M5VM05 Statistické modelování

12. Analýza závislosti dvou veličin

Jan Koláček (kolacek@math.muni.cz)

Ústav matematiky a statistiky, Přírodovědecká fakulta, Masarykova univerzita, Brno

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.

Testování nezávislosti nominálních veličin

Nechť X, Y jsou dvě nominální náhodné veličiny. 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:

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

Testování nezávislosti nominálních veličin

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 $K \geq \chi^2_{1-\alpha}((r-1)(s-1))$.

Definice 1

Výraz $\frac{n_{j.}n_{.k}}{n}$ se nazývá **teoretická četnost**.

Poznámka 2 (Podmínka dobré aproximace)

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.

Měření síly závislosti

Definice 3

Cramérův koeficient je tvaru

$$V = \sqrt{\frac{K}{n(m-1)}},$$

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.

Příklad

Příklad 1

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.

Příklad

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:

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

Čtyřpolní tabulky

Speciálním případem kontingenčních tabulek, kdy $r = s = 2$ jsou čtyřpolní tabulky. Zavádí se pro ně jiné značení.

Definice 4

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

Čtyřpolní tabulky

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 prvních okolností je $\frac{a}{c}$, za druhých okolností je $\frac{b}{d}$.

Definice 5

Podíl šancí (odds ratio) ve čtyřpolní tabulce je definován jako $OR = \frac{ad}{bc}$.

Čtyřpolní tabulky

Věta 6

Pomocí $100(1-\alpha)\%$ *asymptotického intervalu spolehlivosti pro podíl šancí lze na asymptotické hladině významnosti* α *testovat hypotézu o nezávislosti nominálních veličin* X *a* Y*. Asymptotický* $100(1-\alpha)\%$ *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*.*

Příklad

Příklad 2

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

Příklad

Řešení

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

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

$\ln dm = 0,832 - 0,439 \cdot 1,96 = -0,028,\ \ln hm = 0,832 + 0,439 \cdot 1,96 = 1,692 \ \Rightarrow\ dm = e^{-0,28} = 0,972,\ hm = e^{1,692} = 5,433$

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.

Testování nezávislosti ordinálních veličin

Nechť X, Y jsou dvě ordinální náhodné veličiny. Pořídíme dvourozměrný náhodný výběr $(X_1, Y_1), \ldots, (X_n, Y_n)$ z rozložení, jímž se řídí náhodný vektor (X, Y). Označíme R_i pořadí náhodné veličiny X_i a Q_i pořadí náhodné veličiny $Y_i, i = 1, \ldots, n$. 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:

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

Testování nezávislosti ordinálních veličin

H_0 zamítáme na hladině významnosti α

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

$r_{S,1-\alpha}(n)$ 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

$$|r_s| \geq \frac{u_{1-\alpha/2}}{\sqrt{n-1}}$$

Poznámka 7

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.

Příklad

Příklad 3

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.*

Příklad

Řešení

$$r_s = 1 - \frac{6}{7(7^2-1)} \Big[(4-4)^2 + (1-2)^2 + (6-5)^2 + (3-1)^2 + (2-3)^2 + (7-7)^2 \Big] = 0,857$$

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

Testování nezávislosti intervalových či poměrových veličin

Pearsonův koeficient korelace

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

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

Připomeneme jeho vlastnosti:

1. $R(X,X) = 1$
2. $R(X,Y) = R(Y,X)$
3. $R(a+bX, c+dY) = \operatorname{sgn}(bd) R(X,Y)$
4. $-1 \leq R(X,Y) \leq 1$ a rovnosti je dosaženo tehdy a jen tehdy, když existují reálné konstanty a, b, kde $b \neq 0$ 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.

Výběrový koeficient korelace

Definice 8

Z dvourozměrného náhodného výběru $(X_1, Y_1), \ldots, (X_n, Y_n)$ můžeme stanovit:

1. výběrové průměry $M_1 = \frac{1}{n}\sum\limits_{i=1}^{n} X_i$, $M_2 = \frac{1}{n}\sum\limits_{i=1}^{n} Y_i$,
2. výběrové rozptyly

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

3. výběrovou kovarianci

$$S_{12} = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - M_1)(Y_i - M_2)$$

S jejich pomocí zavedeme **výběrový koeficient korelace**

$$R_{12} = \frac{S_{12}}{S_1 S_2} \qquad \text{pro } S_1 S_2 > 0.$$

Koeficient korelace dvourozměrného normálního rozdělení

Věta 9

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

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

přičemž $\mu_1 = E(X), \mu_2 = E(Y), \sigma_1^2 = D(X), \sigma_2^2 = D(Y), \rho = R(X,Y).$
Pak marginální hustoty jsou:

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

Věta 10

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

Koeficient korelace dvourozměrného normálního rozdělení

Věta 11

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:*

$$W = (-\infty, -t_{1-\alpha/2}(n-2)\rangle \cup \langle t_{1-\alpha/2}(n-2), \infty),$$

proti levostranné alternativě: $W = (-\infty, -t_{1-\alpha}(n-2)\rangle$
a proti pravostranné alternativě: $W = \langle t_{1-\alpha}(n-2), \infty)$. H_0 *zamítáme na hladině významnosti* α*, když* $T \in W$*.*

Příklad

Příklad 4

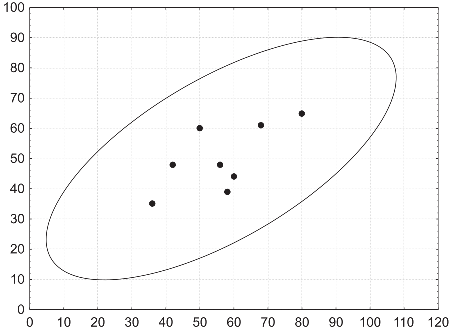
Máme k dispozici výsledky testů ze dvou předmětů zjištěné u osmi náhodně vybraných studentů určitého oboru.

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.

Příklad

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 : Dvourozměrný tečkový diagram

Příklad

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: $W = \langle 1,9432; \infty)$. Protože $T \in W$, hypotézu o neexistenci kladné korelace výsledků z 1. a 2. testu zamítáme na hladině významnosti 0,05.

Porovnání koeficientu korelace s danou konstantou

Věta 12

Nechť c je reálná konstanta. Testujeme $H_0 : \rho = c$ proti $H_1 : \rho \neq c$. 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 $\mathrm{N}(0,1)$, 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

Příklad 5

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.

Porovnání dvou koeficientů korelace

Věta 13

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 koeficient korelace 1. výběru a R_{12}^* výběrový koeficient korelace 2. výběru. Položme

$$Z = \frac{1}{2} \ln \frac{1 + R_{12}}{1 - R_{12}} \quad \text{a} \quad Z^* = \frac{1}{2} \ln \frac{1 + R_{12}^*}{1 - R_{12}^*}.$$

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 tedy je

$$W = (-\infty, -u_{1-\alpha/2}\rangle \cup \langle u_{1-\alpha/2}, \infty).$$

Příklad

Příklad 6

Lékařský výzkum se zabýval sledováním koncentrací látek A a B v moči pacientů trpících určitou ledvinovou chorobou. U 100 zdravých jedinců činil výběrový koeficient korelace 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 se koeficienty korelace v obou skupinách neliší.

Řešení

$Z = \frac{1}{2}\ln\frac{1+0{,}65}{1-0{,}65} = 0,7753,$

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

$U = \frac{0{,}7753-0{,}3884}{\sqrt{\frac{1}{100-3}+\frac{1}{142-3}}} = 2,9242,$

$u_{0{,}975} = 1,96, W = (-\infty, -1,96\rangle \cup \langle 1,96, \infty).$

Protože $U \in W, H_0$ zamítáme na asymptotické hladině významnosti 0,05.

Interval spolehlivosti pro koeficient korelace

Věta 14

Jestliže dvourozměrný náhodný výběr rozsahu n pochází z dvourozměrného normálního rozložení, jehož koeficient korelace se příliš neliší od nuly $(|\rho| < 0{,}5)$ a rozsah výběru je dostatečně velký $(n \geq 100)$, lze odvodit, že $100(1-\alpha)\%$ 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 následujícího tvrzení.

Interval spolehlivosti pro koeficient korelace

Věta 15

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 $100(1-\alpha)\%$ 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.

Interval spolehlivosti pro koeficient korelace

Poznámka 16

Jelikož $Z = \text{arctgh}\, R_{12}$*, dostáváme* $R_{12} = \text{tgh}\, Z$ *a meze intervalu spolehlivosti pro* ρ *můžeme psát ve tvaru*

$$\text{tgh}\left(Z \pm \frac{u_{1-\alpha/2}}{\sqrt{n-3}}\right), \text{ přičemž } \text{tgh}\, x = \frac{e^x - e^{-x}}{e^x + e^{-x}}.$$

Příklad

Příklad 7

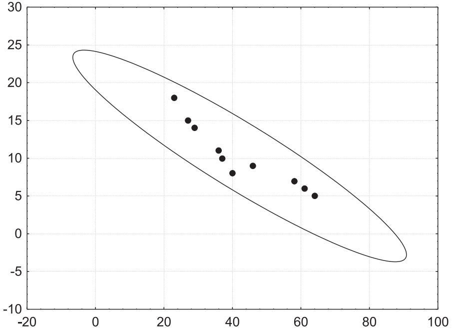
Pracovník personálního oddělení určité firmy zkoumá, zda existuje vztah mezi počtem dní absence za rok (veličina Y) a věkem pracovníka (veličina X). Proto náhodně vybral údaje o 10 pracovnících.

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 koeficient korelace a na hladině významnosti 0,05 testujte hypotézu, že X a Y jsou nezávislé náhodné veličiny. Sestrojte 95% asymptotický interval spolehlivosti pro skutečný koeficient korelace ρ.

Příklad

Řešení. Předpoklad o dvourozměrné normalitě dat ověříme orientačně pomocí dvourozměrného tečkového diagramu, viz. Obr. 2.



Obrázek : Dvourozměrný tečkový diagram

Vzhled diagramu svědčí o tom, že předpoklad je oprávněný.

Příklad

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 $W = (-\infty, -2,306\rangle \cup \langle 2,306, \infty)$. 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.

Úlohy k procvičení

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*

[hypotézu o nezávislosti pohlaví a pedagogické hodnosti nezamítáme, Cramérův koeficient: 0,187]

Úlohy k procvičení

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.

[$r_S = 0{,}715$, nulovou hypotézu zamítáme]

Úlohy k procvičení

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

[$r_{12} = \frac{s_{12}}{s_1 s_2} = 0{,}927$, hypotézu o nezávislosti veličin X a Y zamítáme, IS pro $\rho :\ (0{,}7131; 0{,}983)$]

Úlohy k procvičení

Příklad 4

Nechť $(X_1, Y_1), \ldots, (X_{16}, Y_{16})$ *je náhodný výběr z dvourozměrného normálního rozložení. Výběrový koeficient korelace* R_{XY} *nabyl hodnoty* $-0,87$. *Jestliže provedeme transformaci* $U_i = 1 + 3X_i$, $V_i = -3 - Y_i$, $i = 1, \ldots, 16$, *jakou hodnotu nabude výběrový koeficient korelace* R_{UV}*?*

$[R_{UV} = 0,87]$

Úlohy k procvičení

Příklad 5

400 náhodně vybraných pracovníků potravinářského podniku bylo dotázáno na příčiny nespokojenosti na pracovišti. Výsledky jsou uvedeny v tabulce:

kategorie pracovníků	*hlavní příčina nespokojenosti*				
	pracovní prostředí	*špatné vztahy*	*organizace práce*	*výdělek*	*jiné*
dělníci	*80*	*50*	*75*	*40*	*55*
THP	*10*	*10*	*25*	*30*	*25*

Na hladině významnosti 0,05 testujte hypotézu, že hlavní příčina nespokojenosti nezávisí na kategorii, do níž je pracovník zařazen. Vypočtěte Cramérův koeficient.

[hypotézu o nezávislosti zamítáme, Cramérův koeficient je $V = 0,25$]