Testování nezávislosti

Má dvojí význam: a) věcný – chceme např. zjistit, zda výdaje domácnosti za kulturu závisí či nezávisí na výši jejich příjmů

b) teoretický – využijeme nezávislost, abychom věděli, zda můžeme či nemůžeme k porovnání výběrů použít testy, pro které je nezávislost předpokladem

Zatím nás nebude zajímat konkrétní tvar závislosti, tj. jestli hodnoty jedné sledované veličiny rostou v závislosti na druhé lineárně, kvadraticky atd., ale pouze, zda mezi nimi nějaká vazba je či není.

V teorii pravděpodobnosti se nezávislost definuje následovně:

Náhodné veličiny $X_1, \dots, X_n$ jsou nezávislé, platí-li pro všechna reálná $x_1, \dots, x_n$ vztah

$$F(x_1, \dots, x_n) = F_1(x_1) F_2(x_2) \dots F_n(x_n)$$

resp. $$f(x_1, \dots, x_n) = f_1(x_1) f_2(x_2) \dots f_n(x_n).$$

Zde $F(x_1, \dots, x_n) = P(X_1 < x_1, X_2 < x_2, \dots, X_n < x_n)$ je sdružená distribuční fce,

$F_i(x_i) = P(X_i < x_i)$ jsou marginální distr. fce (podobně pro hustoty).

Pro nezávislé náhodné veličiny platí

$$E(X_1 \cdot X_2 \cdot \ldots \cdot X_n) = EX_1 \cdot EX_2 \dots EX_n$$

$$\operatorname{var}(X_1 + X_2 + \dots + X_n) = \operatorname{var} X_1 + \operatorname{var} X_2 + \dots + \operatorname{var} X_n$$

$$\operatorname{cov}(X_i, X_j) = E(X_i - EX_i)(X_j - EX_j) = 0$$

Mírou závislosti dvou veličin X a Y je pak kovariance nebo ještě lépe korelace $$\rho_{X,Y} = \frac{\operatorname{cov}(X,Y)}{\sqrt{\operatorname{var} X \operatorname{var} Y}}.$$

Pro nezávislé náhodné veličiny je $\rho = 0$. Je-li $\rho \neq 0$, jsou veličiny zcela jistě závislé.

Nejvyšší možná závislost je $\rho_{XX} = \dfrac{\operatorname{cov}(X,X)}{\operatorname{var} X} = 1$.

Pozor! Platí-li $\rho_{X,Y} = 0$, pak veličiny X a Y ještě nemusí být nezávislé (ledaže by měly normální rozdělení). V tom případě jsou pouze nekorelované, což je slabší vlastnost než nezávislost.

Vlastnosti korelačního koeficientu:

a) $-1 \leq \rho_{X,Y} \leq 1$

b) $\rho_{X,Y} = 1$ právě když $Y = aX + b$, kde $a > 0$

c) $\rho_{X,Y} = -1$ —"— $Y = aX + b$, kde $a < 0$

Kovarianční matice náhodného vektoru $\mathbb{X} = (X_1, \ldots, X_n)'$ je matice

$$\mathbb{V} = E(\mathbb{X} - E\mathbb{X})(\mathbb{X} - E\mathbb{X})' = \left(\operatorname{cov}(X_i, X_j)\right)_{i,j=1}^{n}$$

Také se jí říká varianční matice, protože na diagonále má rozptyly veličin $X_1, \ldots, X_n$. Je pozitivně semidefinitní a symetrická.

Korelační matice náhodného vektoru $\mathbb{X}$ je matice $\mathbb{P} = \left(\rho_{X_i, X_j}\right)_{i,j=1}^{n}$.
Je symetrická a na diagonále má jedničky.

Analogicky se definují kovarianční a korelační matice pro dva náh. vektory $\mathbb{X}, \mathbb{Y}$.

$$\operatorname{cov}(\mathbb{X}, \mathbb{Y}) = E(\mathbb{X} - E\mathbb{X})(\mathbb{Y} - E\mathbb{Y})' = \left(\operatorname{cov}(X_i, Y_j)\right)_{i,j=1}^{m,n}$$

$$\operatorname{cor}(\mathbb{X}, \mathbb{Y}) = \left(\rho_{X_i, Y_j}\right)_{i,j=1}^{m,n}$$

Empirickým protějškem korelačního koeficientu je výběrový korelační koeficient.

$$r = \frac{S_{X,Y}}{\sqrt{S_X^2 \, S_Y^2}}$$

kde $S_{X,Y} = \frac{1}{n-1} \sum_i (X_i - \bar{X})(Y_i - \bar{Y})$ je výběrová kovariance

$S_X^2 = \frac{1}{n-1} \sum_i (X_i - \bar{X})^2$, $S_Y^2 = \frac{1}{n-1} \sum_i (Y_i - \bar{Y})^2$ jsou výběrové rozptyly

pro výběry $X_1, \ldots, X_n$ a $Y_1, \ldots, Y_n$.

Platí, že $ES_{XY} = \mathrm{cov}(X,Y)$, tedy S_{XY} je nestranným odhadem kovariance, ale r není nestranným odhadem korelačního koeficientu.

Dále se dá spočítat, že

$$r = \frac{\sum_i X_i Y_i - n\bar{X}\bar{Y}}{\sqrt{\left(\sum_i X_i^2 - n\bar{X}^2\right)\left(\sum_i Y_i^2 - n\bar{Y}\right)}}$$

a že $-1 \le r \le 1$.

Test nezávislosti dvou výběrů se provádí jako test hypotézy $H_0: \rho = 0$ proti alternativě $A: \rho \neq 0$. Hypotéza se zamítá, když $|r| \geq r_n(\alpha)$, kde $r_n(\alpha)$ je kritická hodnota, která bývá tabelovaná.

Nejsou-li v tabulkách přímo kritické hodnoty pro korelační koeficient, dá se využít toho, že za platnosti hypotézy $\rho = 0$ má veličina

$T = \frac{r}{\sqrt{1-r^2}}\sqrt{n-2}$ rozdělení t_{n-2} a zamítnout H_0, když $|T| \geq t_{n-2}(\alpha)$.

Pro oba výsledky je ale podstatný předpoklad, že výběr $\binom{X_1}{Y_1}, \dots, \binom{X_n}{Y_n}$ pochází z dvourozměrného <u>normálního rozdělení</u>. %

Dá se testovat i obecnější hypotéza $H_0: \rho = \rho_0$, kde $\rho_0 \neq \pm 1$, proti $A: \rho \neq \rho_0$, ale je to složitější. Nejprve se vypočte

$$Z = \frac{1}{2}\ln\frac{1+r}{1-r}, \quad Z_0 = \frac{1}{2}\ln\frac{1+\rho_0}{1-\rho_0}.$$

Za platnosti H_0 má veličina $U = \sqrt{n-3}\,(Z - Z_0)$ přibližně rozdělení $N(0,1)$.

H_0 se pak zamítá, když $|U| \geq u\left(\frac{\alpha}{2}\right)$.

Vícerozměrné normální rozdělení: $\boldsymbol{X} = (X_1, \dots, X_n)'$

$$f(\boldsymbol{x}) = \frac{1}{(2\pi)^{n/2} |\mathbb{V}|^{1/2}} \exp\left\{ -\frac{1}{2} (\boldsymbol{x}-\boldsymbol{\mu})' \mathbb{V}^{-1} (\boldsymbol{x}-\boldsymbol{\mu}) \right\}$$

dvourozměrné normální rozdělení:

$$f(x_1, x_2) = \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}} \exp\left\{ -\frac{1}{2(1-\rho^2)} \left(\frac{(x_1-\mu_1)^2}{\sigma_1^2} - 2\rho \frac{(x_1-\mu_1)(x_2-\mu_2)}{\sigma_1\sigma_2} + \frac{(x_2-\mu_2)^2}{\sigma_2^2} \right) \right\}$$

je-li $\rho = 0$, pak $f(x_1, x_2) = f(x_1) f(x_2)$

tj. X_1 a X_2 jsou nezávislé, $X_1 \sim N(\mu_1, \sigma_1^2)$, $X_2 \sim N(\mu_2, \sigma_2^2)$

12649 Svépomoc II/91-100 – TZ 13-322-91-G

Často je třeba sledovat závislost náhodné veličiny Y ne na jedné, ale na více náhodných veličinách $X_1, \dots, X_p$ současně. Teoretickou mírou této závislosti je koeficient mnohonásobné korelace mezi Y a náhodným vektorem $\mathbb{X}$ definovaný jako

$$\rho^2_{Y,\mathbb{X}} = \mathrm{cor}(Y,\mathbb{X})\,(\mathrm{cor}\,\mathbb{X})^{-1}\,\mathrm{cor}(\mathbb{X},Y)$$

(ona to vlastně není definice, ale vztah, který platí a ze kterého se počítá).

Pokud speciálně vektor $\mathbb{X}$ má jen dvě složky X_1, X_2 a označíme-li $Y = X_0$, pak

$$\rho^2_{0,12} = \frac{\rho^2_{01} + \rho^2_{02} - 2\rho_{01}\rho_{02}\rho_{12}}{1-\rho^2_{12}}$$

kde $\rho^2_{0,12} = \rho^2_{Y,(X_1,X_2)}$ atd.

Empirickým protějškem je výběrový koeficient mnohonásobné korelace

$$r^2_{Y,\mathbb{X}} = \mathbb{R}_{Y\mathbb{X}}\,\mathbb{R}^{-1}_{\mathbb{X}\mathbb{X}}\,\mathbb{R}_{\mathbb{X}Y},$$

kde $\mathbb{R}_{\mathbb{X}\mathbb{X}} = (r_{ij})^p_{i,j=1}$, $\mathbb{R}_{Y\mathbb{X}} = (r_{0i})^p_{i=1}$, $\mathbb{R}_{\mathbb{X}Y} = \mathbb{R}'_{Y\mathbb{X}}$ jsou matice výběrových korelačních koeficientů mezi složkami vektoru $\mathbb{X}$, resp. mezi Y a složkami vektoru $\mathbb{X}$.

Opět pro $\mathbb{X} = (X_1, X_2)$ dostáváme $r^2_{0,12} = \dfrac{r^2_{01} + r^2_{02} - 2r_{01}r_{02}r_{12}}{1-r^2_{12}}$

Test hypotézy $H_0: r_{Y,\mathbb{X}} = 0$ je založen na veličině

$$Z = \frac{n-p-1}{p}\cdot\frac{r^2_{Y,\mathbb{X}}}{1-r^2_{Y,\mathbb{X}}},$$

která má za předpokladu normality a za platnosti H_0 rozdělení $F_{p,n-p-1}$.

Význam tohoto koeficientu. Měří celkovou závislost veličiny Y na celém vektoru $\mathbb{X}$, která může být velká, i když závislost Y na každé složce X_i zvlášť je přitom velmi malá (na některé třeba i nulová).

<u>Koeficient parciální korelace</u> $\rho_{Y,Z:\mathbb{X}}$ měří závislost veličin Y a Z při zafixování vektoru $\mathbb{X}$. Používá se pro případy, kdy veličiny Y a Z se jeví jako závislé, ale tato závislost může být pouze zprostředkovaná tím, že na obě působí skrytě nějaká třetí veličina (vektor $\mathbb{X}$). Cílem je vliv tohoto třetího faktoru vyeliminovat.

Obecně se počítá ze vztahu

$$\rho_{Y,Z:\mathbb{X}} = \frac{\rho_{YZ} - cor(Y,\mathbb{X})\,(cor(\mathbb{X}))^{-1}\,cor(\mathbb{X},Z)}{\sqrt{\left(1-cor(Y,\mathbb{X})(cor\,\mathbb{X})^{-1}cor(\mathbb{X},Y)\right)\left(1-cor(Z,\mathbb{X})(cor\,\mathbb{X})^{-1}cor(\mathbb{X},Z)\right)}}$$

Má-li vektor $\mathbb{X}$ jen jednu složku, pak

$$\rho_{Y,Z.X} = \frac{\rho_{YZ} - \rho_{XY}\,\rho_{XZ}}{\sqrt{(1-\rho^2_{XY})(1-\rho^2_{XZ})}}.$$

<u>Výběrový koeficient parciální korelace</u> pro tři veličiny je

$$r_{Y,Z:X} = \frac{r_{YZ} - r_{YX}\,r_{ZX}}{\sqrt{(1-r^2_{YX})(1-r^2_{ZX})}}$$

Pro vícerozměrný vektor $\mathbb{X}$ je to analogické, do vzorce se místo teoretických korelačních matic dosadí jejich výběrové protějšky.

Test hypotézy $H_0: \rho_{Y,Z:\mathbb{X}} = 0$ se dá založit na veličině

$$T = \frac{r_{Y,Z:\mathbb{X}}}{\sqrt{1-r^2_{Y,Z:\mathbb{X}}}}\sqrt{n-p-2} \qquad (p \text{ je rozměr vektoru } \mathbb{X}),$$

která má za předpokladu normality a za platnosti hypotézy rozdělení t_{n-p-2}.

Nebo se dá přímo porovnat $|r_{Y,Z:\mathbb{X}}|$ s kritickou hodnotou $r_{n-p}(\alpha)$.

Příklad: X_i ... počet členů domácnosti

Y_i ... výdaje domácnosti za potraviny a nápoje (v tisících Kčs na 3 měsíce)

Z_i ... celkový čistý příjem domácnosti (— " —)

X_i	4	2	4	1	5	3	4
Y_i	4	3	4	1	6	4	5
Z_i	10	8	12	3	15	12	13

Údaje jsou z roku 1978!

Vyjde $r_{X,Y} = 0,94$ $\quad r_{X,Z} = 0,92$ $\quad r_{Y,Z} = 0,97$

kritická hodnota $r_7(0,05) = 0,75$.

Zamítá se nezávislost pro všechny dvojice.

$r_{Y,(X,Z)} = 0,98$ $\quad Z = 56,2$ $\quad F_{2,4}(0,05) = 6,9$

Zamítá se nezávislost výdajů na potraviny na současném vlivu počtu členů domácnosti a příjmu.

$r_{Y,Z:X} = 0,83$ $\quad r_{Y,X:Z} = 0,52$ $\quad r_6(0,05) = 0,81$

Zamítá se nezávislost výdajů na příjmech, když se eliminuje vliv počtu členů rodiny.

Nezamítá se nezávislost mezi výdaji a počtem členů domácnosti, když se eliminuje vliv příjmů – tohle vypadá divně, ale může to být tak tím, že těch sledovaných rodin bylo málo. Kritické hodnoty totiž po rostoucím n klesají, takže třeba pro $n = 15$ je $r_{15}(0,05) = 0,51$, pro $n = 30$ je $r_{30}(0,05) = 0,36$. ./.

7. Velikost vzorku je ve statistice zásadní problém. Obecně platí, že čím víc mám pozorování, tím jsou závěry spolehlivější. Kde je ale hranice? To se obtížně řeší, navíc někdy může být problém dostatečný počet hodnot získat.

V data miningu tenhle problém není až tak palčivý, protože datové soubory, se kterými se pracuje, bývají obvykle velké.

12649 Svépomoc II/91-100 – TZ 13-322-91-G

Výpočet podobnosti:

$$r_{X,Y} = \frac{S_{XY}}{\sqrt{S_X^2 S_Y^2}} = \frac{\frac{1}{m-1}\sum_i (X_i-\bar{X})(Y_i-\bar{Y})}{\sqrt{\frac{1}{m-1}\sum_i (X_i-\bar{X})^2 \frac{1}{m-1}\sum_i (Y_i-\bar{Y})^2}} = \frac{\sum_i X_i Y_i - m\bar{X}\bar{Y}}{\sqrt{(\sum_i X_i^2 - m\bar{X}^2)(\sum_i Y_i - m\bar{Y}^2)}}$$

$$\bar{X} = \frac{1}{7}\sum_{i=1}^{7} X_i = \frac{23}{7} = 3,2857 \qquad \bar{Y} = \frac{1}{7}\sum_{i=1}^{7} Y_i = \frac{27}{7} = 3,8571$$

$$\bar{X}^2 = 10,7959 \qquad \bar{Y}^2 = 14,8775$$

$$7 \cdot \bar{X}^2 = 75,5714 \qquad 7 \cdot \bar{Y}^2 = 104,1428$$

$$\sum_{i=1}^{7} X_i^2 = 87 \qquad \sum_{i=1}^{7} X_i Y_i = 101 \qquad \sum_{i=1}^{7} Y_i^2 = 119$$

$$7 \cdot \bar{X}\bar{Y} = 88,7129$$

$$r_{X,Y} = \frac{101 - 88,7129}{\sqrt{(87-75,5714)(119-104,1428)}} = \frac{12,2871}{\sqrt{11,4286 \cdot 14,8572}}$$

$$= \frac{12,2871}{\sqrt{169,797}} = \frac{12,2871}{13,0306} = 0,9429$$

podle tabulek kritická hodnota $r_7(0,05) = 0,75$

podobně se vypočte $r_{X,Z} = 0,92$ a $r_{Y,Z} = 0,97$

Hypotéza o nezávislosti se zamítá pro všechny dvojice.

Koeficient mnohonásobné korelace : testujeme nezávislost Y na současném vlivu X a Z

$$r^2_{Y,(X,Z)} = \frac{r^2_{Y,X} + r^2_{Y,Z} - 2\, r_{Y,X}\, r_{Y,Z}\, r_{X,Z}}{1 - r^2_{X,Z}} =$$

$$= \frac{0{,}94^2 + 0{,}97^2 - 2 \cdot 0{,}94 \cdot 0{,}97 \cdot 0{,}92}{1 - 0{,}92^2} = \frac{0{,}8836 + 0{,}9409 - 1{,}6777}{0{,}1536}$$

$$= \frac{0{,}1468}{0{,}1536} = 0{,}9557$$

$$z = \frac{n-p-1}{p} \cdot \frac{r^2}{1-r^2} = \frac{7-2-1}{2} \cdot \frac{0{,}9557}{1-0{,}9557} = \frac{2 \cdot 0{,}9557}{0{,}0443} =$$

$$= \frac{1{,}9114}{0{,}0443} = 43{,}1467$$

kritická hodnota $F_{2,4}(0{,}05) = 6{,}9$ hypotéza se zamítá

Koeficient parciální korelace : nezávislost Y a Z při pevném X

$$r_{Y,Z:X} = \frac{r_{Y,Z} - r_{YX}\, r_{ZX}}{\sqrt{(1-r^2_{Y,X})(1-r^2_{Z,X})}} = \frac{0{,}97 - 0{,}94 \cdot 0{,}92}{\sqrt{(1-0{,}94^2)\cdot(1-0{,}92^2)}} =$$

$$= \frac{0{,}1052}{\sqrt{0{,}1164 \cdot 0{,}1536}} = \frac{0{,}1052}{\sqrt{0{,}0178}} = \frac{0{,}1052}{0{,}1337} = 0{,}7868$$

kritická hodnota $r_{n-p}(\alpha) = r_6(0{,}05) = 0{,}81$

Hypotéza se nezamítá.

To je divné, ale může to být a) nepřesným výpočtem (zaokrouhlovací chyby)

b) malým počtem domácností (kdyby jich bylo o 1 víc, byla by kritická hodnota 0,75 a hypotéza by se zamítala) data file: csu2002.sta [14 cases with 3 variables]

0 CASE NAME	1 VEK	2 PLAT	3 DETI
Praha	41.500	19879.00	8.400
Stredocesk	39.500	15534.00	9.300
Jihocesky	39.000	14029.00	9.200
Plzensky	39.700	14840.00	8.900
Karlovarsk	38.300	13535.00	9.700
Ustecky	38.300	14269.00	10.000
Liberecky	38.600	14081.00	9.700
Kralovehra	39.600	13753.00	9.000
Pardubicky	39.000	13539.00	9.200
Vysocina	38.700	13441.00	9.200
Jihomoravs	39.500	14059.00	9.000
Olomoucky	39.000	13373.00	8.900
Zlinsky	39.000	13843.00	8.800
Moravskosl	38.500	14923.00	9.100

Údaje Českého statistického úřadu na rok 2002:

průměrný věk

průměrná mzda v Kč

počet živě narozených dětí na 1000 obyvatel

Jiný příklad:

STAT. BASIC STATS	Correlations (csu2002.sta) Marked correlations are significant at p < .05000 N=14 (Casewise deletion of missing data)		
Variable	VEK	PLAT	DETI
VEK	1.00	.82 *	-.77 *
PLAT	.82 *	1.00	-.49
DETI	-.77 *	-.49	1.00

$r_{věk,\,plat} = 0{,}82$

$r_{věk,\,děti} = -0{,}77$

$r_{plat,\,děti} = -0{,}49$

kritická hodnota $r_{14}(0{,}05) = 0{,}5324$

Nezávislost platu na věku se zamítá.

—"— dětí —"— —"—

—"— dětí na platu se nezamítá.

Koeficient mnohonásobné korelace $r_{děti,\,(věk,\,plat)}$:

$$r^2 = \frac{r^2_{děti,\,věk} + r^2_{děti,\,plat} - 2\, r_{děti,\,věk}\, r_{děti,\,plat}\, r_{věk,\,plat}}{1 - r^2_{věk,\,plat}} =$$

$$= \frac{(-0{,}77)^2 + (-0{,}49)^2 - 2(-0{,}77)(-0{,}49)\cdot 0{,}82}{1 - 0{,}82^2} =$$

$$= \frac{0{,}5929 + 0{,}2401 - 0{,}618772}{0{,}3276} = \frac{0{,}214228}{0{,}3276} = 0{,}6539$$

$$\chi = \frac{14-2-1}{2} \cdot \frac{0,6539}{1-0,6539} = \frac{11 \cdot 0,6539}{2 \cdot 0,3461} = 10,39$$

kritická hodnota $F_{2,11}(0,05) = 3,98$

Nezávislost počtu dětí na souhrnném vlivu věku a platu se zamítá.

Zkusme nezávislost počtu dětí na platu, když se eliminuje věk:

$$r_{děti,plat \cdot věk} = \frac{r_{děti,plat} - r_{děti,věk}\, r_{plat,věk}}{\sqrt{(1-r^2_{děti,věk})(1-r^2_{plat,věk})}} =$$

$$= \frac{-0,49 - (-0,77)(0,82)}{\sqrt{(1-(-0,77)^2)(1-0,82^2)}} = \frac{0,1414}{\sqrt{0,4071 \cdot 0,3276}} = \frac{0,1414}{0,3652} =$$

$$= 0,3872$$

kritická hodnota $r_{13}(0,05) = 0,5529$

Ani tady se hypotéza nezamítá.

Opět je třeba brát v úvahu rozsah výběru: $r_{17}(0,05) = 0,48$

$r_{30}(0,05) = 0,36$

Poznámka: Korelace při eliminaci věku vychází sice nevýznamná, ale kladná (tj. výše platu působí příznivě na počet narozených dětí). Jednoduchá korelace byla naopak záporná a je možné, že vyjadřovala zprostředkovanou závislost na věku (čím vyšší věk, tím vyšší plat, ale tím méně narozených dětí).

Co v případech, kdy data nepocházejí z normálního rozdělení?

Nechť $\binom{X_1}{Y_1}, \dots, \binom{X_n}{Y_n}$ je náhodný výběr z nějakého dvourozměrného rozdělení,

$R_1, \dots, R_n$ jsou pořadí veličin $X_1, \dots, X_n$

$Q_1, \dots, Q_n$ — " — $Y_1, \dots, Y_n$

Vypočte se tzv. <u>Spearmanův korelační koeficient</u>

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

Ve prospěch hypotézy $H_0: \rho = 0$ svědčí malé hodnoty r_S, protože pokud by si byla pořadí X-ových a Y-ových veličin hodně podobná, bylo by asi závislé.
H_0 se zamítá, když $|r_S| \geq r_S(\alpha)$, kde $r_S(\alpha)$ je tabelovaná kritická hodnota.

Příklad: Spotřeba alkoholu a úmrtnost, Anděl str. 204

Na tomto testu je navíc sympatické, že ani přesné hodnoty sledovaných veličin nemusíme znát, stačí ta pořadí.

Např. nám stačí pořadí zemí ve spotřebě alkoholu a pořadí v úmrtnosti na cirhózu jater a nepotřebujeme vědět, kolik se tam přesně vypilo a kolik lidí umřelo.

Příklad: Spotřeba alkoholu a úmrtnost na cirhózu jater

země:	spotřeba alkoholu (pořadí):	úmrtnost (pořadí):
Finsko	3,9 (1)	3,6 (3)
Norsko	4,2 (2)	4,3 (5)
Irsko	5,6 (3)	3,4 (2)
Holandsko	5,7 (4)	3,7 (4)
Švédsko	6,6 (5)	7,2 (7)
Anglie a Wales	7,2 (6)	3,0 (1)
Belgie	10,8 (7)	12,3 (8)
Rakousko	10,9 (8)	7,0 (6)
NSR	12,3 (9)	23,7 (10)
Itálie	15,7 (10)	23,6 (9)
Francie	24,7 (11)	46,1 (11)

(údaje jsou přepočtené na 100 000 obyvatel, pocházejí z jakési zahraniční sbírky příkladů z r. 1979)

hodnota $r_s = 0,773$

kritická hodnota $r_s(0,05) = 0,6091$

Zamítá se hypotéza, že veličiny jsou nezávislé.

Pozn: I tady se ale mohou skrytě projevovat další faktory, které nezkoumáme -

Opět vyzkouším na příkladu ČSÚ: (1 = nejnižší)

Kraj:	pořadí věk	pořadí plat	pořadí děti
Praha	14	14	1
StřČ.	10	13	11
JČ	6	7	8
Plzeň	13	11	3
KV	1	3	12
Ústí	1	10	14
Liberec	4	9	12
HK	12	5	5
Pardubice	6	4	8
Vysočina	5	2	8
JMor.	10	8	5
Olomouc	6	1	3
Zlín	6	6	2
Mor. Sl.	3	12	7

Korelace věk-děti: $r_s = 1 - \frac{6}{14(14^2-1)}\left(13^2 + 1 + 2^2 + 10^2 + 11^2 + 13^2 + 8^2 + 7^2 + 2^2 + 3^2 + 5^2 + 9^2 + 4^2 + 4^2\right) =$

$= 1 - \frac{6}{2730} \cdot 828 = -0{,}8197$

kritická hodnota: $r_s(14;\, 0{,}05) = 0{,}5341$

zamítá se

Korelace plat – děti:

$$r_s = 1 - \frac{6}{2730}\left(13^2 + 2^2 + 1 + 8^2 + 9^2 + 4^2 + 3^2 + 0 + 4^2 + 6^2 + \right.$$
$$\left. + 3^2 + 2^2 + 4^2 + 5^2\right) = 1 - \frac{6}{2730} \cdot 450 = 0{,}01$$

Hypotéza o nezávislosti se nezamítá.