,2,3,\dots,k} = \sqrt{1 - \frac{\det(\mathbf{R})}{\det(\mathbf{R}_{11})}},$$

kde  $\mathbf{R}_{11}$  vznikne z  $\mathbf{R}$  vynechání 1. řádku a 1. sloupce.

## Koeficient mnohonásobné korelace

Koeficient mnohonásobné korelace vyjadřuje společné působení nezávisle proměnných  $X_1, X_2, \dots, X_k$  na závisle proměnnou  $Y$  a určuje spolehlivost regresního odhadu. Výběrový koeficient mnohonásobné korelace pro případ regrese se dvěma nezávisle proměnnými ( $Y_i = \beta_0 + \beta_1 x_i + \beta_2 z_i + \epsilon_i$ ) je roven

$$r_{y,xz} = \sqrt{\frac{r_{yx}^2 + r_{yz}^2 + 2r_{yx}r_{yz}r_{xz}}{1 - r_{xz}^2}},$$

kde  $r_{yx}$  je výběrový korelační koeficient mezi hodnotami  $y_i$  a  $x_i$ ,  $r_{yz}$  je výběrový korelační koeficient mezi  $y_i$  a  $z_i$  a  $r_{xz}$  je výběrový korelační koeficient mezi  $x_i$  a  $z_i$ . Jeho druhou mocninou je index determinace.