11. Analýza závislosti dvou veličin

11.1. 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.

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.

11.2. Testování nezávislosti nominálních veličin

11.2.1. Popis testu

Nechť X,Y jsou dvě nominální náhodné veličiny (tj. obsahová interpretace je možná jenom u relace rovnosti). Nechť X nabývá variant $x_{[1]}, \ldots, x_{[r]}$ a Y nabývá variant $y_{[1]}, \ldots, y_{[s]}$. 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}	$n_{1.}$
⋮		…	…	…	…
$x_{[r]}$		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 $K \geq \chi^2_{1-\alpha}((r-1)(s-1))$.

11.2.2. 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.

11.2.3. 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.

11.2.4. 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: $V = \sqrt{\frac{76{,}4}{360\cdot 2}} = 0{,}3267$.

11.2.5. 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.

11.2.6. 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.

11.3. Testování nezávislosti ordinálních veličin

11.3.1. 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:

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

H_0 zamítáme na hladině významnosti α

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

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 $\alpha = 0{,}05$ nebo $0{,}01$ a $n \leq 30$ najdeme v tabulkách. Pro $n > 30$ H_0 zamítáme na asymptotické hladině významnosti α ve prospěch oboustranné alternativy, když $\left|r_S\right| \geq \frac{u_{1-\alpha}}{\sqrt{n-1}}$

(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.

11.3.2. 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:

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

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

11.4. Testování nezávislosti intervalových či poměrových veličin

11.4.1. 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) = 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.

11.4.2. 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 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 $S_1 S_2 > 0$). Vlastnosti a), b), c), d) koeficientu korelace se přenášejí i na výběrový koeficient korelace.

11.4.3. 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 $(X_1, Y_1), ..., (X_n, Y_n)$ pochází z dvourozměrného normálního rozložení s parametry $\mu_1, \mu_2, \sigma_1^2, \sigma_2^2, \rho$

11.4.4. Testování hypotézy o nezávislosti

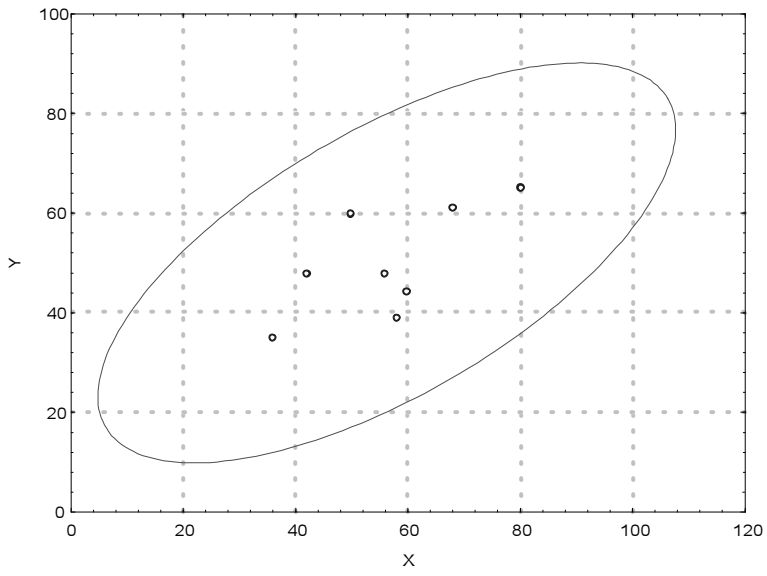
Testujeme H_0: $\rho = 0$ proti oboustranné alternativě H_1: $\rho \neq 0$ (resp. proti levostranné alternativě H_1: $\rho < 0$ resp. proti pravostranné alternativě H_1: $\rho > 0$). Testová statistika má tvar: $T = \frac{R_{12}\sqrt{n-2}}{\sqrt{1-R_{12}^2}}$. Platí-li nulová hypotéza, pak $T \sim t(n-2)$.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

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.

11.4.5. Porovnání koeficientu korelace s danou konstantou

Nechť c je reálná konstanta. Testujeme H_0: $\rho = c$ proti H_1: $\rho \neq c$. (Tento test se provádí např. tehdy, když experimentátor porovnává vlastnosti svých dat s vlastnostmi uváděnými v literatuře.) Test je založen na statistice $U = \left(Z - \frac{1}{2}\ln\frac{1+c}{1-c} - \frac{c}{2(n-1)} \right)\sqrt{n-3}$, která má za platnosti H_0 pro $n \geq 10$ asymptoticky rozložení N(0,1), přičemž $Z = \frac{1}{2}\ln\frac{1+R_{12}}{1-R_{12}}$ je tzv. Fisherova Z-transformace. Kritický obor pro test H_0 proti oboustranné alternativě tedy je $W = (-\infty, -u_{1-\alpha/2}\rangle \cup \langle u_{1-\alpha/2}, \infty)$. H_0 zamítáme na asymptotické hladině významnosti α, když $U \in W$.

Příklad: U 600 vzorků rudy byl stanoven obsah železa dvěma analytickými metodami s výběrovým koeficientem korelace 0,85. V literatuře se uvádí, že koeficient korelace těchto dvou metod má být 0,9. Na asymptotické hladině významnosti 0,05 testujte hypotézu H_0: $\rho = 0{,}9$ proti H_1: $\rho \neq 0{,}9$.

Řešení: $Z = \frac{1}{2}\ln\frac{1+0{,}85}{1-0{,}85} = 1{,}2562$,

$U = \left(1{,}2562 - \frac{1}{2}\ln\frac{1+0{,}9}{1-0{,}9} - \frac{0{,}9}{2(600-1)} \right)\sqrt{600-3} = -5{,}2976$, $u_{0{,}975} = 1{,}96$,

$W = (-\infty, -1{,}96\rangle \cup \langle 1{,}96, \infty)$. Protože $U \in W$, H_0 zamítáme na asymptotické hladině významnosti 0,05.

11.4.6. Porovnání dvou korelačních koeficientů

Nechť jsou dány dva nezávislé náhodné výběry o rozsazích n a n^* z dvourozměrných normálních rozložení s korelačními koeficienty ρ a ρ^*. Testujeme H_0: $\rho = \rho^*$ proti H_1: $\rho \neq \rho^*$. Označme R_{12} výběrový korelační koeficient 1. výběru a R_{12}^* výběrový korelační koeficient 2. výběru. Položme $Z = \frac{1}{2}\ln\frac{1+R_{12}}{1-R_{12}}$ a $Z^* = \frac{1}{2}\ln\frac{1+R_{12}^*}{1-R_{12}^*}$. Platí-li H_0, pak testová statistika $U = \frac{Z - Z^*}{\sqrt{\frac{1}{n-3} + \frac{1}{n^*-3}}}$ má asymptoticky rozložení N(0,1). Kritický obor pro test H_0 proti oboustranné alternativě tedy je $W = (-\infty, -u_{1-\alpha/2}\rangle \cup \langle u_{1-\alpha/2}, \infty)$. H_0 zamítáme na asymptotické hladině významnosti α, když $U \in W$.

Příklad: Lékařský výzkum se zabýval sledováním koncentrací látek A a B v moči pacientů trpících určitou ledvinovou chorobou. U 100 zdravých jedinců činil výběrový korelační koeficient mezi koncentracemi obou látek 0,65 a u 142 osob trpících zmíněnou chorobou byl 0,37. Na asymptotické hladině významnosti 0,05 testujte hypotézu, že korelační koeficienty v obou skupinách se neliší.

Řešení: $Z = \frac{1}{2}\ln\frac{1+0{,}65}{1-0{,}65} = 0{,}7753,\ Z^* = \frac{1}{2}\ln\frac{1+0{,}37}{1-0{,}37} = 0{,}3884$,

$U = \frac{0{,}7753 - 0{,}3884}{\sqrt{\frac{1}{100-3}+\frac{1}{142-3}}} = 2{,}9242$, $u_{0{,}975} = 1{,}96$, $W = \langle -\infty, -1{,}96\rangle \cup \langle 1{,}96, \infty\rangle$. Protože $U \in W$, H_0 zamítáme na asymptotické hladině významnosti 0,05.

11.4.7. Interval spolehlivosti pro korelační koeficient

Jestliže dvourozměrný náhodný výběr rozsahu n pochází z dvourozměrného normálního rozložení, jehož korelační koeficient se příliš neliší od nuly ($|\rho| < 0{,}5$) a rozsah výběru je dostatečně velký ($n \geq 100$), lze odvodit, že 100(1-α)% interval spolehlivosti pro ρ má meze $R_{12} \pm u_{1-\alpha/2}\frac{1-R_{12}^2}{\sqrt{n-3}}$.

Nejsou-li uvedené podmínky splněny, pak nelze tento vzorec použít, protože rozložení výběrového korelačního koeficientu je příliš zešikmené. V takovém případě využijeme toho, že náhodná veličina $Z = \frac{1}{2}\ln\frac{1+R_{12}}{1-R_{12}}$ má i při malém rozsahu výběru přibližně normální rozložení se střední hodnotou $E(Z) = \frac{1}{2}\ln\frac{1+\rho}{1-\rho} + \frac{\rho}{2(n-1)}$ (2. sčítanec lze při větším n zanedbat) a rozptylem $D(Z) = \frac{1}{n-3}$. Standardizací veličiny Z dostaneme veličinu $U = \frac{Z - E(Z)}{\sqrt{D(Z)}}$, která má asymptoticky rozložení N(0,1). Tudíž 100(1-α)% asymptotický interval spolehlivosti pro $\frac{1}{2}\ln\frac{1+\rho}{1-\rho}$ bude mít meze $Z \pm \frac{u_{1-\alpha/2}}{\sqrt{n-3}}$. Interval spolehlivosti pro ρ pak dostaneme zpětnou transformací.

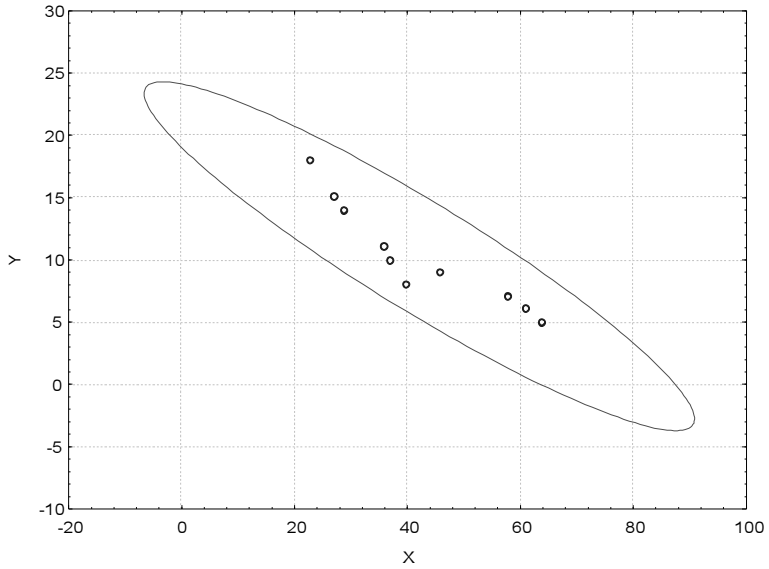
Poznámka: Jelikož Z = arctgh R_{12}, dostáváme R_{12} = tgh Z a meze intervalu spolehlivosti pro ρ můžeme psát ve tvaru $\mathrm{tgh}\left(Z \pm \frac{u_{1-\alpha/2}}{\sqrt{n-3}}\right)$, přičemž $\mathrm{tgh}\ x = \frac{e^x - e^{-x}}{e^x + e^{-x}}$.

Příklad: Pracovník personálního oddělení určité firmy zkoumá, zda existuje vztah mezi počtem dní absence za rok (veličina Y) a věkem pracovníka (veličina X). Proto náhodně vybral údaje o 10 pracovnících.

Č.prac.	1	2	3	4	5	6	7	8	9	10
X	27	61	37	23	46	58	29	36	64	40
Y	15	6	10	18	9	7	14	11	5	8

Za předpokladu, že uvedené údaje tvoří číselné realizace náhodného výběru rozsahu 10 z dvourozměrného normálního rozložení, vypočtěte výběrový korelační koeficient a na hladině významnosti 0,05 testujte hypotézu, že X a Y jsou nezávislé náhodné veličiny. Sestrojte 95% asymptotický interval spolehlivosti pro skutečný korelační koeficient ρ.

Řešení: Předpoklad o dvourozměrné normalitě dat ověříme orientačně pomocí dvourozměrného tečkového diagramu.



Vzhled diagramu svědčí o tom, že předpoklad je oprávněný.

Testujeme H_0: $\rho = 0$ proti H_1: $\rho \neq 0$. Vypočítáme $R_{12} = -0{,}9325$, tedy mezi věkem pracovníka a počtem dnů pracovní neschopnosti existuje silná nepřímá lineární závislost. Testová statistika: $T = -7{,}3053$, kvantil $t_{0{,}975}(8) = 2{,}306$, kritický obor $W = \langle -\infty, -2{,}306 \rangle \cup \langle 2{,}306, \infty \rangle$. Jelikož $T \in W$, zamítáme na hladině významnosti 0,05 hypotézu o nezávislosti veličin X a Y.

Vypočítáme $Z = \frac{1}{2} \ln \frac{1 + R_{12}}{1 - R_{12}} = \frac{1}{2} \ln \frac{1 - 0{,}9325}{1 + 0{,}9325} = -1{,}6772$. Meze 95% asymptotického intervalu spolehlivosti pro ρ jsou $\operatorname{tgh}\left(-1{,}6772 \pm \frac{1{,}96}{\sqrt{7}} \right)$, tedy $-0{,}9842 < \rho < -0{,}7336$ s pravděpodobností přibližně 0,95.

Příklady k 11. kapitole

Příklad 1.: **Testování nezávislosti nominálních veličin**
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

Řešení: Tabulku doplníme o marginální četnosti.

pohlaví	pedagogická hodnost			$n_{j.}$
	odb. asistent	docent	profesor	
muž	32	15	8	55
žena	34	8	3	45
$n_{.k}$	66	23	11	n=100

Vypočteme teoretické četnosti:

$$\frac{n_{1.}n_{.1}}{n}=\frac{55\cdot 66}{100}=36{,}3,\ \frac{n_{1.}n_{.2}}{n}=\frac{55\cdot 23}{360}=12{,}65,\ \frac{n_{1.}n_{.3}}{n}=\frac{55\cdot 11}{100}=6{,}05,$$

$$\frac{n_{2.}n_{.1}}{n}=\frac{45\cdot 66}{100}=29{,}7,\ \frac{n_{2.}n_{.2}}{n}=\frac{45\cdot 23}{100}=10{,}35,\ \frac{n_{2.}n_{.3}}{n}=\frac{45\cdot 11}{100}=4{,}95.$$

Vypočteme testovou statistiku: $K=\sum_{j=1}^{r}\sum_{k=1}^{s}\frac{\left(n_{jk}-\frac{n_{j.}n_{.k}}{n}\right)^2}{\frac{n_{j.}n_{.k}}{n}}$, r = 2, s = 3, tedy

$K=\frac{(32-36{,}3)^2}{36{,}3}+\frac{(15-12{,}65)^2}{12{,}65}+\ldots+\frac{(3-4{,}95)^2}{4{,}95}=3{,}5$, $\chi^2_{0,95}(2)=5{,}991$. Protože $K<5{,}991$, hypotézu o nezávislosti pohlaví a pedagogické hodnosti nezamítáme na asymptotické hladině významnosti 0,05.

Cramérův koeficient: $V=\sqrt{\frac{K}{n(m-1)}}$, kde m = min{r,s}, tedy $V=\sqrt{\frac{3{,}5}{100\cdot 1}}=0{,}187$.

Příklad 2.: **Testování nezávislosti ordinálních veličin**
12 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.

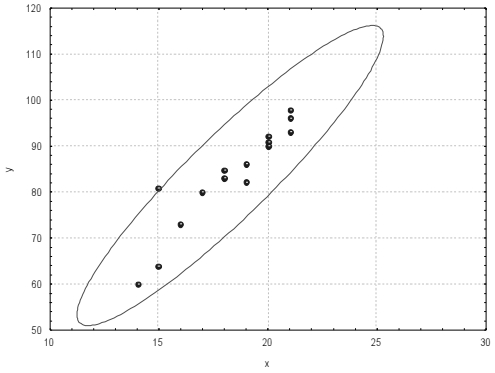
Řešení: $r_S=1-\frac{6}{n(n^2-1)}\sum_{i=1}^{n}(R_i-Q_i)^2$, tedy

$$r_S=1-\frac{6}{12(12^2-1)}\left[\begin{array}{l}(6-4)^2+(7-5)^2+(1-2)^2+(8-10)^2+(4-6)^2+(2{,}5-1)^2+(9-7)^2+\\+(12-11)^2+(10-8)^2+(2{,}5-3)^2+(5-12)^2+(11-9)^2\end{array}\right]=0{,}715$$

Kritická hodnota: $r_{S,0,95}(12)=0{,}576$. Protože $r_S\geq 0{,}576$, nulovou hypotézu zamítáme na hladině významnosti 0,05.

Příklad 3.: **Testování nezávislosti intervalových a poměrových veličin**
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, sestrojte pro něj 99% asymptotický interval spolehlivosti a na hladině 0,01 testujte hypotézu o nezávislosti 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.

Řešení: Dvourozměrnou normalitu dat ověříme pomocí dvourozměrného tečkového diagramu.



Vidíme, že předpoklad dvourozměrné normality je oprávněný.

Vypočteme výběrové průměry: $M_1 = \frac{1}{n}\sum_{i=1}^{n} X_i = 18{,}267$, $M_2 = \frac{1}{n}\sum_{i=1}^{n} Y_i = 83{,}6$, výběrové rozptyly: $S_1^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - M_1)^2 = 5{,}6381$, $S_2^2 = \frac{1}{n-1}\sum_{i=1}^{n}(Y_i - M_2)^2 = 121{,}4$, výběrovou kovarianci: $S_{12} = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - M_1)(Y_i - M_2) = 24{,}2571$ a výběrový koeficient korelace: $R_{12} = \frac{S_{12}}{S_1 S_2} = 0{,}927$. Testová statistika: $T = \frac{R_{12}\sqrt{n-2}}{\sqrt{1 - R_{12}^2}} = 8{,}597$, kvantil $t_{0,995}(13) = 3{,}012$, kritický obor $W = \langle -\infty, -3{,}012\rangle \cup \langle 3{,}012, \infty)$. Protože $T \in W$, hypotézu o nezávislosti veličin X a Y zamítáme na hladině významnosti 0,01.

Stanovení 99% asymptotického intervalu spolehlivosti pro ρ: Vypočítáme $Z = \frac{1}{2}\ln\frac{1+R_{12}}{1-R_{12}} = \frac{1}{2}\ln\frac{1+0{,}927}{1-0{,}927} = 1{,}637$. Meze 99% asymptotického intervalu spolehlivosti pro ρ jsou $\mathrm{tgh}\left(Z \pm \frac{u_{1-\alpha/2}}{\sqrt{n-3}}\right) = \mathrm{tgh}\left(1{,}637 \pm \frac{2{,}5758}{\sqrt{12}}\right)$, tedy $0{,}7131 < \rho < 0{,}983$ s pravděpodobností přibližně 0,99.

Práce se systémem STATISTICA

Téma: Analýza závislosti dvou veličin

Příklad 1.: **Testování nezávislosti nominálních veličin**
V roce 1950 zkoumali Yule a Kendall barvu očí a vlasů u 6800 mužů.

Barva očí	Barva vlasů			
	světlá	kaštanová	černá	rezavá
modrá	1768	807	180	47
šedá nebo zelená	946	1387	746	53
hnědá	115	438	288	16

Na asymptotické hladině významnosti 0,05 testujte hypotézu o nezávislosti barvy očí a barvy vlasů. Vypočtěte Cramérův koeficient. Simultánní četnosti znázorněte graficky.
Návod: Vytvořte nový datový soubor o 12 případech a třech proměnných (OCI, VLASY, CETNOST). Do proměnné OCI napište varianty barvy očí $x_{[1]} = 1$ (modrá), $x_{[2]} = 2$ (šedá nebo zelená), $x_{[3]} = 3$ (hnědá), přičemž každá varianta se objeví čtyřikrát pod sebou. Do proměnné VLASY napište třikrát pod sebe všechny varianty $y_{[1]} = 1$ (světlá), $y_{[2]} = 2$ (kaštanová), $y_{[3]} = 3$ (černá), $y_{[4]} = 4$ (rezavá). Statistics - Basic Statistics/Tables - Tables and Banners - Specify Tables – List 1 OCI, List 2 VLASY, OK, Weight - CETNOST Status On, OK – Statistics for two way tables - zaškrtněte Pearson & M-L Chi -square, Phi & Cramer's V – Advanced - Detailed two-way tables. Ve výstupní tabulce najdete mj. hodnotu testové stastistiky (Chi-square = 1073,51) s počtem stupňů volnosti (df = 6) a odpovídající p-hodnotou (p = 0,0000) i Cramérův koeficient (V = 0,281). Pro grafické znázornění četností se vraťte do Crosstabulation Table Results – Advanced – 3D histograms. Po vytvoření grafu je nutné manuálně zvětšit rozsah zobrazovaných hodnot na osách x a y.
Pomocí STATISTIKY je možno lehce ověřit splnění podmínek dobré aproximace (tzn., že teoretické četností mají být aspoň v 80% případů větší než 5 a ve zbylých 20% případů nemají klesnout pod 2. Teoretické četnosti se vypočítají tak, že v Options zaškrtneme Expected frequencies. V našem případě jsou podmínky dobré aproximace splněny.

Příklad 2.: **Fisherův faktoriálový test**
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.
Návod: Podle návodu z předešlého příkladu vytvořte datový soubor. Proměnné budou mít názvy POHLAVI, NAPOJ a CETNOST. Statistics - Basic Statistics/Tables - Tables and Banners - Specify Tables – List 1 POHLAVI, List 2 NAPOJ, OK, Weight - CETNOST Status On, OK – Options - Statistics for two way tables - zaškrtněte Fisher exact, Yates, McNemar (2x2) – Advanced - Detailed two-way tables. Ve výstupní tabulce je mimo jiné uvedena p-hodnota pro oboustranný (Fisher exact, two tailed) a jednostranný test (Fisher exact, one tailed). 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.

Příklad 3.: **Podíl šancí**
18 mužů onemocnělo určitou chorobou. Někteří z nich se léčili, jiní ne. Někteří se uzdravili, jiní zemřeli. Údaje jsou uvedeny ve čtyřpolní kontingenční tabulce.

přežití	léčení	
	ano	ne
ano	5	3
ne	6	4

Vypočtěte podíl šancí a sestrojte 95% asymptotický interval spolehlivosti pro podíl šancí. Pomocí tohoto intervalu spolehlivosti testujte na asymptotické hladině významnosti 0,05 hypotézu, že přežití nezávisí na léčení.

Návod: Podíl šancí vypočteme ručně. $OR = \frac{ac}{bd} = \frac{5 \cdot 4}{3 \cdot 6} = \frac{20}{18} = \frac{10}{9} = 1,\dot{1}$. (Protože podíl šancí je větší než 1, je zřejmě výhodnější se nechat léčit.) Dolní a horní mez intervalu spolehlivosti pro OR zjistíme pomocí STATISTIKY. Vytvoříme datový soubor o dvou proměnných DM a HM a dvou případech. Do Long Name proměnné DM napíšeme vzorec pro dolní mez: =exp(log(10/9)-sqrt((1/5+1/3+1/6+1/4)*VNormal(0,975;0;1)) a analogicky zjistíme horní mez. Výsledek: 0,1645 < OR < 7,506 s pravděpodobností přibližně 0,95. Protože tento interval spolehlivosti obsahuje 1, nelze na asymptotické hladině významnosti 0,05 zamítnout hypotézu, že přežití nezávisí na léčení.

Příklad 4.: **Testování nezávislosti ordinálních veličin**
Dva lékaři hodnotili stav sedmi pacientů po témž chirurgickém zákroku tak, aby nejvyšší pořadí měl nejtěžší případ.

č. pacienta	1	2	3	4	5	6	7
1. lékař	4	1	6	5	3	2	7
2. lékař	4	2	5	6	1	3	7

Vypočtěte Spearmanův koeficient pořadové korelace a na hladině významnosti 0,05 testujte hypotézu, že hodnocení obou lékařů jsou nezávislá.
Návod: Vytvořte datový soubor o sedmi případech a dvou proměnných X a Y. Statistics - Nonparametrics - Correlations – First variable list X, Second variable list Y, OK, Spearman R. Ve výstupní tabulce najdete Spearmanův koeficient a p-hodnotu. Nulová hypotéza se zamítá na hladině významnosti 0,05, protože p-hodnota = 0,013697 ≤ 0,05.

Příklad 5.: **Testování nezávislosti intervalových a poměrových veličin**
Máme k dispozici výsledky testů ze dvou předmětů zjištěné u osmi náhodně vybraných studentů určitého oboru.
Počet bodů z 1. testu: 80 50 36 58 72 60 56 68
Počet bodů z 2. testu: 65 60 35 39 48 44 48 61
Nakreslete dvourozměrný tečkový diagram, vypočtěte výběrový korelační koeficient, sestrojte 95% interval spolehlivosti pro korelační koeficient a na hladině významnosti 0,05 testujte hypotézu o nezávislosti výsledků obou testů.
Návod: Vytvořte datový soubor o dvou proměnných X a Y a osmi případech. Obvyklým způsobem zobrazte dvourozměrný tečkový diagram, s jehož pomocí posoudíte dvourozměrnou normalitu dat.
Testování hypotézy o nezávislosti: Statistics - Basic Statistics /Tables - Correlation matrices – OK - One variable list X,Y, OK – OK - Display r, p-levels and N's - Summary. Ve výstupní tabulce je hodnotu výběrového korelačního koeficientu R_{12} (r=0,6264, tzn. že mezi X a Y

existuje nepříliš silná přímá lineární závislost) a p-hodnotu pro test hypotézy o nezávislosti ($p=0,097$, H_0 tedy nelze zamítnout na hladině významnosti 0,05).
Výpočet mezí intervalu spolehlivosti: k datovému souboru přidáte dvě nové proměnné DM a HM. Do Long name proměnné DM zapište příkaz
= TanH(0,5*log((1+0,6264)/(1-0,6264))-VNormal(0,975;0;1)/sqrt(8-3))
a do Long name proměnné HM zapište příkaz
= TanH(0,5*log((1+0,6264)/(1-0,6264))+VNormal(0,975;0;1)/sqrt(8-3))
95% interval spolehlivosti pro ρ má tedy meze –0,14013 a 0,923454, tudíž hypotézu o nezávislosti veličin X, Y nelze zamítnout na hladině významnosti 0,05.
Poznámka: Pokud známe výběrový koeficient korelace a rozsah výběru, můžeme test nezávislosti veličin X, Y provést pomocí Probabilty Calculator. Statistics – Probability Calculator – Correlation – zadáme n a r, zaškrtneme Compute ρ from r – Compute.

Příklad 6.: **Porovnání dvou korelačních koeficientů**
Lékařský výzkum se zabýval sledováním koncentrací látek A a B v moči pacientů trpících určitou ledvinovou chorobou. U 100 zdravých osob činil výběrový koeficient korelace mezi koncentracemi obou látek 0,65 a u 142 osob trpících zmíněnou chorobou byl 0,37. Za předpokladu dvourozměrné normality dat testujte na hladině významnosti 0,05 hypotézu, že korelační koeficienty se neliší.
Návod: Statistics - Basic Statistics and Tables – Difference tests: r, %, means – OK – Difference between two correlation coefficients a vyplňte hodnoty korelačních koeficientů a rozsahy obou výběrů. Vypočtenou p-hodnotu porovnejte se zvolenou hladinou významnosti.