Kapitola 8.: Analýza závislosti dvou náhodných veličin

8.1. Motivace

Při zpracování dat se velmi často setkáme s úkolem zjistit, zda dvě náhodné veličiny jsou stochasticky nezávislé. Např. nás může zajímat, zda ve sledované populaci je barva očí a barva vlasů nezávislá nebo zda počet dnů absence a věk pracovníka jsou nezávislé.Testování hypotézy o nezávislosti se provádí různými způsoby podle toho, jakého typu jsou dané náhodné veličiny – zda jsou nominální, ordinální, intervalové či poměrové.

Při zkoumání závislosti je nesmírně 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.

Zpravidla chceme také zjistit intenzitu případné závislosti sledovaných dvou veličin. K tomuto účelu byly zkonstruovány různé koeficienty, které nabývají hodnot od 0 do 1 (resp. od -1 do 1). Čím je takový koeficient bližší 1 (resp. -1), tím je závislost mezi danými dvěma veličinami silnější a čím je bližší 0, tím je slabší.

8.2. Testování nezávislosti nominálních veličin

8.2.1. Popis testu

Nechť X,Y jsou dvě nominální náhodné veličiny (tj. obsahová interpretace je možná jenom u relace rovnosti). Nechť X nabývá variant $x_{[1]}, ..., x_{[r]}$ a Y nabývá variant $y_{[1]}, ..., y_{[s]}$. Pořídíme dvourozměrný náhodný výběr rozsahu n z rozložení, kterým se řídí dvourozměrný diskrétní náhodný vektor (X, Y). Zjištěné absolutní četnosti n_{jk} dvojice variant $(x_{[j]}, y_{[k]})$ uspořádáme do kontingenční tabulky:

	y	$y_{[1]}$	...	$y_{[s]}$	$n_{j.}$
x	n_{jk}				
$x_{[1]}$		n_{11}	...	n_{1s}	$n_{1.}$
$\vdots$		...	...	...	...
$x_{[r]}$		n_{r1}	...	n_{rs}	$n_{r.}$
$n_{.k}$		$n_{.1}$	...	$n_{.s}$	n

Testujeme hypotézu H_0: X, Y jsou stochasticky nezávislé náhodné veličiny proti H_1: X, Y nejsou stochasticky nezávislé náhodné veličiny. Testová statistika má tvar:

$$K = \sum_{j=1}^{r}\sum_{k=1}^{s}\frac{\left(n_{jk} - \frac{n_{j.}n_{.k}}{n}\right)^2}{\frac{n_{j.}n_{.k}}{n}}$$. Platí-li H_0, pak K se asymptoticky řídí rozložením $\chi^2((r-1)(s-1))$.

Hypotézu o nezávislosti veličin X, Y tedy zamítáme na asymptotické hladině významnosti α, když $K \geq \chi^2_{1-\alpha}((r-1)(s-1))$.

8.2.2. Podmínky dobré aproximace

Výraz $\frac{n_{j.}n_{.k}}{n}$ se nazývá teoretická četnost. Rozložení statistiky K lze aproximovat rozložením $\chi^2((r-1)(s-1))$, pokud teoretické četnosti aspoň v 80% případů nabývají hodnoty větší nebo rovné 5 a ve zbylých 20% neklesnou pod 2. Není-li splněna podmínka dobré aproximace, doporučuje se slučování některých variant.

8.2.3. Měření síly závislosti

Cramérův koeficient: $V = \sqrt{\frac{K}{n(m-1)}}$, kde $m = \min\{r,s\}$. Tento koeficient nabývá hodnot mezi 0 a 1. Čím blíže je 1, tím je těsnější závislost mezi X a Y, čím blíže je 0, tím je tato závislost volnější.

8.2.4. Příklad

V sociologickém průzkumu byl z uchazečů o studium na vysokých školách pořízen náhodný výběr rozsahu 360. Mimo jiné se zjišťovala sociální skupina, ze které uchazeč pochází a typ školy, na kterou se hlásí. Výsledky jsou zaznamenány v kontingenční tabulce:

Typ školy	Sociální skupina				$n_{j.}$
	I	II	III	IV	
univerzitní	50	30	10	50	140
technický	30	50	20	10	110
ekonomický	10	20	30	50	110
$n_{.k}$	90	100	60	110	360

Na asymptotické hladině významnosti 0,05 testujte hypotézu o nezávislosti typu školy a sociální skupiny. Vypočtěte Cramérův koeficient.

Řešení:

$$\frac{n_{1.}n_{.1}}{n} = \frac{140\cdot 90}{360} = 35,\ \frac{n_{1.}n_{.2}}{n} = \frac{140\cdot 100}{360} = 38{,}9,\ \frac{n_{1.}n_{.3}}{n} = \frac{140\cdot 60}{360} = 23{,}3,\ \frac{n_{1.}n_{.4}}{n} = \frac{140\cdot 110}{360} = 42{,}8,$$
$$\frac{n_{2.}n_{.1}}{n} = \frac{110\cdot 90}{360} = 27{,}5,\ \frac{n_{2.}n_{.2}}{n} = \frac{110\cdot 100}{360} = 30{,}6,\ \frac{n_{2.}n_{.3}}{n} = \frac{110\cdot 60}{360} = 18{,}3,\ \frac{n_{2.}n_{.4}}{n} = \frac{110\cdot 110}{360} = 33{,}6,$$
$$\frac{n_{3.}n_{.1}}{n} = \frac{110\cdot 90}{360} = 27{,}5,\ \frac{n_{3.}n_{.2}}{n} = \frac{110\cdot 100}{360} = 30{,}6,\ \frac{n_{3.}n_{.3}}{n} = \frac{110\cdot 60}{360} = 18{,}3,\ \frac{n_{3.}n_{.4}}{n} = \frac{110\cdot 110}{360} = 33{,}6$$

$K = \frac{(50-35)^2}{35} + \frac{(30-38{,}9)^2}{38{,}9} + \ldots + \frac{(50-33{,}6)^2}{33{,}6} = 76{,}84$, $r = 3$, $s = 4$, $\chi^2_{0,95}(6) = 12{,}6$. Protože $K \geq 12{,}6$, hypotézu o nezávislosti typu školy a sociální skupiny zamítáme na asymptotické hladině významnosti 0,05. Cramérův koeficient: $V = \sqrt{\frac{76{,}4}{360\cdot 2}} = 0{,}3267$.

8.2.5. Čtyřpolní tabulky

Nechť $r = s = 2$. Pak hovoříme o čtyřpolní kontingenční tabulce a používáme označení: $n_{11} = a$, $n_{12} = b$, $n_{21} = c$, $n_{22} = d$.

X	Y		$n_{j.}$
	$y_{[1]}$	$y_{[2]}$	
$x_{[1]}$	a	b	a+b
$x_{[2]}$	c	d	c+d
$n_{.k}$	a+c	b+d	n

Pro tuto tabulku navrhl R. A. Fisher přesný (exaktní) test nezávislosti známý jako Fisherův faktoriálový test. (Je popsán např. v knize K. Zvára: Biostatistika, Karolinum, Praha 1998.) STATISTICA poskytuje p-hodnotu pro tento test. Jestliže vyjde $p \le \alpha$, pak hypotézu o nezávislosti zamítáme na hladině významnosti α.

Ve čtyřpolních tabulkách používáme charakteristiku $OR = \frac{ad}{bc}$, která se nazývá podíl šancí (odds ratio). Můžeme si představit, že pokus se provádí za dvojích různých okolností a může skončit buď úspěchem nebo neúspěchem.

Výsledek pokusu	okolnosti		$n_{j.}$
	I	II	
úspěch	a	b	a+b
neúspěch	c	d	c+d
$n_{.k}$	a+c	b+d	n

Poměr počtu úspěchů k počtu neúspěchů (tzv. šance) za 1. okolností je $\frac{a}{c}$, za druhých okolností je $\frac{b}{d}$. Podíl šancí je $OR = \frac{ad}{bc}$. Pomocí 100(1-α)% asymptotického intervalu spolehlivosti pro podíl šancí lze na asymptotické hladině významnosti α testovat hypotézu o nezávislosti nominálních veličin X a Y. Asymptotický 100(1-α)% interval spolehlivosti pro přirozený logaritmus skutečného podílu šancí má meze:

$\ln OR \pm \sqrt{\frac{1}{a}+\frac{1}{b}+\frac{1}{c}+\frac{1}{d}}\,u_{1-\alpha/2}$. Jestliže po odlogaritmování nezahrne interval spolehlivosti 1, pak hypotézu o nezávislosti zamítneme na asymptotické hladině významnosti α.

8.2.6. Příklad

U 135 uchazečů o studium na jistou fakultu byl hodnocen dojem, jakým zapůsobili na komisi u ústní přijímací zkoušky. Na asymptotické hladině významnosti 0,05 testujte hypotézu, že přijetí na fakultu nezávisí na dojmu u přijímací zkoušky.

přijetí	dojem		$n_{j.}$
	dobrý	špatný	
ano	17	11	28
ne	39	58	97
$n_{.k}$	56	69	125

Řešení:

$$OR = \frac{ad}{bc} = \frac{17 \cdot 58}{11 \cdot 39} = 2{,}298,\ \ln OR = 0{,}832,$$

$$\sqrt{\frac{1}{a}+\frac{1}{b}+\frac{1}{c}+\frac{1}{d}} = \sqrt{\frac{1}{17}+\frac{1}{11}+\frac{1}{39}+\frac{1}{58}} = 0{,}439,\ u_{0{,}975} = 1{,}96$$

$$\ln d = 0{,}832 - 0{,}439 \cdot 1{,}96 = -0{,}028,\ \ln h = 0{,}832 + 0{,}439 \cdot 1{,}96 = 1{,}692$$

$$d = e^{-0{,}028} = 0{,}972,\ h = e^{1{,}692} = 5{,}433$$

Protože interval (0,972; 5,433) obsahuje číslo 1, na asymptotické hladině významnosti 0,05 nezamítáme hypotézu o nezávislosti dojmu u přijímací zkoušky a přijetí na fakultu.

8.3. Testování nezávislosti ordinálních veličin

8.3.1. 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. Testujeme hypotézu H_0: X, Y jsou pořadově nezávislé náhodné veličiny proti oboustranné alternativě H_1: X, Y jsou pořadově závislé náhodné veličiny (resp. proti levostranné alternativě H_1: mezi X a Y existuje nepřímá pořadová závislost resp. proti pravostranné alternativě H_1: mezi X a Y existuje přímá pořadová závislost).

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

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

H_0 zamítáme na hladině významnosti α

a) ve prospěch oboustranné alternativy, když $|r_S| \geq r_{S,1-\alpha}(n)$

b) ve prospěch levostranné alternativy, když $r_S \leq - r_{S,1-\alpha}(n)$

c) ve prospěch pravostranné alternativy, když $r_S \geq r_{S,1-\alpha}(n)$, kde $r_{S,1-\alpha}(n)$ je kritická hodnota, kterou pro α = 0,05 nebo 0,01 a n ≤ 30 najdeme v tabulkách. Pro n > 30 H_0 zamítáme na asymptotické hladině významnosti α ve prospěch oboustranné alternativy, když $|r_S| \geq \frac{u_{1-\alpha}}{\sqrt{n-1}}$ (analogicky pro jednostranné alternativy).

Spearmanův koeficient r_S současně měří sílu pořadové závislosti náhodných veličin X, Y. Nabývá hodnot z intervalu $\langle -1,1 \rangle$. Čím je jeho hodnota bližší -1 (resp.1), tím je silnější nepřímá (resp. přímá) pořadová závislost veličin X, Y. Čím je jeho hodnota bližší 0, tím je slabší pořadová závislost veličin X, Y.

8.3.2. 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:

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

Kritická hodnota: $r_{S,0,95}(7) = 0,745$. Protože 0,857 ≥ 0,745, nulovou hypotézu zamítáme na hladině významnosti 0,05.

8.4. Testování nezávislosti intervalových či poměrových veličin

8.4.1. Pearsonův koeficient korelace

V teorii pravděpodobnosti byl zaveden Pearsonův koeficient korelace náhodných veličin X, Y (které jsou aspoň intervalového charakteru) vztahem
$R(X,Y) = \frac{C(X,Y)}{\sqrt{D(X)}\sqrt{D(Y)}}$ pro $\sqrt{D(X)}\sqrt{D(Y)} > 0$, = 0 jinak . Připomeneme jeho vlastnosti:

a) $R(X,X) = 1$
b) $R(X,Y) = R(Y,X)$
c) $R(a + bX, c + dY) = \mathrm{sgn}(bd)R(X,Y)$
d) $-1 \leq R(X,Y) \leq 1$ 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.

8.4.2. Výběrový koeficient korelace

R(X,Y) většinou nemůžeme počítat přímo, protože to vyžaduje znalost simultánního rozložení náhodného vektoru (X, Y). V praxi jsme většinou odkázáni na náhodný výběr rozsahu n z dvourozměrného rozložení daného distribuční funkcí $\Phi(x,y)$. 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.

8.4.3. Koeficient korelace dvourozměrného normálního rozložení

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

$\varphi(x,y) = \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}} e^{-\frac{1}{2(1-\rho^2)}\left[\left(\frac{x-\mu_1}{\sigma_1}\right)^2 - 2\rho\frac{x-\mu_1}{\sigma_1}\frac{y-\mu_2}{\sigma_2} + \left(\frac{y-\mu_2}{\sigma_2}\right)^2\right]}$, 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 $(X_1, Y_1), \dots, (X_n, Y_n)$ pochází z dvourozměrného normálního rozložení s parametry $\mu_1, \mu_2, \sigma_1^2, \sigma_2^2, \rho$

8.4.4. Testování hypotézy o nezávislosti

Testujeme H_0: $\rho = 0$ proti oboustranné alternativě H_1: $\rho \neq 0$ (resp. proti levostranné alternativě H_1: $\rho < 0$ resp. proti pravostranné alternativě H_1: $\rho > 0$). Testová statistika má tvar: $T = \dfrac{R_{12}\sqrt{n-2}}{\sqrt{1-R_{12}^2}}$. Platí-li nulová hypotéza, pak $T \sim t(n-2)$. Kritický obor pro test H_0 proti oboustranné alternativě: $W = \left(-\infty, -t_{1-\alpha/2}(n-2)\right\rangle \cup \left\langle t_{1-\alpha/2}(n-2), \infty\right)$, proti levostranné alternativě: $W = \left(-\infty, -t_{1-\alpha}(n-2)\right\rangle$ a proti pravostranné alternativě: $W = \left\langle t_{1-\alpha}(n-2), \infty\right)$. H_0 zamítáme na hladině významnosti α, když $T \in W$.

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

8.4.5. Příklad

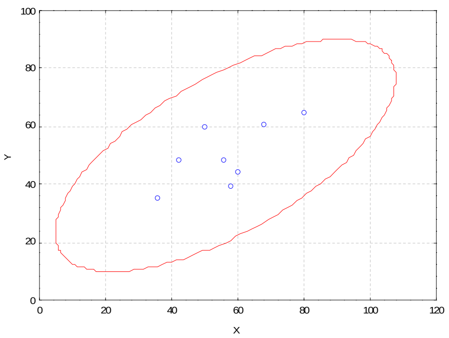
Máme k dispozici výsledky testů ze dvou předmětů zjištěné u osmi náhodně vybraných studentů určitého oboru.

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

Kontrolní otázky

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

Příklady

1. Na hladině významnosti 0,05 testujte hypotézu o nezávislosti pedagogické hodnosti a pohlaví a vypočtěte Cramérův koeficient, jsou-li k dispozici následující údaje:

pohlaví	pedagogická hodnost		
	odb. asistent	docent	profesor
muž	32	15	8
žena	34	8	3

Výsledek:
Podmínky dobré aproximace jsou splněny, pouze jedna teoretická četnost klesne pod 5. Testová statistika se realizuje hodnotou 3,5 , počet stupňů volnosti = 2, kritický obor je $W = \langle 5{,}991; \infty \rangle$. Hypotézu o nezávislosti pohlaví a pedagogické hodnosti tedy nezamítáme na asymptotické hladině významnosti 0,05. Cramérův koeficient V = 0,187.

2. 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.

3. 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.

4. (S) 100 náhodně vybraných mužů a žen bylo dotázáno, zda dávají přednost nealkoholickému nápoji A či B. Údaje jsou uvedeny ve čtyřpolní kontingenční tabulce.

pohlaví	nápoj	
	A	B
muž	20	30
žena	30	20

Na hladině významnosti 0,05 testujte pomocí Fisherova faktoriálového testu hypotézu, že preferovaný typ nápoje nezáleží na pohlaví respondenta.
Výsledek:
V našem případě se jedná o jednostrannou závislost, zajímáme se tedy o Fisher exact, one tailed. Ta je 0,03567. Protože p-hodnota je menší nebo rovna 0,05, zamítáme na hladině významnosti hypotézu, že preferovaný typ nápoje nezáleží na pohlaví respondenta.

5. (S) 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.