evropský sociální fond v ČR
EVROPSKÁ UNIE
MINISTERSTVO ŠKOLSTVÍ, MLÁDEŽE A TĚLOVÝCHOVY
OP Vzdělávání pro konkurenceschopnost
UNIVERZITA OBRANY
INVESTICE DO ROZVOJE VZDĚLÁVÁNÍ

Charakteristiky statistické vazby a jejich výpočet

1 Kategoriální (nominální) znaky

1.1 χ^2-test nezávislosti

Mějme kategoriální proměnné X a Y. Vytvoříme tzv. **kontingenční tabulku**. Budeme tedy testovat hypotézu H : proměnné X a Y jsou nezávislé, proti alternativní hypotéze, že jsou závislé. Pro nezávislé jevy A, B platí: $P(A \cap B) = P(A) \cdot P(B)$. Budeme porovnávat empiricky zjištěné četnosti n_{ij} četnostmi teoretickými

$$n \cdot \pi_{ij} = n \cdot \pi_{i\bullet} \cdot \pi_{\bullet j}.$$

Odhady teoretických četností jsou

$$\hat{\pi}_{i\bullet} = \frac{n_{i\bullet}}{n} \quad \text{a} \quad \hat{\pi}_{\bullet j} = \frac{n_{\bullet j}}{n},$$

odhad teoretické sdružené pravděpodobnosti je

$$\hat{\pi}'_{ij} = \hat{\pi}_{i\bullet} \cdot \hat{\pi}_{\bullet j} = \frac{n_{i\bullet}}{n} \cdot \frac{n_{\bullet j}}{n} = \frac{n_{i\bullet} \cdot n_{\bullet j}}{n^2}.$$

Potom odhad teoretické četnosti je

$$n'_{ij} = n \cdot \hat{\pi}'_{ij} = n \cdot \frac{n_{i\bullet} \cdot n_{\bullet j}}{n^2} = \frac{n_{i\bullet} \cdot n_{\bullet j}}{n}$$

Pro test H : proměnné X a Y jsou nezávislé $\rightarrow A$: proměnné X a Y jsou závislé užijeme testovou statistiku

$$\chi^2 = \sum_{i=1}^{r} \sum_{j=1}^{s} \frac{(n_{ij} - n'_{ij})^2}{n'_{ij}},$$

která má za předpokladu nezávislosti znaků X a Y pro dostatečně velké n přibližně Pearsonovo $\chi^2(\nu)$ rozdělení se stupni volnosti $\nu = (r-1)(s-1)$. (n_{ij} jsou empirické četnosti, n'_{ij} jsou teoretické četnosti). Hypotézu o nezávislosti znaků X a Y zamítáme, jestliže

$$\chi^2 \geq \chi^2_{1-\alpha}(\nu), \text{ kde } \nu = (r-1)(s-1).$$

Test χ^2 lze korektně použít tehdy, pokud jsou všechny buňky tabulky dostatečně obsazené, tj. když pro alespoň 80 % teoretických četností platí $n'_{ij} \geq 5$ a zbývající teoretické četnosti jsou $n'_{ij} > 1$. při nesplnění této podmínky se doporučuje spojování „sousedních“ obměn u jedné nebo druhé proměnné (sčítáme celé řádky nebo sloupce a při opakovaném testu s nimi zacházíme jakou 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

Operační program Vzdělávání pro konkurenceschopnost
Název projektu: Inovace magisterského studijního programu Fakulty ekonomiky a managementu
Registrační číslo projektu: CZ.1.07/2.2.00/28.0326

PROJEKT JE SPOLUFINANCOVÁN EVROPSKÝM SOCIÁLNÍM FONDEM A STÁTNÍM ROZPOČTEM ČESKÉ REPUBLIKY.

	j					
i	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

Tab. 1: Tabulka teoretických četností

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.

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

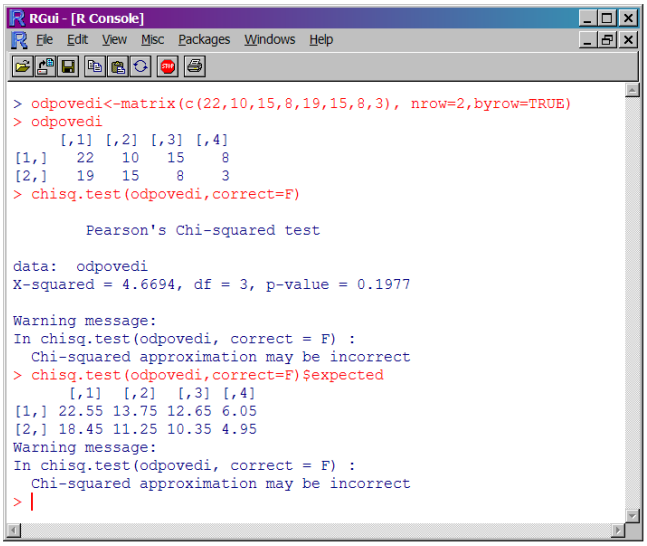
	j				
i	1	2	3	4	celkem
1	22,55	13,75	12,65	6,05	55,00
2	18,45	11,25	10,35	4,95	45,00
celkem	41,00	25,00	23,00	11,00	100,00
	j				
i	1	2	3	4	celkem
1	0,013	1,023	0,437	0,629	2,101
2	0,016	1,250	0,534	0,768	2,568
celkem	0,03	2,273	0,97	1,397	4,669

Tab. 2: Tabulka teoretických četností a výpočet testové statistiky

Hodnota testové statistiky je tedy

$$\chi^2 = \sum_{i=1}^{r}\sum_{j=1}^{s}\frac{(n_{ij}-n'_{ij})^2}{n'_{ij}} = 4{,}66,$$

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 $\chi^2_{1-\alpha}(3) = 7{,}815$. Hodnota testového kritéria nepatří do kritického oboru, tedy se s 95% pravděpodobností neprokázalo, že odpověď na danou otázku závisí na pohlaví. Test nezávislosti v kontingenční tabulce lze v programu R spočítat pomocí funkce **chisq.test**.



1.2 Koeficienty kontingence

Těsnost 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

2 Ordinální znaky

2.1 Spearmanův korelační koeficient

V případě dvourozměrného souboru kvalitativních údajů, které jsou po složkách ordinálního typu, je možno zjistit stupeň závislosti těchto dvou znaků. K měření takovýchto závislostí se používá **Spearmanův korelační koeficient**. Hodnotám x_i, y_i přiřadíme pořadová čísla p_i, q_i (pořadí jednotlivých hodnot při uspořádaní podle velikosti). Spearmanův koeficient (koeficient pořadové korelace) je potom definován vztahem

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

	Spotřeba cigaret		Úmrtnost		
Stát USA	x_i	p_i	y_i	q_i	$(p_i - q_i)^2$
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 Yersy	2900	5	26	6	1
Washington	2100	1	20	3	4

Příklad. 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. Suma kvadrátů v posledním sloupci je 8,

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

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

2.2 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.

3 Měřitelné znaky

3.1 Pearsonův korelační koeficient

Mějme dvourozměrný datový soubor $\begin{pmatrix} x_1 & y_1 \\ \vdots & \vdots \\ x_n & y_n \end{pmatrix}$, označme $\overline{x}$ a $\overline{y}$ průměry znaků a s_x, s_y směrodatné odchylky znaků X, Y. **Koeficient korelace (Pearsonův)** definujeme vztahem

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

kde $s_{xy} = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})$ je výběrová kovariance znaků X a Y, $s_x = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2}$ je výběrová směrodatná odchylka znaku X a $s_y = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(y_i - \overline{y})^2}$ je výběrová směrodatná odchylka znaku Y. Pearsonův korelační koeficient lze jej vyjádřit ve tvaru

$$r_{xy} = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \overline{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \overline{y})^2}} = \\ = \frac{n\sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{\sqrt{n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2}\sqrt{n\sum_{i=1}^{n} y_i^2 - \left(\sum_{i=1}^{n} y_i\right)^2}}.$$

4 Popis náhodných vektorů

4.1 Sdružená distribuční funkce

Zaměříme se především na popis dvourozměrných náhodných veličin (vektorů)

Definice 4.1 Nechť X a Y jsou náhodné veličiny, $\mathbf{X} = (X, Y)$ se nazývá (dvourozměrný) **náhodný vektor**.

Definice 4.2 **Sdružená distribuční funkce** vektoru $\mathbf{X} = (X, Y)$ je definována jako

$$F(\mathbf{X}) = F(X, Y) = P(X \leq x, Y \leq y) \quad \text{pro } x, y \in \mathbb{R}.$$

Věta 4.1 Má-li $\mathbf{X} = (X, Y)'$ sdruženou distribuční funkci F, potom X a Y mají distribuční funkce

$$F_X(x) = F(x, \infty) \quad \text{a} \quad F(y) = F(\infty, y).$$

Distribuční funkce náhodných veličin X a Y se nazývají **marginální distribuční funkce**.

4.2 Diskrétní náhodný vektor

Definice 4.3 Jsou-li X a Y diskrétní náhodné veličiny, potom $\mathbf{X} = (X, Y)'$ se nazývá **diskrétní náhodný vektor**.

Definice 4.4 Je -li $\mathbf{X} = (X, Y)$ diskrétní náhodný vektor $\{(x_j, y_k) : j, k = 1, 2, \dots\}$, funkce

$$p(x_j, y_k) = P(X = x_j, Y = y_k)$$

se nazývá **sdružená pravděpodobnostní funkce** vektoru $\mathbf{X} = (X, Y)'$.

Věta 4.2 Jestliže $\mathbf{X} = (X, Y)'$ má sdruženou pravděpodobnostní funkci p, pak marginální pravděpodobnostní funkce X a Y jsou

$$p_X(x_j) = \sum_{k=1}^{\infty} p(x_j, y_k), \quad j = 1, 2, \dots,$$

$$p_Y(y_k) = \sum_{j=1}^{\infty} p(x_j, y_k), \quad k = 1, 2, \dots.$$

4.3 Spojitý náhodný vektor

Definice 4.5 Existuje-li taková funkce f, pro kterou platí

$$P((X,Y) \in B) = \iint_B f(x,y)dx\,dy$$

pro všechny podmnožiny $B \subseteq \mathbb{R}^2$, potom funkce f se nazývá **sdružená hustota pravděpodobnosti** vektoru $\mathbf{X} = (X,Y)'$.

Je-li $B = \{(s,t) : s \le x, t \le y\} = (-\infty, x\rangle \times (-\infty, y\rangle$, potom

$$F(x,y) = \int_\infty^y \int_{-\infty}^x f(s,t)ds\,dt.$$

Věta 4.3 Jsou-li X a Y spojité náhodné veličiny se sdruženou distribuční funkcí F a sdruženou hustotou f, potom

$$f(x,y) = \frac{\partial^2}{\partial x \partial y} F(x,y), \quad x,y \in \mathbb{R}.$$

Sdružená funkce hustoty pravděpodobnosti musí mít následující vlastnosti:

- $f(x,y) \ge 0$ pro všechna $x, y \in \mathbb{R}$,
- $\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f(x,y)dx\,dy = 1$.

Věta 4.4 Nechť X a Y spojité náhodné veličiny se sdruženou hustotou f, potom jejich marginální hustoty jsou

$$f_X(x) = \int_{-\infty}^{\infty} f(x,y)dy, \quad x \in \mathbb{R},$$
$$f_Y(y) = \int_{-\infty}^{\infty} f(x,y)dx, \quad y \in \mathbb{R}.$$

4.4 Podmíněné rozdělení a nezávislost

Definice 4.6 Nechť X a Y jsou diskrétní náhodné veličiny s obory hodnot $\{x_1, x_2, \dots\}$ a $\{y_1, y_2, \dots\}$, nechť p je jejich sdružená distribuční funkce. Podmíněná pravděpodobnostní funkce Y za podmínky $X = x$ je definována vztahem

$$p_Y(y_k|x_j) = \frac{p(x_j, y_k)}{p_X(x_j)} \quad \text{pro } y_k \in \{y_1, y_2, \dots\}.$$

$$p_Y(y_k) = \sum_{j=1}^{\infty} p_Y(y_k|x_j) p_X(x_j)$$

Definice 4.7 Nechť spojité náhodné veličiny X a Y mají sdruženou hustotu f. Podmíněná funkce hustoty pravděpodobnosti Y za podmínky $X = x$ je definována vztahem

$$f_Y(y|x) = \frac{f(x,y)}{f_X(x)} \quad y \in \mathbb{R}.$$

Definice 4.8 Náhodné veličiny X a Y jsou **nezávislé**, jestliže

$$P(X \in A, Y \in B) = P(X \in A)P(Y \in B) \quad \text{pro všechny } A, B \subseteq \mathbb{R}.$$

Věta 4.5 Náhodné veličiny X a Y jsou **nezávislé** právě když

$$F(x, y) = F_X(x)F_Y(y) \quad \text{pro každé } x, y \in \mathbb{R}.$$

Věta 4.6 Diskrétní náhodné veličiny X a Y se sdruženou pravděpodobnostní funkcí p jsou **nezávislé** právě když

$$p(x, y) = p_X(x)p_Y(y) \quad \text{pro každé } x, y \in \mathbb{R}.$$

Věta 4.7 Spojité náhodné veličiny X a Y se sdruženou hustotou f jsou **nezávislé** právě když

$$f(x, y) = f_X(x)f_Y(y) \quad \text{pro každé } x, y \in \mathbb{R}.$$

4.5 Charakteristiky náhodného vektoru

Definice 4.9 Existují-li střední hodnoty $E(X)$, $E(Y)$ náhodných veličin X a Y, pak střední hodnota vektoru $\mathbf{X} = (X, Y)'$ je vektor

$$E(\mathbf{X}) = (E(X), E(Y))'.$$

Věta 4.8 Jestliže $E(X^2) < \infty$ a $E(Y^2) < \infty$, definujeme **kovarianci** náhodných veličin vztahem

$$C(X, Y) = E[X - E(X)][Y - E(Y)].$$

Pro kovarianci platí:

- $C(Y, X) = C(X, Y)$,
- $C(X, X) = D(X)$,
- $C(X, Y) = E(XY) - E(X)E(Y)$,
- $C(a_1 + a_2X, b_1 + b_2Y) = a_2b_2C(X, Y)$, kde $a_1, a_2, b_1, b_2 \in \mathbb{R}$.

Definice 4.10 **Korelační koeficient** náhodných veličin X a Y je definován vztahem

$$\rho(X, Y) = \frac{C(X, Y)}{\sqrt{D(X)D(Y)}}.$$

Pro korelační koeficient platí:

- $-1 \leq \rho(X, Y) \leq 1$,
- jestliže jsou X a Y nezávislé, pak $\rho(X, Y) = 0$,
- $\rho(X, Y) = 1$ právě když $Y = aX + b$, kde $a > 0$,
- $\rho(X, Y) = -1$ právě když $Y = aX + b$, kde $a < 0$.

Definice 4.11 **Varianční maticí** vektoru $\mathbf{X} = (X, Y)'$ rozumíme matici

$$\operatorname{var}\mathbf{X} = E[\mathbf{X} - E(\mathbf{X})][\mathbf{X} - E(\mathbf{X})]' = E(\mathbf{X}\mathbf{X}') - [E(\mathbf{X})][E(\mathbf{X})]'.$$

$$\operatorname{var}\mathbf{X} = \begin{pmatrix} C(X,X) & C(X,Y) \\ C(Y,X) & C(Y,Y) \end{pmatrix} = \begin{pmatrix} D(X) & C(X,Y) \\ C(X,Y) & D(Y) \end{pmatrix}.$$

Definice 4.12 **Korelační maticí** vektoru $\mathbf{X} = (X, Y)'$ rozumíme matici

$$\operatorname{cor}\mathbf{X} = \begin{pmatrix} 1 & \rho(X,Y) \\ \rho(X,Y) & 1 \end{pmatrix}.$$

Věta 4.9 Nechť X a Y jsou náhodné veličiny, pak

$$E(X+Y) = E(X) + E(Y).$$

Pro spojité veličiny platí

$$\begin{aligned} E(X+Y) &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}(x+y)f(x,y)dx\,dy \\ &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}xf(x,y)dx\,dy + \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}yf(x,y)dx\,dy \\ &= \int_{-\infty}^{\infty}xf_X(x)dx + \int_{-\infty}^{\infty}yf_Y(y)dy = E(X) + E(Y) \end{aligned}$$

Věta 4.10 Nechť X a Y jsou náhodné veličiny, $a, b \in \mathbb{R}$, pak

$$E(aX + bY) = aE(X) + bE(Y).$$

Věta 4.11 Nechť X a Y jsou náhodné veličiny, pak

$$D(X+Y) = D(X) + D(Y) + 2C(X,Y).$$

$$\begin{aligned} D(X+Y) &= E[(X+Y)^2] - [E(X+Y)]^2 \\ &= E(X^2) + 2E(XY) + E(Y^2) - E(X)^2 - 2E(X)E(Y) - E(Y)^2 \\ &= E(X^2) - E(X^2) + E(Y) - E(Y^2) + 2(E(XY) - E(X)E(Y)) \\ &= D(X) + D(Y) + 2C(X,Y) \end{aligned}$$

Věta 4.12 Nechť X a Y jsou nezávislé náhodné veličiny, pak platí

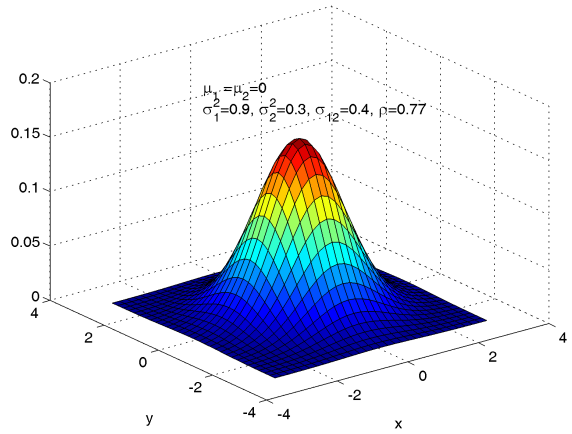
- $E(XY) = E(X)E(Y)$,
- $D(X+Y) = D(X) + D(Y)$.

$$\begin{aligned} E(XY) &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}xyf(x,y)dx\,dy \\ &= \int_{-\infty}^{\infty}xf_X(x)dx\int_{-\infty}^{\infty}yf_Y(y)dy = E(X)E(Y) \end{aligned}$$

$$C(X,Y) = E(XY) - E(X)E(Y) = E(X)E(Y) - E(X)E(Y) = 0$$

Věta 4.13 Nechť X a Y jsou nezávislé náhodné veličiny, $a, b \in \mathbb{R}$, pak

$$D(aX + bY) = a^2D(X) + b^2D(Y).$$



Obr. 1: Graf dvourozměrného normálního rozdělení

5 Dvourozměrné normální rozdělení

Definice 5.1 Má-li náhodný vektor $\mathbf{X} = (X, Y)'$ sdruženou hustotu pravděpodobnosti

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

pro $x, y \in \mathbb{R}$, pak říkáme, že má **dvourozměrné normální rozdělení** s parametry $\mu_1, \mu_2, \sigma_1, \sigma_2, \rho$.

Věta 5.1 Nechť $\mathbf{X} = (X, Y)'$ má dvourozměrné normální rozdělení s parametry $\mu_1, \mu_2, \sigma_1, \sigma_2, \rho$, potom

- $X \sim N(\mu_1, \sigma_1^2)$ a $\mathrm{Y} \sim N(\mu_2, \sigma_2^2)$,
- ρ je korelační koeficient X a Y.

Věta 5.2 Nechť $\mathbf{X} = (X, Y)'$ má dvourozměrné normální rozdělení. Pro pevné $x \in \mathbb{R}$ platí

$$Y|X = x \sim N\left(\mu_2 + \rho\frac{\sigma_2}{\sigma_1}(x-\mu_1), \sigma_2^2(1-\rho^2)\right)$$

Z uvedené věty plyne, že

$$E(Y|X) = \mu_2 + \rho\frac{\sigma_2}{\sigma_1}(x-\mu_1).$$

Tato podmíněná střední hodnota se nazývá **regresní přímka.** Empirickým protějškem korelačního koeficientu ρ **výběrový korelační koeficient** (koeficient korelace) r

$$r = \frac{s_{xy}}{s_x \cdot s_y},$$

kde $s_{xy} = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})$ je **výběrová kovariance**, s_x a s_y jsou výběrové směrodatné odchylky. Korelační koeficient r lze vyjádřit ve tvaru

$$r = \frac{n\sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{\sqrt{n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2}\sqrt{n\sum_{i=1}^{n} y_i^2 - \left(\sum_{i=1}^{n} y_i\right)^2}}$$

Koeficient determinace je pro závislost popsanou regresní přímkou zvláštním případem indexu determinace, tedy platí $r_{yx}^2 = \frac{S_T}{S_Y}$. Tato míra těsnosti závislosti má zcela stejné vlastnosti jako i_{yx}^2.

Výběrový koeficient determinace r_{yx}^2 lze použít jako odhad teoretického koeficientu determinace ρ^2 v základním souboru. Úpravou

$$r_{kor}^2 = 1 - (1 - r^2)\frac{n-1}{n-2}$$

získáme nestranný odhad ρ^2.

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

Testové kritérium je statistika

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

Kritický obor je dán

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

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

5.1 Koeficient mnohonásobné korelace

Koeficient mnohonásobné korelace vyjadřuje společné působení nezávisle proměnných $X_1, X_2, \ldots X_k$ 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_{yx} je výběrový korelační koeficient mezi x_i a z_i. Jeho druhou mocninou je index determinace.

Příklady k procvičení

1. V následující tabulce je výsledek žákovského průzkumu o oblibě přírodovědných předmětů na gymnáziu ve třech čtvrtých ročnících. Zajímá nás, zda obliba předmětů závisí na pohlaví dotazovaných studentů. Ověřte χ^2-testem, zda jsou veličiny: obliba přírodovědných předmětů a pohlaví nezávislé. Formulujte nulovou hypotézu a výsledek testu komentujte slovně. Zkontrolujte, zda četnosti splňují podmínku pro platnost χ^2-test a zdůvodněte slovně.

	Matematika	Fyzika	Chemie	Biologie	Celkem
Dívky	26	18	12	24	80
Chlapci	8	16	12	4	40
Celkem	34	34	24	28	120

2. Ve studii nozokomiálních infekcí se sledovaly veličiny, které obecně platí za rizikový faktor. V následující tabulce je výsledek šetření ve 100 nemocnicích v ČR a za rizikový faktor je považováno pohlaví pacienta. Zajímá nás, zda se potvrdí závislost výskytu NI na pohlaví. Ověřte testem χ^2-testem, zda jsou tyto veličiny nezávislé. Formulujte nulovou hypotézu a výsledek testu komentujte slovně.

Pohlaví	Pacienti s NI	Pacienti bez NI	Celkem
Muž	238	715	953
Žena	131	531	662
Celkem	369	1246	1615

3. Z dlouhodobého sledování dětí u dětského lékaře vyšly tyto údaje:

	Vrozené vady kyčlí		
Pohlaví	Ano	Ne	Celkem
Chlapci	325	1589	1914
Dívky	389	1525	1914
Celkem	714	3114	3828

   Ověřte χ^2-testem, zda jsou vrozená vada kyčlí a pohlaví dítěte nezávislé veličiny. Formulujte nulovou hypotézu a výsledek testu komentujte slovně. Určete Pearsonův a Cramerův koeficient kontingence.

4. V následující tabulce je výsledek žákovského průzkumu o oblibě předmětů na jazykové škole ve druhých ročnících. Zajímá nás, zda obliba předmětů závisí na pohlaví dotazovaných studentů. Ověřte χ^2-testem, zda jsou veličiny: obliba humanitních předmětů a pohlaví nezávislé. Formulujte nulovou hypotézu a výsledek testu komentujte slovně. Zkontrolujte, zda četnosti splňují podmínku pro platnost testu χ^2-testem a zdůvodněte slovně. Určete Pearsonův, Cramerův a Čuprovův koeficient kontingence.

	Čeština	Francouzština	Angličtina	Latina	Celkem
Dívky	24	18	21	4	67
Chlapci	15	14	22	2	53
Celkem	39	32	43	6	120

5. Byla zjištěna výška otců a výška jejich nejstarších synů [v cm].

otec	165	178	158	170	180	160	170	167	185	165	173	175
syn	162	184	163	170	189	165	177	170	187	176	171	183

Určete Pearsonův, Spearmanův a Kendallův korelační koeficient. Proveďte test významnosti Pearsonova korelačního koeficientu. [Datový soubor: vyska_otec_syn.txt]

6. O 7 vybraných strojích v určitém podniku máme informace o jejich stáří (v letech) a týdenních nákladech na jejich údržbu (v Kč):

stáří stroje	1	1	3	3	5	6	7
náklady	35	52	81	105	100	125	120

Určete Pearsonův, Spearmanův a Kendallův korelační koeficient. Proveďte test významnosti Pearsonova korelačního koeficientu. [Datový soubor: stari_stroje_naklady.txt]

7. U 30 žáků byly zjištěny hodnoty znaků X – známka z fyziky, Y – známka z chemie a Z – pohlaví (0. . . dívka, 1. . . chlapec).

Fyzika	Chemie	Pohlaví	Fyzika	Chemie	Pohlaví	fyzika	Chemie	Pohlaví
1	2	1	2	3	1	3	4	0
2	3	1	2	3	0	5	4	0
2	3	0	2	2	0	2	1	0
4	5	1	4	5	0	2	2	0
2	1	1	3	3	1	3	1	1
4	3	1	3	4	1	3	4	0
2	2	1	2	3	1	2	1	1
4	2	0	2	2	0	1	1	0
3	3	0	5	3	1	1	1	1
4	5	0	3	2	1	1	2	0

Určete Pearsonův, Spearmanův a Kendallův korelační koeficient pro známky z fyziky a chemie. Výpočty proveďte zvlášť pro dívky, pro hochy a dohromady bez závislosti na pohlaví.

[Datový soubor: znamky.txt]

8. U 40 pracovníků byla sledována závislost počtu chybných operací za směnu (Y) na délce zapracování v hodinách (X) s těmito výsledky:

x_i	y_i	x_i	y_i	x_i	y_i	x_i	y_i	x_i	y_i	x_i	y_i	x_i	y_i	x_i	y_i
3	2	2	4	1	5	2	6	4	3	3	6	2	5	3	5
4	3	2	6	3	3	4	4	1	6	5	2	4	1	4	2
1	4	4	1	2	7	4	3	2	4	3	3	1	6	1	7
4	5	2	5	3	4	3	5	2	5	2	3	5	4	4	4
4	2	5	3	5	1	3	4	5	1	3	4	3	4	5	2

Určete Pearsonův, Spearmanův a Kendallův korelační koeficient. Proveďte test významnosti Pearsonova korelačního koeficientu. [Datový soubor: chyby_zapracovani.txt]

9. Data popisují výsledky vstupních zdravotních testů uchazečů o službu u policie.

Tlak	66	87	85	59	76	77	70	66	75	66
Hmotnost	87,36	117,6	82,85	62,32	82	102	70,12	88,07	77,96	74,33
Tuk	16,98	27,6	6,61	3,26	19	27	6,88	18,8	18,87	8,15
Tlak	74	68	72	76	94	63	80	67	77	78
Hmotnost	56,2	81,75	80,24	74,81	61,98	95,23	72,48	92,45	104,56	66,2
Tuk	3,44	20,31	12,96	12,42	3,58	12,91	11,34	17,5	18,93	10,94
Tlak	77	67	78	78	80	95	76	78	73	80
Hmotnost	87,16	82,42	64,11	81,57	99,85	78,49	87,13	65,64	51,76	67,14
Tuk	17,72	9,55	9,54	13,1	17,75	9,57	18,52	6,4	2,86	4,31
Tlak	81	61	65	69	66	75	72	66	93	77
Hmotnost	78,74	86,83	70,48	72,67	85,86	84,86	66,97	68,33	63,34	85,72
Tuk	16,26	9,72	6,29	4,37	14,43	17	5,8	8,14	3,63	23,61
Tlak	68	71	84	81	74	79	89	79	80	67
Hmotnost	89	95,17	84,19	63,12	70,01	82,11	71	94,56	70,91	79,19
Tuk	18,83	19,16	15,83	8,77	6,61	22,22	8,29	26,82	9,32	19,9

Určete Pearsonův, Spearmanův a Kendallův korelační koeficient pro jednotlivé dvojice proměnných. Spočtete koeficienty mnohonásobné korelace. [Datový soubor: vstupni_testy.txt]