Kapitola 8.: Jednoduchá korelační analýza

Cíl kapitoly

Po prostudování této kapitoly budete umět
- provádět test pořadové nezávislosti dvou náhodných veličin ordinálního typu pomocí Spearmanova koeficientu pořadové korelace
- testovat hypotézu o nezávislosti dvou náhodných veličin intervalového či poměrového typu, které se řídí dvourozměrným normálním rozložením

Časová zátěž

Na prostudování této kapitoly a splnění úkolů s ní spojených budete potřebovat asi 9 hodin studia.

8.1. Motivace

Uvažme náhodné veličiny X, Y, které jsou aspoň ordinálního typu. Mezi těmito náhodnými veličinami může existovat různý vztah:

- Deterministická (funkční) závislost: jedna náhodná veličina je spjata s druhou náhodnou veličinou funkční závislostí vyjádřenou předpisem Y = g(X). Např. X je poloměr náhodně vybrané sériově vyráběné kuličky do kuličkových ložisek, $Y = \frac{4}{3}\pi X^3$ je objem této kuličky. Každé realizaci náhodné veličiny X (vysvětlující proměnná) je přiřazena právě jedna realizace náhodné veličiny Y (vysvětlovaná proměnná).

- Stochastická závislost: jedna náhodná veličina ovlivňuje v různé míře druhou náhodnou veličinu. Např. X je věk pracovníka v letech, Y je počet dnů absence za rok. Každé realizaci náhodné veličiny X může být přiřazeno více realizací náhodné veličiny Y. Závislost může být jednostranná i oboustranná.

- Stochastická nezávislost: náhodné veličiny se navzájem neovlivňují. Např. házíme-li naráz dvěma kostkami a označíme X počet ok padlých na jedné kostce a Y počet ok padlých na druhé kostce, pak náhodné veličiny X, Y jsou stochasticky nezávislé.

Úkolem korelační analýzy je právě zkoumání stochastické závislosti náhodných veličin X, Y a měření těsnosti této závislosti. Přitom se požaduje, aby míra těsnosti stochastické závislosti nabývala hodnot z určitého přesně vymezeného intervalu, uvnitř tohoto intervalu monotónně rostla se zvyšováním stupně závislosti a nebyla závislá na velikosti hodnot či používaných jednotkách zkoumaných veličin. Tyto požadavky splňuje Spearmanův koeficient pořadové korelace a Pearsonův koeficient korelace.

Při zkoumání závislosti je velmi důležité provést logický rozbor problému. Nemá smysl se zabývat hledáním závislosti v případech, když
- z logických důvodů nemůže existovat,
- závislost je způsobena formálními vztahy mezi veličinami,
- soubor dvourozměrných dat je nehomogenní,
- závislost je způsobena společnou příčinou.

8.2. Testování nezávislosti ordinálních veličin

8.2.1. Popis testu

Nechť X,Y jsou dvě ordinální náhodné veličiny (tj. obsahová interpretace je možná jenom u relace rovnosti a relace uspořádání). Pořídíme dvourozměrný náhodný výběr $(X_1, Y_1), ..., (X_n, Y_n)$ z rozložení, jímž se řídí náhodný vektor (X, Y). Označíme R_i pořadí náhodné veličiny X_i a Q_i pořadí náhodné veličiny Y_i, i = 1, ..., n.

Na hladině významnosti α testujeme hypotézu H_0: X, Y jsou pořadově nezávislé náhodné veličiny proti

- oboustranné alternativě H_1: X, Y jsou pořadově závislé náhodné veličiny
- levostranné alternativě H_1: mezi X a Y existuje nepřímá pořadová závislost
- pravostranné alternativě H_1: mezi X a Y existuje přímá pořadová závislost).

Testová statistika se nazývá Spearmanův koeficient pořadové korelace a má tvar:

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

Tento koeficient nabývá hodnot mezi –1 a 1. Čím je bližší 1, tím je silnější přímá pořadová závislost mezi veličinami X a Y, čím je bližší –1, tím je silnější nepřímá pořadová závislost mezi veličinami X a Y. Teoretická hodnota Spearmanova koeficientu se značí ρ_S.

Nulovou hypotézu zamítáme na hladině významnosti α ve prospěch

- oboustranné alternativy, když $|r_S| \geq r_{S,1-\alpha}(n)$
- levostranné alternativy, když $r_S \leq - r_{S,1-2\alpha}(n)$
- pravostranné alternativy, když $r_S \geq r_{S,1-2\alpha}(n)$,

kde $r_{S,1-\alpha}(n)$ je kritická hodnota, kterou pro α = 0,05 nebo 0,01 a n ≤ 30 najdeme v tabulkách. Pozor – kritické hodnoty pro jednostranné alternativy se v běžně dostupných tabulkách nenajdou.

8.2.2. Asymptotické varianty testu

Pro n > 20 lze použít testovou statistiku $T_0 = \frac{r_S\sqrt{n-2}}{\sqrt{1-r_S^2}}$, která se v případě platnosti nulové hypotézy asymptoticky řídí rozložením t(n-2).

Kritický obor pro oboustrannou alternativu: $W = \left(-\infty, -t_{1-\alpha/2}(n-2)\right\rangle \cup \left\langle t_{1-\alpha/2}(n-2), \infty\right)$

Kritický obor pro levostrannou alternativu: $W = \left(-\infty, -t_{1-\alpha}(n-2)\right\rangle$

Kritický obor pro pravostrannou alternativu: $W = \left\langle t_{1-\alpha}(n-2), \infty\right)$.

Hypotézu o pořadové nezávislosti náhodných veličin X, Y zamítáme na asymptotické hladině významnosti α, když $t_0 \in W$.

Upozornění: Systém STATISTICA používá tuto variantu testu pořadové nezávislosti bez ohledu na rozsah náhodného výběru.

Pro n > 30 lze použít testovou statistiku $r_s\sqrt{n-1}$. Platí-li H_0, pak $r_s\sqrt{n-1} \approx N(0, 1)$.

Nulovou hypotézu tedy zamítáme na asymptotické hladině významnosti α ve prospěch

oboustranné alternativy, když $r_S\sqrt{n-1} \in \left(-\infty, -u_{1-\alpha/2}\right\rangle \cup \left\langle u_{1-\alpha/2}, \infty\right)$,

levostranné alternativy, když $r_S\sqrt{n-1} \in \left(-\infty, -u_{1-\alpha}\right\rangle$,

pravostranné alternativy, když $r_S\sqrt{n-1} \in \left\langle u_{1-\alpha}, \infty\right)$

8.2.3. Příklad

Dva lékaři hodnotili stav sedmi pacientů po témž chirurgickém zákroku. Postupovali tak, že nejvyšší pořadí dostal nejtěžší případ.

Číslo pacienta	1	2	3	4	5	6	7
Hodnocení 1. lékaře	4	1	6	5	3	2	7
Hodnocení 2. lékaře	4	2	5	6	1	3	7

Vypočtěte Spearmanův koeficient r_S a na hladině významnosti 0,05 testujte hypotézu, že hodnocení obou lékařů jsou pořadově nezávislá.

Řešení:
Na hladině významnosti 0,05 testujeme hypotézu H_0: X, Y jsou pořadově nezávislé náhodné veličiny proti oboustranné alternativě H_1: X, Y jsou pořadově závislé náhodné veličiny.
V tomto příkladě přímo známe pořadí R_i (tj. hodnocení 1. lékaře) a pořadí Q_i (tj. hodnocení 2. lékaře). Vypočteme

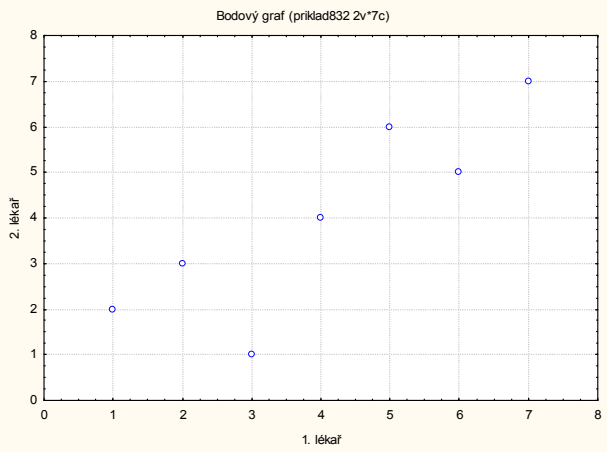
$$r_S = 1 - \frac{6}{7\left(7^2-1\right)}\left[(4-4)^2 + (1-2)^2 + (6-5)^2 + (5-6)^2 + (3-1)^2 + (2-3)^2 + (7-7)^2\right] = 0{,}857 .$$

Kritická hodnota: $r_{S,0,95}(7) = 0{,}745$. Protože $0{,}857 \geq 0{,}745$, nulovou hypotézu zamítáme na hladině významnosti 0,05. S rizikem omylu nejvýše 0,05 jsme tedy prokázali, že hodnocení obou lékařů jsou pořadově závislá.

Řešení pomocí sytému STATISTICA:
Otevřeme nový datový soubor se dvěma proměnnými X, Y a sedmi případy. Do těchto proměnných zapíšeme zjištěná hodnocení.
Statistika – Neparametrická statistika – Korelace – OK, Vytvořit Detailní report, Proměnné - 1. seznam proměnných X, 2. seznam proměnných Y – OK – Spearman R.

	Spearmanovy korelace (dva lekari.sta) ChD vynechány párově Označ. korelace jsou významné na hl. p <,05000			
Dvojice proměnných	Počet plat.	Spearman R	t(N-2)	Úroveň p
X & Y	7	0,857143	3,721042	0,013697

Spearmanův koeficient korelace nabyl hodnoty 0,857143, asymptotická testová statistika se realizovala číslem 3,721042, odpovídající p-hodnota je 0,013697, tedy na asymptotické hladině významnosti 0,05 zamítáme hypotézu o pořadové nezávislosti hodnocení obou lékařů. Pokud bychom chtěli provést přesný test, nikoliv asymptotický test, museli bychom použít statistické tabulky a vyhledat v nich kritickou hodnotu $r_{S,0,95}(7)$ – viz výše.
Výpočet ještě doplníme dvourozměrným tečkovým diagramem. Grafy – Bodové grafy - vypneme Typ proložení – Proměnné – X, Y - OK, OK.



Vidíme, že s rostoucím hodnocením 1. lékaře roste hodnocení 2. lékaře a naopak. Tedy mezi oběma proměnnými existuje určitý stupeň přímé pořadové závislosti.

8.3. Testování nezávislosti intervalových či poměrových veličin

8.3.1. Pearsonův koeficient korelace

V teorii pravděpodobnosti byl zaveden Pearsonův koeficient korelace náhodných veličin X, Y (které jsou aspoň intervalového typu) vztahem

$$R(X,Y)=\frac{C(X,Y)}{\sqrt{D(X)}\sqrt{D(Y)}} \text{ pro } \sqrt{D(X)}\sqrt{D(Y)}>0, =0 \text{ jinak} .$$

Připomeneme jeho vlastnosti:

a) R(X,X) = 1

b) R(X,Y) = R(Y,X)

c) R(a + bX, c + dY) = sgn(bd)R(X,Y)

d) $-1 \le R(X,Y) \le 1$ a rovnosti je dosaženo tehdy a jen tehdy, když existují reálné konstanty a, b, $b \neq 0$ tak, že P(Y = a + bX) = 1, přičemž R(X,Y) = 1 pro b > 0 a R(X,Y) = -1 pro b < 0.

Z těchto vlastností plyne, že R(X,Y) je vhodnou mírou těsnosti lineárního vztahu náhodných veličin X, Y.

Pomocí koeficientu korelace zavádíme nekorelovanost náhodných veličin X, Y.

Je-li R(X, Y) = 0, pak řekneme, že náhodné veličiny jsou nekorelované. (Znamená to, že mezi X a Y neexistuje žádná lineární závislost. Jsou-li náhodné veličiny X,Y stochasticky nezávislé, pak jsou samozřejmě i nekorelované.)

Je-li R(X, Y) > 0, pak řekneme, že náhodné veličiny jsou kladně korelované. (Znamená to, že s růstem hodnot veličiny X rostou hodnoty veličiny Y a s poklesem hodnot veličiny X klesají hodnoty veličiny Y.)

Je-li R(X, Y) < 0, pak řekneme, že náhodné veličiny jsou záporně korelované. (Znamená to, že s růstem hodnot veličiny X klesají hodnoty veličiny Y a s poklesem hodnot veličiny X rostou hodnoty veličiny Y.)

8.3.2. Výběrový koeficient korelace

R(X,Y) většinou nemůžeme počítat přímo, protože to vyžaduje znalost simultánního rozložení náhodného vektoru (X, Y). V praxi jsme zpravidla odkázáni na náhodný výběr rozsahu n z dvourozměrného rozložení daného distribuční funkcí Φ(x,y). Z tohoto dvourozměrného náhodného výběru můžeme stanovit:

výběrové průměry $M_1 = \frac{1}{n}\sum_{i=1}^{n} X_i$, $M_2 = \frac{1}{n}\sum_{i=1}^{n} Y_i$,

výběrové rozptyly $S_1{}^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - M_1)^2$, $S_2{}^2 = \frac{1}{n-1}\sum_{i=1}^{n}(Y_i - M_2)^2$,

výběrovou kovarianci $S_{12} = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - M_1)(Y_i - M_2)$ a s jejich pomocí zavedeme

výběrový koeficient korelace $R_{12} = \frac{S_{12}}{S_1 S_2}$ (pro $S_1S_2 > 0$). Vlastnosti a), b), c), d) koeficientu korelace se přenášejí i na výběrový koeficient korelace.

8.3.3. Koeficient korelace dvourozměrného normálního rozložení

Nechť náhodný vektor (X, Y) má dvourozměrné normální rozložení s hustotou

$\varphi(x,y) = \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}} e^{-\frac{1}{2(1-\rho^2)}\left[\left(\frac{x-\mu_1}{\sigma_1}\right)^2 - 2\rho\frac{(x-\mu_1)(y-\mu_2)}{\sigma_1\ \sigma_2} + \left(\frac{y-\mu_2}{\sigma_2}\right)^2\right]}$, přičemž

$\mu_1 = E(X)$, $\mu_2 = E(Y)$, $\sigma_1{}^2 = D(X)$, $\sigma_2{}^2 = D(Y)$, $\rho = R(X,Y)$.

Marginální hustoty jsou:

$\varphi_1(x) = \frac{1}{\sigma_1\sqrt{2\pi}} e^{-\frac{(x-\mu_1)^2}{2\sigma_1{}^2}}$, $\varphi_2(y) = \frac{1}{\sigma_2\sqrt{2\pi}} e^{-\frac{(y-\mu_2)^2}{2\sigma_2{}^2}}$.

Je-li $\rho = 0$, pak pro $\forall(x,y) \in R^2 : \varphi(x,y) = \varphi_1(x)\varphi_2(y)$, tedy náhodné veličiny X, Y jsou stochasticky nezávislé. Jinými slovy: stochastická nezávislost složek X, Y normálně rozloženého náhodného vektoru je ekvivalentní jejich nekorelovanosti.

Je-li $\rho \neq 0$, jsou náhodné veličiny X, Y stochasticky závislé. Je-li $\rho > 0$, říkáme, že jsou kladně korelované, je-li $\rho < 0$, říkáme, že jsou záporně korelované.

Upozornění: V dalším textu budeme předpokládat, že náhodný výběr $(X_1, Y_1), ..., (X_n, Y_n)$ pochází z dvourozměrného normálního rozložení s parametry $\mu_1, \mu_2, \sigma_1{}^2, \sigma_2{}^2, \rho$. Předpoklad dvourozměrné normality lze orientačně ověřit pomocí dvourozměrného tečkového diagramu: tečky by měly zhruba rovnoměrně vyplnit vnitřek elipsovitého obrazce, neboť vrstevnice hustoty dvourozměrného normálního rozložení jsou elipsy.

8.3.4. Testování hypotézy o nezávislosti

Na hladině významnosti α testujeme hypotézu H_0: X, Y jsou stochasticky nezávislé náhodné veličiny (tj. $\rho = 0$) proti

- oboustranné alternativě H_1: X, Y nejsou stochasticky nezávislé náhodné veličiny (tj. $\rho \neq 0$)
- levostranné alternativě H_1: X, Y jsou záporně korelované náhodné veličiny (tj. $\rho < 0$)
- pravostranné alternativě H_1: X, Y jsou kladně korelované náhodné veličiny (tj. $\rho > 0$).

Testová statistika má tvar: $T_0 = \frac{R_{12}\sqrt{n-2}}{\sqrt{1-R_{12}{}^2}}$. Platí-li nulová hypotéza, pak $T \sim t(n-2)$.

Kritický obor pro test H_0 proti

- oboustranné alternativě: $W = (-\infty, -t_{1-\alpha/2}(n-2)\rangle \cup \langle t_{1-\alpha/2}(n-2), \infty)$,
- levostranné alternativě: $W = (-\infty, -t_{1-\alpha}(n-2)\rangle$,
- pravostranné alternativě: $W = \langle t_{1-\alpha}(n-2), \infty)$.

H_0 zamítáme na hladině významnosti α, když $t_0 \in W$.

Není-li splněn předoklad dvourozměrné normality, použijeme Spearmanův koeficient pořadové korelace.

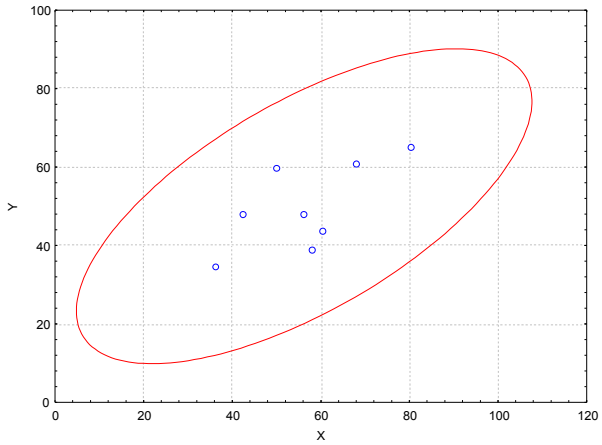
8.3.5. Příklad

Máme k dispozici výsledky testů ze dvou předmětů zjištěné u osmi náhodně vybraných studentů určitého oboru.

Číslo studenta	1	2	3	4	5	6	7	8
Počet bodů v 1. testu	80	50	36	58	42	60	56	68
Počet bodů ve 2. testu	65	60	35	39	48	44	48	61

Na hladině významnosti 0,05 testujte hypotézu, že výsledky obou testů nejsou kladně korelované.

Řešení:
Nejprve se musíme přesvědčit, že uvedené výsledky lze považovat za realizace náhodného výběru z dvourozměrného normálního rozložení. Lze tak učinit orientačně pomocí dvourozměrného tečkového diagramu. Tečky by měly vytvořit elipsovitý obrazec.



Obrázek svědčí o tom, že předpoklad dvourozměrné normality je oprávněný a že mezi počty bodů z 1. a 2. testu bude existovat určitý stupeň přímé lineární závislosti.

Testujeme H_0: $\rho = 0$ proti pravostranné alternativě H_1: $\rho > 0$.

Výpočtem zjistíme: $r_{12} = 0{,}6668$, $t_0 = 2{,}1917$. V tabulkách najdeme $t_{0,95}(6) = 1{,}9432$. Kritický obor: $W = \langle 1{,}9432; \infty)$. Protože $t_0 \in W$, hypotézu o neexistenci kladné korelace výsledků z 1. a 2. testu zamítáme na hladině významnosti 0,05.

Řešení pomocí sytému STATISTICA:
Otevřeme nový datový soubor o dvou proměnných 1.TEST a 2.TEST a osmi případech. Zobrazíme dvourozměrný tečkový diagram s proloženou elipsou 95% konstantní hustoty pravděpodobnosti, s jehož pomocí posoudíme dvourozměrnou normalitu dat: Grafy – Bodové grafy – vypneme Typ proložení – Proměnné X 1.TEST, Y 2.TEST - OK . Na záložce Detaily vybereme Elipsa Normální – OK. Ve vzniklém dvourozměrném tečkovém diagramu změníme rozsah zobrazených hodnot na vodorovné a svislé ose, abychom viděli celou elipsu (viz obrázek výše).

Formát – Vš. Možnosti – Osa:Měřítka – Osa X – automatický mód změníme na manuální s minimem 0 a maximem 120. Totéž pro osu Y, ale stačí maximum 100.
Testování hypotézy o nezávislosti: Statistika – Základní statistiky /Tabulky - Korelační matice – OK – 1.seznam proměnných 1.TEST, 2.TEST, OK. Na záložce Možnosti zaškrtneme Zobrazit detailní tabulku výsledků – Souhrn.

	Korelace (priklad845) Označ. korelace jsou významné na hlad. p < ,05000 (Celé případy vynechány u ChD)										
Prom. X & prom. Y	Průměr	Sm.Odch.	r(X,Y)	r^2	t	p	N	Konst. záv.: Y	Směr. záv: Y	Konst. záv.: X	Směrnic záv.: X
1. test	56,25000	13,99745									
1. test	56,25000	13,99745	1,000000	1,000000			8	0,00000	1,000000	0,00000	1,000000
1. test	56,25000	13,99745									
2. test	50,00000	10,92834	0,666802	0,444625	2,191693	0,070909	8	20,71637	0,520598	13,54665	0,854067
2. test	50,00000	10,92834									
1. test	56,25000	13,99745	0,666802	0,444625	2,191693	0,070909	8	13,54665	0,854067	20,71637	0,520598
2. test	50,00000	10,92834									
2. test	50,00000	10,92834	1,000000	1,000000			8	0,00000	1,000000	0,00000	1,000000

Ve výstupní tabulce najdeme relizaci výběrového korelačního koeficientu ($r_{12} = 0{,}666802$, tzn. že mezi X a Y existuje nepříliš silná přímá lineární závislost), realizaci testové statistiky $t_0 = 2{,}191693$ a p-hodnotu pro test hypotézy o nezávislosti ($p = 0{,}070909$). Tato p-hodnota je však vypočítána pro testování nulové hypotézy proti oboustranné alternativě, proto ji musíme dělit 2. Dostaneme $p = 0{,}035455$, H_0 tedy zamítáme na hladině významnosti 0,05. S rizikem omylu nejvýše 5% jsme prokázali, že mezi výsledky 1. a 2. testu existuje přímá lineární závislost.

Poznámka: Pokud známe výběrový koeficient korelace a rozsah výběru, můžeme test nezávislosti veličin X, Y provést pomocí Pravděpodobnostního kalkulátoru. Statistika - Pravděpodobnostní kalkulátor – Korelace – zadáme n a r, zaškrtneme Počítat p pomocí r – Výpočet. V našem případě navíc ještě odškrtneme Dvojité, protože proti nulové hypotéze stavíme jednostrannou alternativu. V okénku p se objeví hodnota 0,035455, tedy na hladině významnosti 0,05 zamítáme hypotézu, že výsledky obou testů jsou nekorelované ve prospěch pravostranné alternativy, která tvrdí, že mezi výsledky obou testů existuje přímá lineární závislost.

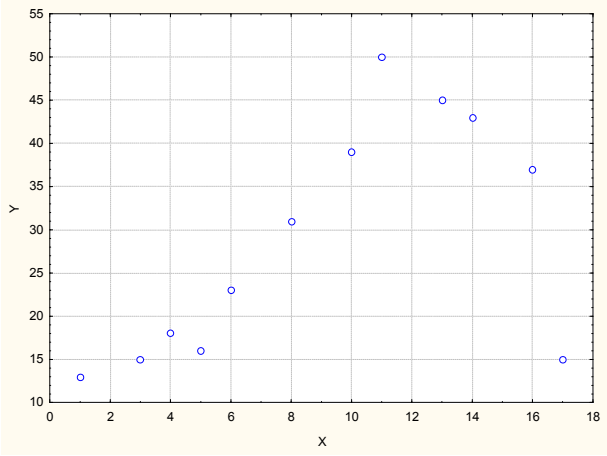
8.3.6. Příklad (Ilustrace postupu při nesplnění předpokladu dvourozměrné normality)
Máme k dispozici realizace náhodného výběru rozsahu 12 z dvourozměrného rozložení:

X	1	3	4	5	6	8	10	11	13	14	16	17
Y	13	15	18	16	23	31	39	50	45	43	37	15

Na hladině významnosti 0,05 testujte hypotézu, že náhodné veličiny X, Y jsou nezávislé proti oboustranné alternativě.
Řešení:
Nejprve ověříme předpoklady použití testu nezávislosti dvou náhodných veličin. Budeme tedy testovat hypotézu o normalitě náhodné veličiny X a náhodné veličiny Y pomocí Lilieforsovy varianty K - S testu a S – W testu:

Proměnná	Testy normality (Tabulka1)				
	N	max D	Lilliefors p	W	p
X	12	0,130669	p > .20	0,956714	0,736098
Y	12	0,202049	p < ,20	0,885918	0,104405

V obou případech hypotézu o normalitě nezamítáme na hladině významnosti 0,05.
Ověření dvourozměrné normality pomocí dvourozměrného tečkového diagramu:



Dvourozměrná normalita je silně porušena, tečky nevyplňují vnitřek elipsovitého obrazce.
Přejdeme tedy k testování hypotézy o pořadové nezávislosti: Testujeme hypotézu H_0: X, Y jsou pořadově nezávislé náhodné veličiny proti oboustranné alternativě H_1: X, Y jsou pořadově závislé náhodné veličiny.
Vypočítáme Spearmanův koeficient pořadové korelace.

X	1	3	4	5	6	8	10	11	13	14	16	17
Y	13	15	18	16	23	31	39	50	45	43	37	15
R_i	1	2	3	4	5	6	7	8	9	10	11	12
Q_i	1	2,5	5	4	6	7	9	12	11	10	8	2,5

$$r_S = 1 - \frac{6}{12\left(12^2-1\right)}\left[\begin{array}{l}(1-1)^2+(2-2{,}5)^2+(3-5)^2+(4-4)^2+(5-6)^2+(6-7)^2+(7-9)^2+\\ +(8-12)^2+(9-11)^2+(10-10)^2+(11-8)^2+(12-2{,}5)^2\end{array}\right] =$$

$$= 1 - \frac{6}{12\cdot 143}\left(0+0{,}25+4+0+1+1+4+16+4+0+9+90{,}25\right) = 1 - \frac{1}{286}\cdot 129{,}5 = 0{,}5472$$

Stanovíme kritický obor:

$$W = \left\langle -1, -r_{S,1-\alpha}(n)\right\rangle \cup \left\langle r_{S,1-\alpha}(n), 1\right\rangle = \left\langle -1, -r_{S,0,95}(12)\right\rangle \cup \left\langle r_{S,0,95}(12), 1\right\rangle = \left\langle -1, -0{,}5804\right\rangle \cup \left\langle 0{,}5804, 1\right\rangle$$

Testová statistika se nerealizuje v kritickém oboru, nulovou hypotézu nezamítáme na hladině významnosti 0,05.
Pokud bychom neověřili předpoklad dvourozměrné normality, pak obvyklým způsobem vypočteme realizaci výběrového koeficientu korelace $r_{12} = 0{,}5856$ a realizaci testové statistiky $t_0 = 2{,}2843$ pro test nezávislosti. Stanovíme kritický obor:
$W = \left(-\infty, -t_{0,975}(10)\right\rangle \cup \left\langle t_{0,975}(10), \infty\right) = \left(-\infty, -2{,}2281\right\rangle \cup \left\langle 2{,}2281, \infty\right)$. Protože $t_0 \in W$, zamítáme na hladině významnosti 0,05 hypotézu o nezávislosti náhodných veličin X a Y.
Vidíme tedy, že při nerespektování předpokladů testu můžeme dojít k chybným závěrům.

Shrnutí

Máme-li dvě náhodné veličiny ordinálního typu, pak testujeme hypotézu o pořadové nezávislosti těchto dvou veličin pomocí *Spearmanova koeficientu pořadové korelace*, který slouží zároveň jako testová statistika i jako míra intenzity pořadové závislosti daných veličin. Pro menší rozsahy výběrů (orientačně $n < 30$) porovnáváme tento koeficient s tabelovanou kritickou hodnotou, pro větší rozsahy výběrů využijeme jeho asymptotické normality.

Při testování hypotézy o nezávislosti dvou náhodných veličin intervalového či poměrového typu, které se řídí dvourozměrným normálním rozložením, využijeme skutečnosti, že v tomto případě je stochastická nezávislost ekvivalentní nekorelovanosti těchto dvou veličin. Testová statistika vznikne transformací *výběrového koeficientu korelace* a v případě platnosti nulové hypotézy se řídí Studentovým rozložením.

Při zkoumání závislosti dvou náhodných veličin aspoň intervalového typu je vhodné vytvořit dvourozměrný tečkový diagram a s jeho pomocí posoudit intenzitu a směr závislosti, případně orientačně ověřit dvourozměrnou normalitu dat.

Kontrolní otázky

1. K čemu slouží Spearmanův koeficient pořadové korelace?
2. Uveďte vlastnosti výběrového koeficientu korelace.
3. Jak se na vzhledu dvourozměrného tečkového diagramu projeví, jsou-li náhodné veličiny X, Y kladně korelovány?
4. Pro náhodný výběr z dvourozměrného normálního rozložení popište test hypotézy o nezávislosti veličin X, Y.

Autokorekční test

1. Nechť $(X_1, Y_1), ..., (X_{16}, Y_{16})$ je náhodný výběr z dvourozměrného normálního rozložení. Výběrový koeficient korelace R_{12} nabyl hodnoty -0,87. Jestliže provedeme lineární transformaci $U= 1 + 3X$, $V = -3 - Y$, jakou hodnotu nabude výběrový koeficient korelace transformovaných hodnot $(U_1, V_1), ..., (U_{16}, V_{16})$?
a) –0,61
b) 0,87
c) –0,87

2. Pro 12 náhodně vybraných ojetých automobilů byl vypočten výběrový koeficient korelace mezi jejich stářím v měsících a počtem najetých kilometrů. Nabyl hodnoty 0,831. Předpokládáme, že data pocházejí z dvourozměrného normálního rozložení. Jaká je hodnota testové statistiky pro test nezávislosti obou veličin?
a) 4,724
b) 0,831
c) 6,392

3. Pro dvourozměrný náhodný výběr rozsahu $n = 10$ z dvourozměrného normálního rozložení byl vypočten výběrový koeficient korelace. Nabyl hodnoty –0,94. Co lze usoudit o vztahu náhodných veličin X a Y?
a) S růstem hodnot jedné náhodné veličiny hodnoty druhé náhodné veličiny lineárně rostou.
b) Veličiny X a Y jsou nezávislé.
c) S růstem hodnot jedné náhodné veličiny hodnoty druhé náhodné veličiny lineárně klesají.

4. Nechť dvourozměrný náhodný výběr pochází z dvourozměrného rozložení, které je výrazně odlišné od normálního. Chceme-li testovat hypotézu, že náhodné veličiny X a Y, které jsou poměrového typu, jsou nezávislé, použijeme testovou statistiku, která je založena na
a) Cramérově koeficientu
b) Spearmanově koeficientu pořadové korelace
c) výběrovém koeficientu korelace.

5. Na základě dvourozměrného náhodného výběru rozsahu 18 byl vypočten Spearmanův koeficient pořadové korelace 0,4819. Jak vypadá kritický obor pro test hypotézy o pořadové nekorelovanosti proti oboustranné alternativě, pokud hladinu významnosti volíme 0,05?
a) $W = (-\infty, -2{,}1199\rangle \cup \langle 2{,}1199, \infty)$
b) $W = (-\infty, -0{,}4716\rangle \cup \langle 0{,}4716, \infty)$
c) $W = (-\infty, -1{,}96 \cup \langle 1{,}96, \infty)$

Správné odpovědi: 1b) 2a) 3c) 4b) 5b)

Příklady

1. Dvanáct různých softwarových firem nabízí programy pro vedení účetnictví. Programy byly posouzeny odbornou komisí a komisí složenou z profesionálních účetních. Výsledky v 1. a 2. komisi: (6,4), (7,5), (1,2), (8,10), (4,6), (2.5,1), (9,7), (12,11), (10,8), (2.5,3), (5,12), (11,9). Vypočtěte Spearmanův koeficient pořadové korelace a na hladině významnosti 0,05 testujte hypotézu o nezávislosti pořadí v obou komisích.
Výsledek:
Spearmanův koeficent pořadové korelace je 0,715, kritická hodnota pro n = 12 a $\alpha = 0{,}05$ je 0,576. H_0 zamítáme na hladině významnosti 0,05 ve prospěch oboustranné alternativy.

2. V dílně pracuje 15 dělníků, u nichž byl zjištěn počet směn odpracovaných za měsíc (veličina X) a počet zhotovených výrobků (veličina Y). Orientačně ověřte dvourozměrnou normalitu dat, vypočtěte výběrový koeficient korelace mezi X a Y a na hladině 0,01 testujte hypotézu o nezávislosti veličin X a Y.

X	20	21	18	17	20	18	19	21	20	14	16	19	21	15	15
Y	92	93	83	80	91	85	82	98	90	60	73	86	96	64	81

Výsledek:
Vzhled dvourozměrného tečkového diagramu svědčí o tom, že předpoklad dvourozměrné normality je oprávněný. Výběrový koeficient korelace je 0,927, testová statistika se realizuje hodnotou 8,597, kritický obor je $W = (-\infty, -3{,}012\rangle \cup \langle 3{,}012, \infty)$. Hypotézu o nezávislosti veličin X a Y zamítáme na hladině významnosti 0,01.

3. V následující tabulce jsou uvedeny číselné realizace a absolutní četnosti náhodného výběru (X_1, Y_1), (X_1, Y_2), ..., (X_{62}, Y_{62}) z dvourozměrného rozložení:

x	y						
	1	3	5	7	9	11	13
15	0	0	0	0	1	2	1
25	0	0	0	5	4	2	0
35	0	0	5	8	2	0	0
45	0	5	6	4	0	0	0
55	3	5	3	0	0	0	0
65	4	2	0	0	0	0	0

Podle vzhledu dvourozměrného tečkového diagramu orientačně posuďte dvourozměrnou normalitu dat. Vypočtěte výběrový koeficient korelace a interpretujte ho. Na hladině významnosti 0,05 testujte hypotézu o nezávislosti veličin X a Y.
Výsledek:
Protože tečky v dvourozměrném tečkovém diagramu vytvářejí elipsovitý obrazec, lze připustit dvourozměrnou normalitu. Výběrový koeficient korelace nabývá hodnoty –0,899, což znamená, že mezi veličinami X a Y existuje dosti silná nepřímá lineární závislost. Testová statistika se realizuje hodnotou -13,6613, odpovídající p-hodnota je velmi blízká 0, nulovou hypotézu zamítáme na hladině významnosti 0,05.

4. Pro náhodný výběr (X_i, Y_i) , i = 1, ..., 27 z dvourozměrného normálního rozložení byl vypočten výběrový koeficient korelace 0,77. Na hladině významnosti 0,01 testujte hypotézu o nezávislosti veličin X, Y proti pravostranné alternativě.
Výsledek:
Testová statistika se realizuje hodnotou 6,034, kritický obor pro pravostrannou alternativu $W = \langle 2{,}4851, \infty)$. Protože testová statistika se realizuje v kritickém oboru, nulovou hypotézu zamítáme na hladině významnosti 0,01 ve prospěch pravostranné alternativy.