Analýza dat v neurologii

LXXII. Statistické hodnocení Pearsonova korelačního koeficientu v příkladech

V tomto díle seriálu reagujeme na dotazy několika čtenářů k předchozímu dílu, který uvedl postupy pro výpočet Pearsonova korelačního koeficientu (*R*) a pro hodnocení jeho statistické významnosti. Nejprve připomeňme, že pomocí tohoto koeficientu měříme sílu lineární (přímkové) závislosti dvou náhodných veličin s dvourozměrným normálním rozdělením hodnot. Formou příkladů zde rozvedeme postupy testování statistické významnosti koeficientu *R*, které zahrnují jednak statistický test nulové hypotézy $R = 0$ a dále výpočet intervalu spolehlivosti pro odhad hodnoty *R*.

Ze vztahu pro výpočet Pearsonova korelačního koeficientu vyplývá, že jde o statistiku standardizovanou, která může nabývat pouze hodnot od –1 do 1. 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:

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

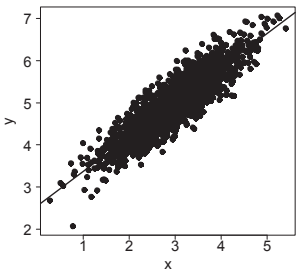
L. Dušek, T. Pavlík,
J. Jarkovský, J. Koptíková

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

N – velikost výběrového souboru; *R* – Pearsonův korelační koeficient; $IS_{0,95}$ – interval spolehlivosti odhadu Pearsonova korelačního koeficientu; $\bar{x}, \bar{y}$ – aritmetický průměr proměnných X, Y; s_x, s_y – směrodatná odchylka proměnných X, Y; t – hodnota testové statistiky Studentova rozdělení; *p* – hladina statistické významnosti odhadu Pearsonova korelačního koeficientu (test hypotézy $R = 0$)

Příklad 1a
Velký výběrový soubor a malá variabilita *X* a *Y* vedou k vysoké významnosti *R*. Tomu odpovídá úzký interval spolehlivosti *R*, který nezahrnuje nulu.

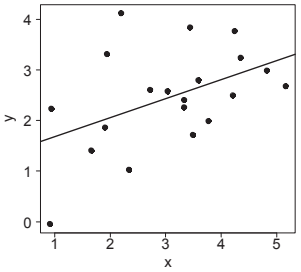


$N = 2\,000$
$\bar{x} = 3{,}00$
$\bar{y} = 5{,}00$
$s_x = 0{,}73$
$s_y = 0{,}62$

$R = 0{,}89$
$t = 86{,}56$
$p < 0{,}001$

$IS_{0,95} = (0{,}88; 0{,}9)$

Příklad 1b
Malý výběrový soubor s hodnotou $R = 0{,}46$. Statistická významnost odhadu *R* je hraniční ($p = 0{,}042$), čemuž odpovídá širší interval spolehlivosti s hraniční hodnotou blízkou nule.

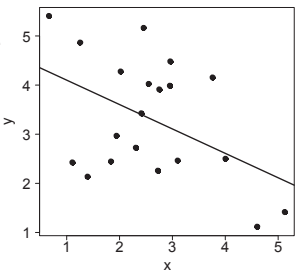


$N = 20$
$\bar{x} = 3{,}07$
$\bar{y} = 2{,}46$
$s_x = 1{,}22$
$s_y = 1$

$R = 0{,}46$
$t = 2{,}19$
$p = 0{,}042$

$IS_{0,95} = (0{,}02; 0{,}75)$

Příklad 1c
Malý výběrový soubor a relativně vysoká variabilita proměnných *X* a *Y*. Příklad opakuje situaci v příkladu 1b, avšak se zápornou korelací proměnných X a Y.



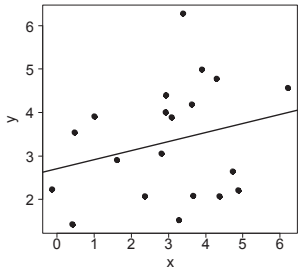
$N = 20$
$\bar{x} = 2{,}6$
$\bar{y} = 3{,}3$
$s_x = 1{,}15$
$s_y = 0{,}23$

$R = -0{,}46$
$t = -2{,}21$
$p = 0{,}04$

$IS_{0,95} = (-0{,}76; -0{,}02)$

Příklad 1. Korelační analýza v příkladech dokumentujících širokou škálu typů vstupních dat.

Příklad 1d
Malý výběrový soubor a relativně vysoká variabilita proměnných *X* a *Y* vedou k nízké a statisticky nevýznamné hodnotě *R*. Interval spolehlivosti odhadu *R* je široký a zahrnuje nulu.

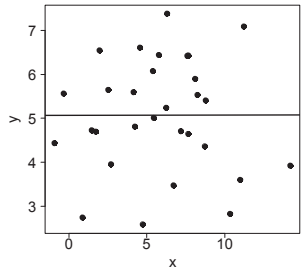


$N = 20$
$\bar{x} = 3{,}00$
$\bar{y} = 3{,}33$
$s_x = 1{,}65$
$s_y = 1{,}31$

$R = 0{,}26$
$t = 1{,}27$
$p = 0{,}259$

$IS_{0{,}95} = (-0{,}21; 0{,}63)$

Příklad 1e
Výběrový soubor s vysokou variabilitou proměnných *X* a *Y* způsobenou i částečně odlehlými hodnotami obou proměnných. Hodnota *R* je velmi nízká, statistiky nevýznamná. Interval spolehlivosti *R* zahrnuje nulu.

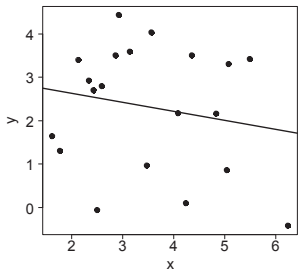


$N = 30$
$\bar{x} = 5{,}82$
$\bar{y} = 5{,}07$
$s_x = 3{,}6$
$s_y = 1{,}28$

$R = 0{,}002$
$t = 0{,}02$
$p = 0{,}98$

$IS_{0{,}95} = (-0{,}36; 0{,}36)$

Příklad 1f
Malý výběrový soubor a relativně vysoká variabilita proměnných *X* a *Y* vedou k nízké a statisticky nevýznamné hodnotě *R*. Příklad opakuje situace v příkladu 1d, avšak se zápornou hodnotou *R*.



$N = 20$
$\bar{x} = 3{,}54$
$\bar{y} = 2{,}31$
$s_x = 1{,}32$
$s_y = 1{,}44$

$R = -0{,}19$
$t = -0{,}82$
$p = 0{,}424$

$IS_{0{,}95} = (-0{,}58; 0{,}28)$

Příklad 1 – pokračování. Korelační analýza v příkladech dokumentujících širokou škálu typů vstupních dat.

V uvedeném vztahu jsou x_i, y_i 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* a s_x, s_y jsou hodnoty směrodatných odchylek obou proměnných.

Statistickou významnost Pearsonova koeficientu hodnotíme pomocí testové statistiky, se Studentovým rozdělením hodnot (t), která ma N – 2 stupňů volnosti. Konkrétně počítáme hodnotu *t* dle následujícího vztahu:

$$t = \frac{R\sqrt{N-2}}{\sqrt{1-R^2}}$$

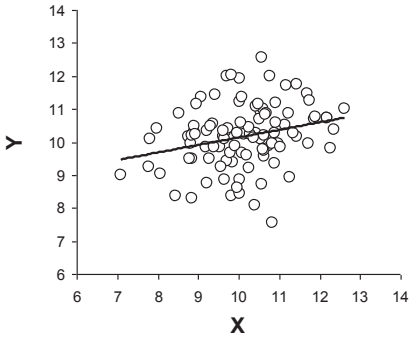
Je zřejmé, že hodnotu testové statistiky ovlivňuje vedle samotné hodnoty R také velikost výběrového souboru, na kterém je hodnota korelačního koeficientu odhadována. Považujeme za nutné tento fakt zdůraznit, neboť hodnota korelačního koeficientu bývá často tendenčně posuzována pouze podle její absolutní hodnoty, tedy bez uvedení statistické významnosti. Je ovšem přirozené, že máme tendenci posuzovat míru korelace již podle samotné hodnoty R, neboť tato má jasně dánu minimální a maximální možnou hodnotu. Hodnotu korelačního koeficientu 0,9 tak považujeme za vysokou a naopak hodnotu 0,2 za nízkou. Avšak chceme-li sílu a průkaznost korelace dvou proměnných posoudit skutečně exaktně, pak musíme současně zvažovat nejen velikost korelačního koeficientu, ale i jeho statistickou významnost. Teoreticky totiž mohou při hodnocení významnosti korelace nastat různě rozporuplné situace, při kterých je třeba interpretaci výsledků analýzy pečlivě zvážit. Při analýze velkého souboru můžeme prokázat jako statisticky významný (významně odlišný od nuly) i korelační koeficient s relativně malou hodnotou. A naopak i velmi vysoká hodnota R nemusí být prokázána jako statisticky významně odlišná od nuly, jde-li o analýzu velmi malého souboru dat. V obou případech je na zvážení analytika, jak silně bude korelaci interpretovat. Přitom neexistují žádná paušálně daná pravidla, jak v dané situaci postupovat. Záleží na zadání dané studie, okolnostech výběru vzorku a jeho reprezentativnosti a v neposlední řadě i na odborném úsudku autora analýzy. Lze však doporučit následující tři pomocné postupy, které interpretaci usnadní a umožní také budoucím čtenářům lépe posoudit skutečný význam zjištěné korelace:

1. **Grafická dokumentace.** Korelační analýzu je vždy možné doplnit bodovým diagramem s hodnotami proměnných *X* a *Y*. Toto doporučujeme zejména, pokud nastane některá ze sporných situací popsaných výše. Autor analýzy i její čtenáři tak mohou snadno přímo posoudit rozdělení hodnot proměnných *X* a *Y*, a také interpretační význam zjištěné korelace.
2. **Výpočet koeficientu determinace.** Tato veličina udává, jaký podíl z celkové variability proměnné *Y* vysvětluje přímkový vztah s proměnnou *X*, nebo naopak jaký podíl variability *X* je vysvětlen lineárním

Příklad dokumentuje vliv hodnoty Pearsonova korelačního koeficientu na šířku jeho intervalu spolehlivosti (při stejné velikosti vzorku a při statistické významnosti korelačního koeficientu). Vyšším hodnotám korelačního koeficientu odpovídá užší interval spolehlivosti.

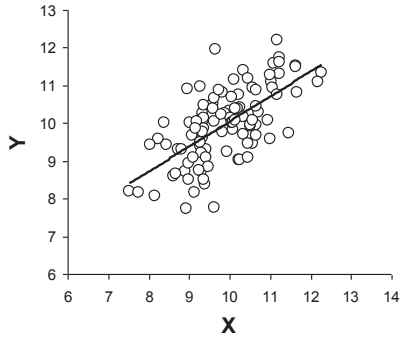
Příklad 2a
Nízké hodnotě *R* odpovídá široký interval spolehlivosti, s hranicí blízkou nule.

N = 100
R = 0,261
95% *IS*: 0,068; 0,435



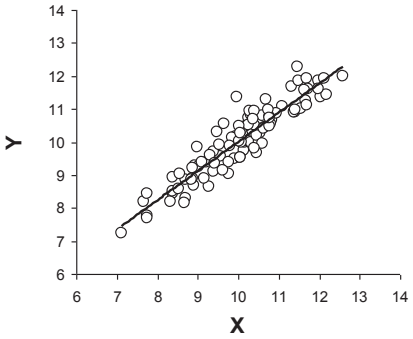
Příklad 2b
Středně silná korelace proměnných *X* a *Y*, hodnota *R* je užší než v příkladu 2a.

N = 100
R = 0,644
95% *IS*: 0,512; 0,746



Příklad 2c
Velmi silná korelace proměnných *X* a *Y*, interval spolehlivosti odhadu *R* je úzký.

N = 100
R = 0,928
95% *IS*: 0,895; 0,951

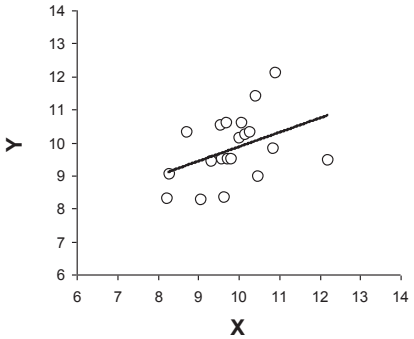


Příklad 2. Interval spolehlivosti při různých hodnotách Pearsonova korelačního koeficientu.

Příklad dokumentuje vliv velikosti vzorku jak na šířku intervalu spolehlivosti (*IS*) odhadu Pearsonova korelačního koeficientu (*R*), tak na statistickou významnost (*p*) při testování hypotézy o korelačním koeficientu (hypotéza: *R* = 0). Příklady dokumentují srovnatelné situace s obdobnou hodnotou korelačního koeficientu 0,4. Se zvětšující se velikostí vzorku se silně zužuje interval spolehlivosti odhadu *R* a tomu odpovídá rostoucí statistická významnost korelačního koeficientu.

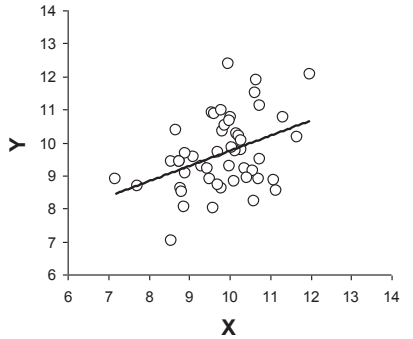
Příklad 3a
Malá velikost vzorku, korelační koeficient není statisticky významný, interval spolehlivosti je široký a obsahuje hodnotu nula.

N = 20
R = 0,403
95% *IS*: –0,048; 0,718
p = 0,078



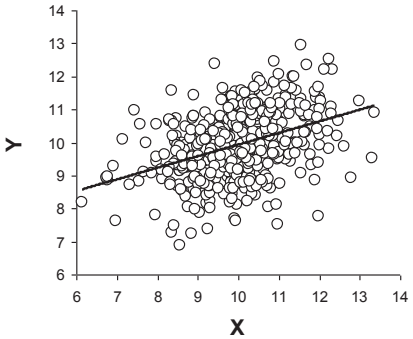
Příklad 3b
Střední velikost vzorku – korelační koeficient je statisticky významný, interval spolehlivosti neobsahuje hodnotu 0 a je užší než v příkladu 3a.

N = 50
R = 0,393
95% *IS*: 0,129; 0,605
p = 0,005



Příklad 3c
Velká velikost vzorku, korelační koeficient je vysoce statisticky významný, interval spolehlivosti je velmi úzký a neobsahuje hodnotu nula.

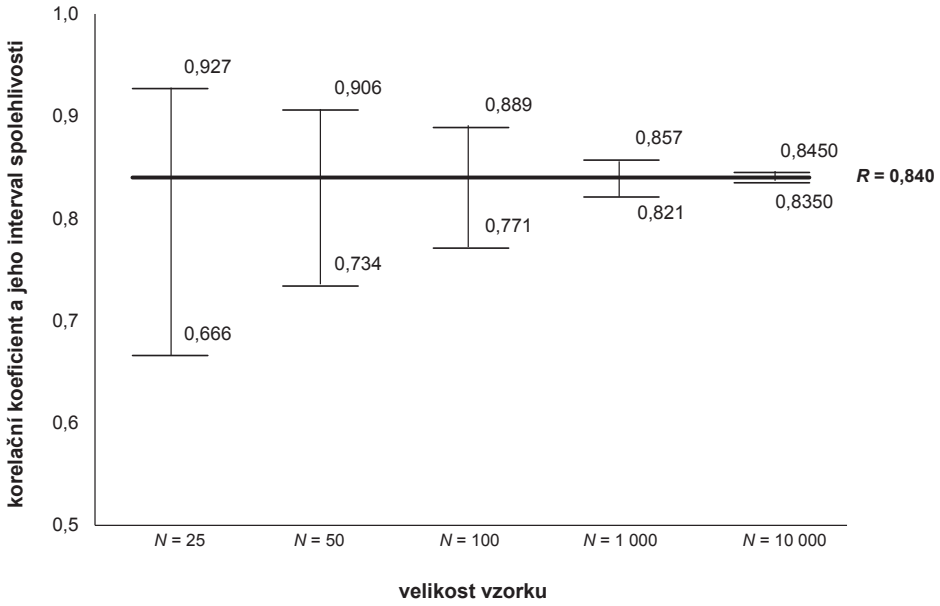
N = 500
R = 0,391
95% IS: 0,314; 0,463
p < 0,001



Příklad 3. Vliv velikosti vzorku na šířku intervalu spolehlivosti Pearsonova korelačního koeficientu; implikace pro hodnocení statistické významnosti korelačního koeficientu.

vztahem s proměnnou *Y*. Koeficient determinace jednoduše spočítáme jako druhou mocninu korelačního koeficientu (R^2). Obvykle se násobí 100 a výsledek je pak uváděn v procentech. V případě, že proměnné *X* a *Y* mají mezi sebou absolutní lineární závislost a jejich body v *X*-*Y* diagramu přesně leží na přímce, pak při znalosti hodnot jedné proměnné můžeme přesně vypočítat hodnotu proměnné druhé. Hodnota korelačního koeficientu je maximální možná (–1 nebo +1) a koeficient determi-

S rostoucí velikostí vzorku se při stejném korelačním koeficientu (*R*) významně zužuje interval spolehlivosti odhadu *R*. Z grafu je rovněž patrná asymetrie intervalu spolehlivosti, který v případě Pearsonova korelačního koeficientu nemůže překročit hranice dané minimální a maximální možnou hodnotou *R* (–1; 1).



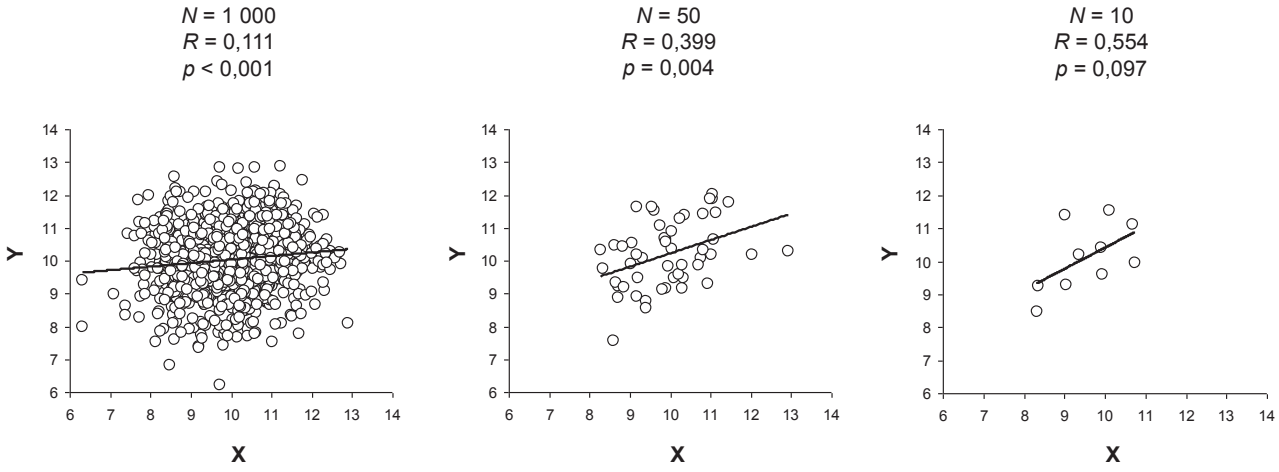
Příklad 4. Vliv velikosti vzorku na šířku intervalu spolehlivosti (95% *IS*).

Příklad dokumentuje vliv velikosti vzorku na statistickou významnost Pearsonova korelačního koeficientu (*R*). S rostoucí velikostí vzorku snáze prokazujeme statistickou významnost *R*, a to i při relativně nízké absolutní hodnotě *R*. A naopak, malá velikost vzorku vede k statisticky nevýznamnému korelačnímu koeficientu i při jeho vysoké absolutní hodnotě.

Příklad 5a
Velmi velký vzorek vede k vysoce statisticky významnému korelačnímu koeficientu, a to i při jeho nízké malé absolutní hodnotě.

Příklad 5b
Střední velikost vzorku a statistická významnost střední hodnoty korelačního koeficientu.

Příklad 5c
Malá velikost vzorku způsobuje statistickou nevýznamnost i relativně vysoké hodnoty korelačního koeficientu.



Příklad 5. Vliv velikosti vzorku na statistickou významnost Pearsonova korelačního koeficientu.

nace je 100 %. Při hodnotě *R* = 0,8 je koeficient determinace 64 % a při *R* = 0,2 již pouze 4 %.

3. **Výpočet intervalu spolehlivosti *R*.** Šířka intervalu spolehlivosti velmi návodně ukazuje míru spolehlivosti odhadu korelačního koeficientu. Vzhledem k úzké provázanosti mezi výpočtem intervalu spolehlivosti a testováním sta-

tistické významnosti stojí za pozornost možnost přímo využít interval spolehlivosti k interpretaci významnosti R. V případě, že 95% interval spolehlivosti nezahrnuje nulu, lze tento výsledek považovat za ekvivalentní zamítnutí nulové hypotézy $R = 0$ na hladině významnosti $\alpha = 0{,}05$.

Příklady 1–5 připravené pro tento díl seriálu dokumentují různé výsledky korelačních analýz a vliv velikosti vzorku na konečný výsledek a jeho interpretaci.