# Analýza dat v neurologii

## LXXI. Pearsonův korelační koeficient

L. Dušek, T. Pavlík,
J. Jarkovský, J. Koptíková
Institut biostatistiky a analýz,
LF MU, Brno

✉
doc. RNDr. Ladislav Dušek, Ph.D.
Institut biostatistiky a analýz,
LF MU, Brno
e-mail: dusek@iba.muni.cz

V sérii výukových článků tento rok řešíme hodnocení vztahu dvou spojitých proměnných. Výklad jsme započali analýzou lineárního vztahu mezi dvěma spojitými, normálně rozloženými veličinami a minulé dva díly jsme věnovali vysvětlení pojmu kovariance značené *cov(X, Y)*. Kovariance je jedním ze základních ukazatelů síly vztahu dvou proměnných. U normálně rozložených veličin pracujeme s aritmetickým průměrem a rozptylem a z těchto statistických ukazatelů středu a variability rozložení vychází také vztah pro výpočet kovariance:

$$cov(X, Y) = \frac{\sum_{i=1}^{N} (x_i - \bar{x}) \times (y_i - \bar{y})}{N - 1}, \text{ kde}$$

- $x_i$, $y_i$ jsou jednotlivé hodnoty proměnných *X* a *Y* naměřené párově u $i = 1$ až $i = N$ jedinců v analyzovaném souboru,
- $\bar{x}$, $\bar{y}$ jsou aritmetické průměry proměnných *X* a *Y*.

Připomeňme, že kovariance je ukazatelem síly lineárního vztahu dvou proměnných, přičemž její kladná hodnota značí vztah pozitivní a záporná hodnota vztah negativní. Kovariance blízká nule dokládá neexistenci vztahu, kdy hodnoty obou proměnných na sobě nijak nezávisí a vyskytují se v pozici vůči svým průměrným hodnotám zcela náhodně.

V minulém díle jsme rovněž rozebírali největší nevýhodu kovariance, a to že její hodnoty závisí na rozptylu obou proměnných, resp. na jednotkách, ve kterých jsou vyjadřovány. Pro odhad kovariance tedy není definována maximální hodnota, která by vyjadřovala nejsilnější možný vztah zkoumaných proměnných (jejich hodnoty by v takovém případě ležely přesně na přímce). To značně omezuje interpretaci odhadu kovariance a snižuje srovnatelnost odhadů kovariance z různých studií. Proto bývá kovariance často

| $x$ | $\bar{x} - x$ | $y$ | $\bar{y} - y$ | $(\bar{x} - x) * (\bar{y} - y)$ |
|---|---|---|---|---|
| 3 | −2,5 | 4 | −4,5 | 11,25 |
| 11 | 5,5 | 8 | −0,5 | −2,75 |
| 4 | −1,5 | 8 | −0,5 | 0,75 |
| 2 | −3,5 | 10 | 1,5 | −5,25 |
| 5 | −0,5 | 10 | 1,5 | −0,75 |
| 8 | 2,5 | 11 | 2,5 | 6,25 |

$\bar{x} = 5{,}5$ $\bar{y} = 8{,}5$ $\sum (\bar{x} - x) * (\bar{y} - y) = 9{,}5$

$$R = \frac{Cov(X,Y)}{s_x s_y}$$

$$Cov(X,Y) = \frac{\sum_{i=1}^{N}(x_i - \bar{x}) * (y_i - \bar{y})}{N - 1}$$

$$s^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{x})^2 \qquad s = \sqrt{s^2}$$

$$Cov(X,Y) = \frac{(3-5{,}5)*(4-8{,}5)}{5} + \frac{(11-5{,}5)*(8-8{,}5)}{5} + \frac{(4-5{,}5)*(8-8{,}5)}{5} + \frac{(2-5{,}5)*(10-8{,}5)}{5} + \frac{(5-5{,}5)*(10-8{,}5)}{5} + \frac{(8-8{,}5)*(11-8{,}5)}{5} = 1{,}9$$

$$s_x = \sqrt{\frac{1}{5} * \left((3-5{,}5)^2 + (11-5{,}5)^2 + (4-5{,}5)^2 + (2-5{,}5)^2 + (5-5{,}5)^2 + (8-5{,}5)^2\right)} = 3{,}39$$

$$s_y = \sqrt{\frac{1}{5} * \left((4-8{,}5)^2 + (8-8{,}5)^2 + (8-8{,}5)^2 + (10-8{,}5)^2 + (10-8{,}5)^2 + (11-8{,}5)^2\right)} = 2{,}51$$

$$R = \frac{Cov(X,Y)}{s_x s_y} = \frac{1{,}9}{3{,}39 * 2{,}51} = 0{,}22$$

Příklad 1. Výpočet výběrového Pearsonova korelačního koeficientu.

Odhad Pearsonova korelačního koeficientu (*r*) může být doplněn intervalem spolehlivosti. Vzhledem k definovanému rozsahu Pearsonova korelačního koeficientu mezi −1 a +1 je jeho interval spolehlivosti při přiblížení k těmto hranicím asymetrický. Výpočet je založen na tzv. Fisherově transformaci hodnot *r* na skóre *z* (standardizované normální rozdělení):

$$z = 0{,}5 \times ln\left(\frac{1+r}{1-r}\right)$$

dále kalkulujeme směrodatnou odchylku dle vztahu

$$směr.odch. = \sqrt{1/(n-3)}$$

Výsledný 95% interval spolehlivosti pro odhad *z* počítáme dle vztahu

$$z \pm 1{,}96 \times směr.odch.$$

Takto spočtené hranice jsou zpětnou transformací převedeny zpět na hranice 95% intervalu spolehlivosti v původních hodnotách Pearsonova korelačního koeficientu.

Rozsah intervalu spolehlivosti koresponduje se statistickou významností Pearsonova korelačního koeficientu (pokud 95% interval spolehlivosti *r* nezahrnuje hodnotu 0, lze korelační koeficient považovat za statisticky významně odlišný od nuly na hladině p < 0,05).

**Interval spolehlivosti pro Pearsonův korelační koeficient může být kalkulován i dle následujícího vztahu, který pracuje přímo s kvantily Fisher-Snedecorova rozdělení (*F*):**

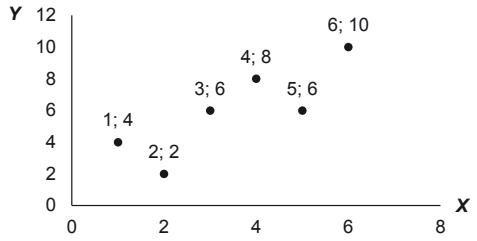
spodní hranice intervalu spolehlivosti: $L_1 = \frac{(1+F_\alpha)r + (1-F_\alpha)}{(1+F_\alpha) + (1-F_\alpha)r}$

horní hranice intervalu spolehlivosti: $L_2 = \frac{(1+F_\alpha)r - (1-F_\alpha)}{(1+F_\alpha) - (1-F_\alpha)r}$

kde $\mathbf{F_\alpha}$ je hodnota Fisher-Snedecorova rozdělení pro $\mathbf{F_{\alpha(2),\nu1,\nu2}}$, např. pro 95% interval spolehlivosti **α(2)** = 0.975); stupně volnosti jsou **ν1 = ν2 = N−2**.

Příklad 2. Výpočet intervalu spolehlivosti výběrového Pearsonova korelačního koeficientu.

**vstupní data**




**potřebné vztahy pro výpočet**

$$z = 0{,}5 \times ln\left(\frac{1+r}{1-r}\right)$$

$$směr.odch. = \sqrt{1/(n-3)}$$

$$(d^*, h^*) = z \pm 1{,}96 \times směr.odch.$$

**výpočet hodnoty *r***

$$R = \frac{Cov(X,Y)}{s_x s_y} = \frac{4{,}4}{1{,}9*2{,}8} = 0{,}8$$

**výpočet Fisherovy transfomace**

$$z = 0{,}5 \times ln\left(\frac{1+r}{1-r}\right) = 0{,}5 \times ln\left(\frac{1+0{,}8}{1-0{,}8}\right) = 1{,}1$$

$$směr.odch. = \sqrt{1/(n-3)} = \sqrt{1/(6-3)} = 0{,}6$$

$$(d^*, h^*) = z \pm 1{,}96 * směr.odch. = 1{,}1 \pm 1{,}96*0{,}6 = (-0{,}08;\ 2{,}3)$$

**zpětná Fisherova transformace**

$$(d,h) = \frac{\exp(2*d^*)-1}{\exp(2*d^*)+1}; \frac{\exp(2*h^*)-1}{\exp(2*h^*)+1} =$$

$$= \frac{\exp(2*(-0{,}08))-1}{\exp(2*(-0{,}08))+1}; \frac{\exp(2*2{,}3)-1}{\exp(2*2{,}3)+1} =$$

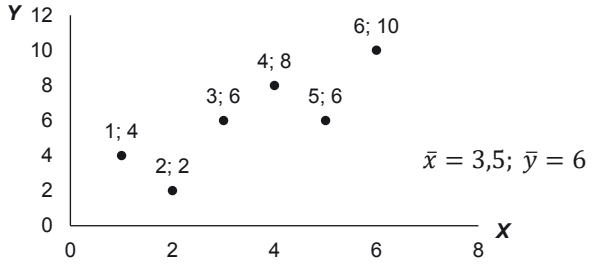
**výsledný 95% interval spolehlivosti *r***

$$= \mathbf{(-0{,}08;\ 0{,}98)}$$

Příklad 2 – pokračování. Výpočet intervalu spolehlivosti výběrového Pearsonova korelačního koeficientu.

Příklad dokládá postup výpočtu Pearsonova korelačního koeficientu a jeho statistické významnosti na reálném datovém souboru o $N = 6$.

**příkladová data**




$\bar{x} = 3{,}5;\ \bar{y} = 6$

**potřebné vztahy a testovaná hypotéza**

$$R = \frac{Cov(X,Y)}{s_x s_y} \qquad t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}}$$

**$H_0$: r = 0**
**$H_1$: r ≠ 0**

$$Cov(X,Y) = \frac{\sum_{i=1}^{N}(x_i - \bar{x}) * (y_i - \bar{y})}{N-1}$$

$$s^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{x})^2 \qquad s = \sqrt{s^2}$$

$$Cov(X,Y) = 4{,}4 \qquad R = \frac{Cov(X,Y)}{s_x s_y} = \frac{4{,}4}{1{,}9*2{,}8} = 0{,}8$$

**výpočet směrodatné odchylky**

$$s_x = \sqrt{\frac{1}{5} * \left((3-3{,}5)^2+(5-3{,}5)^2+(6-3{,}5)^2+(4-3{,}4)^2+(2-3{,}5)^2+(1-3{,}5)^2\right)} = 1{,}9$$

analogicky k $s_x$ se spočítá $s_y = 2{,}8$

**testová statistika a počet stupňů volnosti (df)**

$$t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}} = \frac{0{,}8*\sqrt{6-2}}{\sqrt{1-0{,}8^2}} = 2{,}7$$

$df = N - 2 = 4$

kritická hodnota z tabulek Studentova rozdělení pro 97,5% kvantil a 4 stupně volnosti $t_{0{,}975}(4) = 2{,}776$

**Závěr: jelikož je hodnota testové statistiky *t* menší než kritická hodnota dle Studentova rozdělení (2,7 < 2,776 ) nezamítáme nulovou hypotézu.**

Příklad 3. Testování statistické významnosti Pearsonova korelačního koeficientu na hladině α = 0,05.

citována jako nestandardizovaný ukazatel síly vztahu proměnných.

Výše uvedená nevýhoda kovariance je také důvodem, proč je pro vyjádření síly či „těsnosti" vztahu dvou spojitých proměnných běžně využíván jiný ukazatel, tzv. Pearsonův korelační koeficient (Pearson's correlation coefficient), někdy také označovaný jako párový korelační koeficient. Označuje se *R, r, R(X, Y)* nebo $r_{xy}$. V praxi se běžně vynechává označení Pearsonův a používá se pouze označení korelační koeficient. Korelační koeficient odhadnutý na výběrovém vzorku *N* subjektů je označován jako výběrový korelační koeficient. Jeho cílová populační hodnota je typicky značena řeckým písmenem ρ.

Korelační koeficient je na rozdíl od kovariance statistikou standardizovanou, což pochopíme ze vztahu pro jeho výpočet:

$$R(X,Y) = \frac{\sum_{i=1}^{N}(x_i - \bar{x}) \times (y_i - \bar{y})}{(N-1) \times s_x \times s_y}$$

Je zřejmé, že vztah vychází z výpočtu kovariance, u kterého ve jmenovateli zlomku přibyly hodnoty směrodatných odchylek obou proměnných $s_x$ a $s_y$. Tímto přestala být výsledná hodnota *R* závislá na jednotkách či rozptylu proměnných a může nabývat pouze hodnot v intervalu od –1 do + 1. Dělení směrodatnou odchylkou standardizuje u normálního rozdělení vzdálenost hodnoty $x_i$ od průměru veličiny *X*. Získáváme tak z skóre, např. pro proměnnou *X*:

$$Z = \frac{x_i - \bar{x}}{s_x}$$

Hodnoty *R* blízké nule značí neexistující lineární vztah obou proměnných, hodnoty záporné ukazují na záporný lineární vztah a naopak kladné hodnoty koeficientu ukazují na vztah kladný. Doplníme-li do výše uvedeného vztahu vzorce pro směrodatné odchylky, získáme pro výpočet *R* následující formu zápisu:

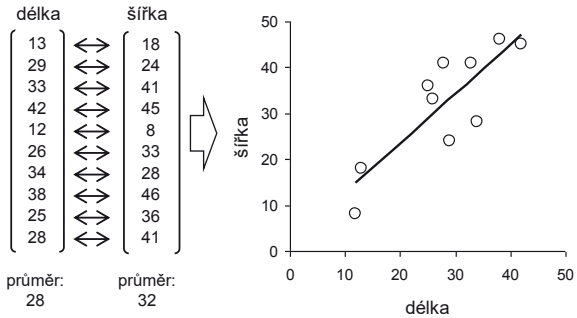
$$R(X,Y) = \frac{\sum_{i=1}^{N}(x_i - \bar{x}) \times (y_i - \bar{y})}{\sqrt{\sum_{i=1}^{N}(x_i - \bar{x})^2 \times \sum_{i=1}^{N}(y_i - \bar{y})^2}}$$

Obecnější formou zápisu je následující vztah, kde *E* a *D* jsou označením výpočtu střední hodnoty a rozptylu:

$$R(X,Y) = \frac{E((X - EX) \times (Y - EY))}{\sqrt{DX} \times \sqrt{DY}}$$

Vlastní výpočet korelačního koeficientu dokládá příklad 1. Jde o bodový odhad hodnoty korelačního koeficientu na daném výběru hodnot o velikosti $N = 6$. Tento výběrový korelační koeficient je možné, tak jako u jiných výběrových statistik, doplnit 100(1 – *α*)% intervalem spolehlivosti (confidence interval), přičemž nejčastěji bývá publikován 95% interval. Postup výpočtu přibližujeme v příkladu 2, ze kterého je patrné, že výpočet zahrnuje poměrně složitou tzv. Fisherovu transformaci. Ačkoli totiž kore-

V analýze hodnotíme vztah mezi délkou a šířkou určité anatomické struktury na výběrovém souboru o velikosti $N = 10$. Cílem je kvantifikovat tento vztah Pearsonovým korelačním koeficientem, jeho 95% intervalem spolehlivosti a otestovat statistickou významnost korelačního koeficientu.




1) výpočet Pearsonova korelačního koeficientu:

$$R = \frac{Cov(X\ Y)}{s_x s_y} = \frac{915}{(852\,.\,1396)^{0{,}5}} = 0{,}839$$

2) výpočet 95% intervalu spolehlivosti Pearsonova korelačního koeficientu pomocí kvantilů Fisher-Snedecorova rozdělení:

$$F_\alpha = 4.443 \quad L_1 = \frac{(1+F_\alpha)r + (1-F_\alpha)}{(1+F_\alpha) + (1-F_\alpha)r} = 0{,}441 \quad L_2 = \frac{(1+F_\alpha)r - (1-F_\alpha)}{(1+F_\alpha) - (1-F_\alpha)r} = 0{,}961$$

3) testování statistické významnosti Pearsonova korelačního koeficientu ($H_0$: $R = 0$):

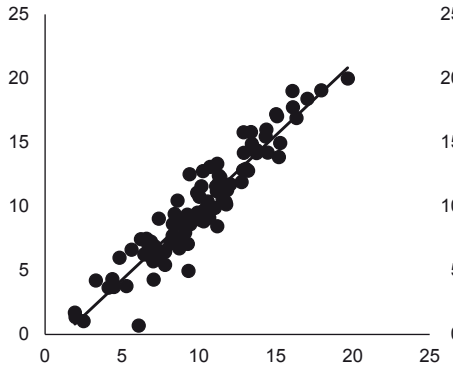
$$t = \left(\frac{r}{\sqrt{1-r^2}}\right)\sqrt{n-2} = 4{,}361 \quad \nu = n-2 = 8 \quad \text{p=0,002}$$

**Závěr: Pearsonův korelační koeficient pro vztah hodnocených proměnných je roven 0,839 s 95% intervalem spolehlivosti 0,441 – 0,961. Korelační koeficient se statisticky významně liší od 0 na hladině významnosti p = 0,002.**

Příklad 4. Výpočet Pearsonova korelačního koeficientu, jeho intervalu spolehlivosti a statistické významnosti.

Pearsonův korelační koeficient ($R$) je standardizovaným ukazatelem síly lineárního vztahu dvou spojitých proměnných. Korelační koeficient je bezrozměrný a může nabývat hodnot od −1 (úplná záporná korelace) do +1 (úplná kladná korelace). Hodnota Pearsonova korelačního koeficientu může být testována na statistickou významnost, kdy nulová hypotéza je $R = 0$, alternativní hypotéza pak $R \neq 0$. Testová statistika má Studentovo rozdělení (t) s n−2 stupni volnosti. Příklady níže ukazují výsledky testu při různých hodnotách Pearsonova korelačního koeficientu. Zkratkou IS je označen interval spolehlivosti, t je hodnota testové statistiky se Studentovým rozdělením.

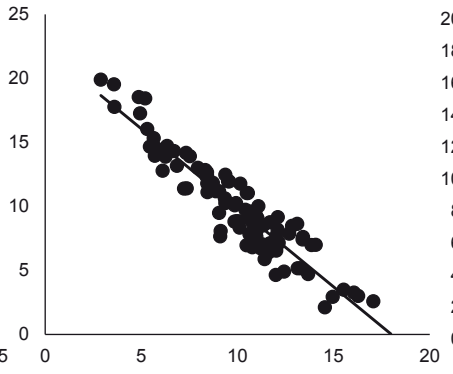
**5a.** významná kladná korelace




Cov = 13,896
**R = 0,945, 95% IS: 0,918; 0,962**
t = 28,462
p < 0,001

statisticky významná kladná korelace

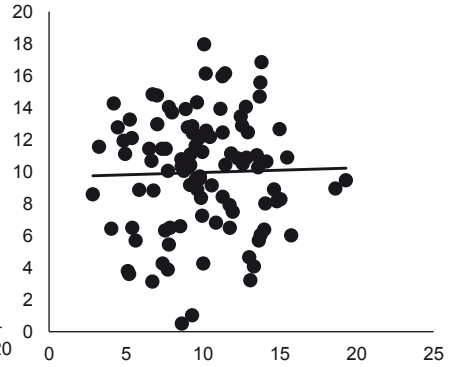
**5b.** významná záporná korelace




Cov = –10,886
**R = –0,941, 95% IS: –0,960; -0,913**
t = –27,447
p < 0,001

statisticky významná záporná korelace

**5c.** nevýznamná korelace

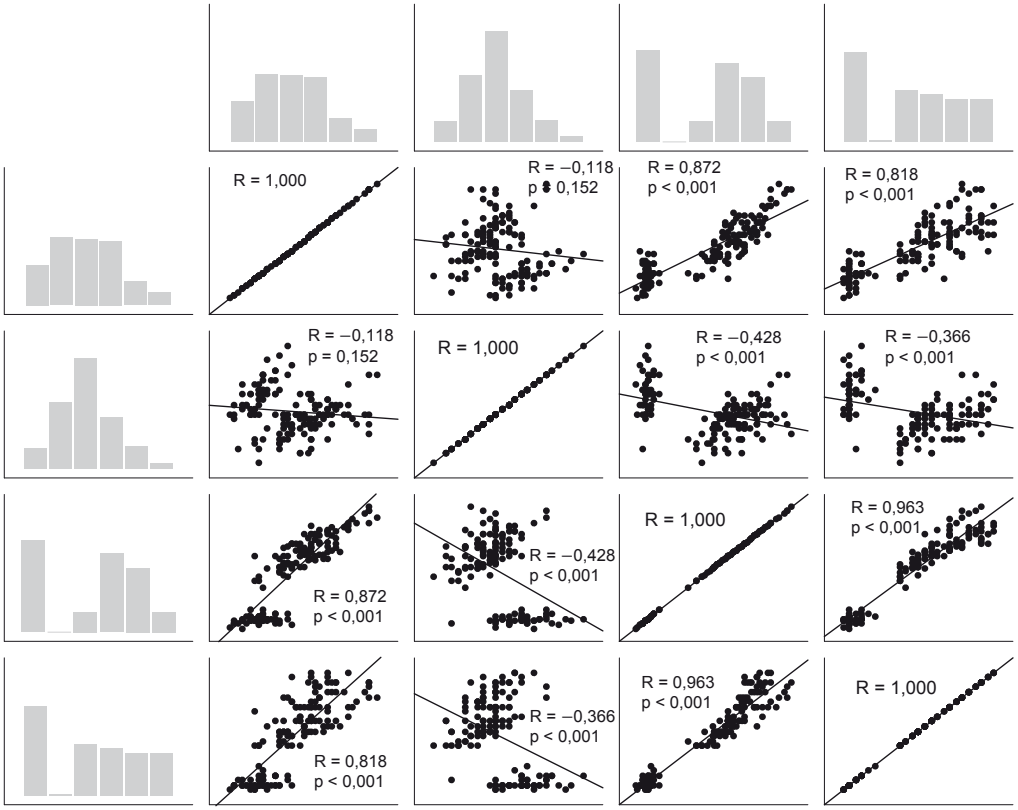



Cov = 0,326
**R = 0,027, 95% IS: –0,171; 0,222**
t = 0,265
p = 0,791

statisticky nevýznamná korelace

Příklad 5. Pearsonův korelační koeficient a jeho statistická významnost.

Pro hodnocení vzájemného vztahu více spojitých proměnných je využívána matice korelačních koeficientů (obdoba matice kovariancí). Jde o čtvercovou matici, jejíž buňky obsahují korelační koeficienty příslušných dvojic proměnných. Matici lze prezentovat i graficky v tzv. korelogramu, jak dokládá ukázka níže.




**Příklad 6. Korelační matice a korelogram.**

lační koeficient vyžaduje splnění předpokladu normality vstupních proměnných *X* a *Y*, hodnoty korelačního koeficientu nemají normální rozdělení. Proto je nutné aplikovat normalizující transformaci, která převádí hodnotu *R* na z skóre.

Interpretace intervalu spolehlivosti pro korelační koeficient se nijak neliší od interpretace pro jakýkoli jiný statistický ukazatel. Tedy např. 95% interval spolehlivosti udává dolní a horní hranici pro hodnoty *R*, v rámci kterých by se vyskytlo 95 výběrových odhadů *R*, pokud bychom odhad 100× nezávisle opakovali. Z hlediska interpretace intervalu spolehlivosti je dále zásadní pozice nuly. Pokud interval spolehlivosti korelačního koeficientu zahrnuje hodnotu nula, nelze tento koeficient označit za významně odlišný od nuly, a tedy nelze potvrdit existenci lineárního vztahu mezi oběma proměnnými.

Samotný interval spolehlivosti by ovšem neměl nahrazovat klasické statistické testy o významnosti korelačního koeficientu. Statistickou významností *R* myslíme situaci, kdy je hodnota *R* statisticky prokazatelně rozdílná od nuly. Testujeme tedy platnost nulové hypotézy $R = 0$, a pokud tu zamítneme pomocí výpočtu testové statistiky, pak platí $R \neq 0$ a mezi oběma veličinami existuje prokazatelný (statisticky významný) lineární vztah. Pro testy týkající se významnosti korelačního koeficientu používáme testovou statistiku Studentova rozdělení (*t*) s $N - 2$ stupni volnosti. Příklad 3 dokumentuje výpočet tohoto statistického testu s výsledkem, který nevedl k zamítnutí nulové hypotézy. Příklad 4 naopak uceleně shrnuje hodnocení odhadu korelace dvou proměnných, která je vysoce statisticky významná. Na tyto příklady navazuje příklad 5, který ukazuje tři kvalitativně rozdílné výsledky korelační analýzy, vč. kalkulovaných 95% intervalů spolehlivosti pro odhad *R* a provedených testů statistické významnosti *R*.

Stejně jako v případě kovariance i u korelační analýzy vzniká v praxi často potřeba vyhodnotit současně korelaci více než dvou proměnných. Při současném zpracování K proměnných hodnotíme korelaci pro $K*(K - 1)/2$ dvojic, které sestavujeme do tzv. korelační matice, jejíž řádky a sloupce jsou věnovány postupně první až *K*-té proměnné. Na průsečíku *i*-tého řádku a *j*-tého sloupce je uvedena korelace *i*-té a *j*-té proměnné. Korelační matice je čtvercová (symetrická podle hlavní diagonály) a na diagonále obsahuje korelační koeficienty rovny jedné, neboť platí, že $R(X, X) = 1$. Příklad 6 dokumentuje grafické znázornění korelační matice, které se nazývá korelogram.

Na závěr tohoto dílu uvádíme několik poznámek, které sice z výše uvedeného výkladu vyplývají, ale měly by být pro svůj význam zdůrazněny:

- Pearsonův korelační koeficient má smysl hodnotit pouze u lineárních (přímkových) vztahů proměnných *X* a *Y*. Pro nelineární vztahy nemá výpočet této korelace žádný smysl.
- Výpočet Pearsonova korelačního koeficientu vyžaduje normální rozdělení obou korelovaných proměnných. Významné

odchylky od normálního rozdělení, zešikmení rozdělení či výskyt odlehlých hodnot, vážným způsobem zkreslují hodnotu korelačního koeficientu a znehodnocují jeho výpočet. Ověření předpokladu normality rozdělení proměnných vstupujících do korelační analýzy je naprostou nutností.

- Test statistické významnosti ověřuje platnost nulové hypotézy $R = 0$ a v případě jejího zamítnutí prokazujeme statisticky významný lineární vztah dvou proměnných. Nic více, nejde o průkaz kauzality vztahu či příčinné závislosti.
- A naopak pokud potvrdíme platnost nulové hypotézy $R = 0$, znamená to, že mezi proměnnými neexistuje prokazatelný lineární vztah. Může však mezi nimi být jiná forma nelineární závislosti. Nekorelovanost neznamená nezávislost.
- Test ověřující platnost hypotézy $R = 0$ je oboustranný. Pokud to daná analýza vyžaduje, můžeme ověřovat i hypotézy jednostranné, jako např. $R < 0$ nebo $R > 0$.