

Hodnocení závislosti dvou náhodných veličin

Motivace

Při zpracování dat se setkáváme s dvojicemi veličin nominálního, ordinálního, intervalového a poměrového typu, např.:

nominální veličiny: rodinný stav ženicha a nevěsty – svobodný/á, rozvedený/á, vdovec/vdova,

ordinální veličiny: hodnocení softwarových produktů odbornou a laickou porotou na pětibodové škále,

intervalové veličiny: teplota měřená ve stupních Celsia na dvou meteorologických stanicích,

poměrové veličiny: roční příjem manžela a manželky.

Máme-li k dispozici n objektů, na nichž zjišťujeme hodnoty dvou veličin X a Y , můžeme testovat hypotézu, že veličiny X a Y jsou nezávislé.

Např. nás zajímá, zda barva očí a barva vlasů jsou ve sledované populaci jedinců nezávislé.

Intenzitu případné závislosti měří různé koeficienty, které nabývají hodnot od -1 do 1 nebo od 0 do 1 . Čím je absolutní hodnota takového koeficientu bližší 1 , tím je závislost mezi danými dvěma veličinami silnější a čím je bližší 0 , tím je slabší.

Závislost dvou veličin nominálního typu

Kontingenční tabulky

Nechť X, Y jsou dvě nominální náhodné veličiny (tj. obsahová interpretace je možná jenom u relace rovnosti). Nechť X nabývá variant $x_{[1]}, \dots, x_{[r]}$ a Y nabývá variant $y_{[1]}, \dots, y_{[s]}$.

Pořídíme dvourozměrný náhodný výběr $(X_1, Y_1), \dots, (X_n, Y_n)$ rozsahu n z rozložení, kterým se řídí dvourozměrný diskrétní náhodný vektor (X, Y) . Zjištěné absolutní simultánní četnosti n_{jk} dvojice variant $(x_{[j]}, y_{[k]})$ uspořádáme do kontingenční tabulky:

	y	$y_{[1]}$	\dots	$y_{[s]}$	$n_{j\cdot}$
x	n_{jk}	n_{j1}	\dots	n_{js}	$n_{j\cdot}$
$x_{[1]}$		n_{11}	\dots	n_{1s}	$n_{1\cdot}$
\dots		\dots	\dots	\dots	\dots
$x_{[r]}$		n_{r1}	\dots	n_{rs}	$n_{r\cdot}$
$n_{\cdot k}$		$n_{\cdot 1}$	\dots	$n_{\cdot s}$	n

$n_{j\cdot} = n_{j1} + \dots + n_{js}$ je marginální absolutní četnost varianty $x_{[j]}$

$n_{\cdot k} = n_{1k} + \dots + n_{rk}$ je marginální absolutní četnost varianty $y_{[k]}$

Testování hypotézy o nezávislosti

Testujeme nulovou hypotézu H_0 : X, Y jsou stochasticky nezávislé náhodné veličiny proti alternativě H_1 : X, Y nejsou stochasticky nezávislé náhodné veličiny.

Kdyby náhodné veličiny X, Y byly stochasticky nezávislé, pak by platil multiplikační vztah

$\forall j = 1, \dots, r, \forall k = 1, \dots, s: \pi_{jk} = \pi_j \cdot \pi_k$ neboli $\frac{n_{jk}}{n} = \frac{n_{j.}}{n} \cdot \frac{n_{.k}}{n}$, tj. $n_{jk} = \frac{n_{j.} \cdot n_{.k}}{n}$. Číslo $\frac{n_{j.} \cdot n_{.k}}{n}$ se nazývá **teoretická**

četnost dvojice variant $(x_{[j]}, y_{[k]})$.

Testová statistika:
$$K = \sum_{j=1}^r \sum_{k=1}^s \frac{\left(n_{jk} - \frac{n_{j.} \cdot n_{.k}}{n} \right)^2}{\frac{n_{j.} \cdot n_{.k}}{n}}.$$

Platí-li H_0 , pak K se asymptoticky řídí rozložením $\chi^2((r-1)(s-1))$.

Kritický obor: $W = \langle \chi^2_{1-\alpha}((r-1)(s-1)), \infty \rangle$.

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

Podmínky dobré aproximace

Rozložení statistiky K lze aproximovat rozložením $\chi^2((r-1)(s-1))$, pokud teoretické četnosti $\frac{n_{j.} \cdot n_{.k}}{n}$ 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

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 k 1, tím je závislost mezi X a Y těsnější, čím blíže je k 0, tím je tato závislost volnější.

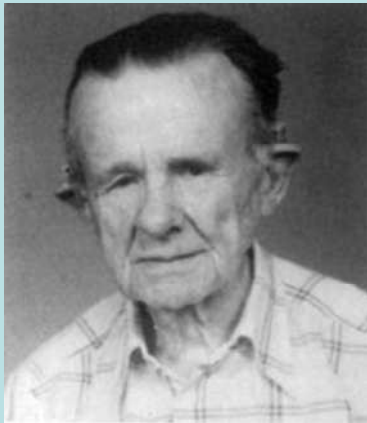
Význam hodnot Cramérova koeficientu:

mezi 0 až 0,1 ... zanedbatelná závislost,

mezi 0,1 až 0,3 ... slabá závislost,

mezi 0,3 až 0,7 ... střední závislost,

mezi 0,7 až 1 ... silná závislost.



Carl Harald Cramér (1893 – 1985): Švédský matematik

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 (veličina X) a typ školy, na kterou se hlásí (veličina Y). Výsledky jsou zaznamenány v kontingenční tabulce:

Sociální skupina	Typ školy			n_j
	univerzitní	technický	ekonomický	
I	50	30	10	90
II	30	50	20	100
III	10	20	30	60
IV	50	10	50	110
n_k	140	110	110	360

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

Řešení:

Nejprve vypočteme všech 12 teoretických četností:

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

$$\frac{n_{1.}n_{.1}}{n} = \frac{90 \cdot 140}{360} = 35, \quad \frac{n_{1.}n_{.2}}{n} = \frac{90 \cdot 110}{360} = 27,5, \quad \frac{n_{1.}n_{.3}}{n} = \frac{90 \cdot 110}{360} = 27,5,$$

$$\frac{n_{2.}n_{.1}}{n} = \frac{100 \cdot 140}{360} = 38,9, \quad \frac{n_{2.}n_{.2}}{n} = \frac{100 \cdot 110}{360} = 30,6, \quad \frac{n_{2.}n_{.3}}{n} = \frac{100 \cdot 110}{360} = 30,6,$$

$$\frac{n_{3.}n_{.1}}{n} = \frac{60 \cdot 140}{360} = 23,3, \quad \frac{n_{3.}n_{.2}}{n} = \frac{60 \cdot 110}{360} = 18,3, \quad \frac{n_{3.}n_{.3}}{n} = \frac{60 \cdot 110}{360} = 18,3,$$

$$\frac{n_{4.}n_{.1}}{n} = \frac{110 \cdot 140}{360} = 42,8, \quad \frac{n_{4.}n_{.2}}{n} = \frac{110 \cdot 110}{360} = 33,6, \quad \frac{n_{4.}n_{.3}}{n} = \frac{110 \cdot 110}{360} = 33,6$$

Kontingenční tabulka teoretických četností:

Sociální skupina	Typ školy		
	univerzitní	technický	ekonomický
I	35	27,5	27,5
II	38,9	30,6	30,6
III	23,3	18,3	18,3
IV	42,8	33,6	33,6

Vidíme, že podmínky dobré aproximace jsou splněny, všechny teoretické četnosti převyšují číslo 5.

$$\text{Dosadíme do vzorce pro testovou statistiku K: } K = \frac{(50 - 35)^2}{35} + \frac{(30 - 27,5)^2}{27,5} + \dots + \frac{(50 - 33,6)^2}{33,6} = 76,84.$$

Dále stanovíme kritický obor:

$$W = \langle \chi^2_{1-\alpha}((r-1)(s-1)), \infty \rangle = \langle \chi^2_{0,95}((4-1)(3-1)), \infty \rangle = \langle \chi^2_{0,95}(6), \infty \rangle = \langle 12,6, \infty \rangle$$

Protože $K \in W$, hypotézu o nezávislosti typu školy a sociální skupiny zamítáme na asymptotické hladině významnosti 0,05.

Výpočet pomocí systému STATISTICA:

Vytvoříme nový datový soubor o třech proměnných (X - sociální skupina, Y – typ školy, četnost) a 12 případech:

	1 X	2 Y	3 četnost
1	I	univerzitní	50
2	I	technický	30
3	I	ekonomický	10
4	II	univerzitní	30
5	II	technický	50
6	II	ekonomický	20
7	III	univerzitní	10
8	III	technický	20
9	III	ekonomický	30
10	IV	univerzitní	50
11	IV	technický	10
12	IV	ekonomický	50

Statistiky – Základní statistiky/tabulky – OK – Specif. Tabulky – List 1 X, List 2 Y – OK, zapneme proměnnou vah četnost – OK, Výpočet – na záložce Možnosti zaškrtneme Očekávané četnosti. Dostaneme kontingenční tabulku teoretických četností:

Souhrnná tab.: Očekávané četnosti (typ školy)				
Četnost označených buněk > 10				
Pearsonův chí-kv. : 76,8359, sv=6, p=,000000				
X	Y univerzitní	Y technický	Y ekonomický	Řádk. součty
I	35,0000	27,5000	27,5000	90,0000
II	38,8889	30,5556	30,5556	100,0000
III	23,3333	18,3333	18,3333	60,0000
IV	42,7778	33,6111	33,6111	110,0000
Vš.skup.	140,0000	110,0000	110,0000	360,0000

Všechny teoretické četnosti jsou větší než 5, podmínky dobré aproximace jsou splněny. V záhlaví tabulky je uvedena hodnota testové statistiky $K = 76,8359$, počet stupňů volnosti 6 a odpovídající p-hodnota. Je velmi blízká 0, tedy na asymptotické hladině významnosti 0,05 zamítáme hypotézu o nezávislosti typu školy a sociální skupiny.

Hodnotu testové statistiky a Cramérův koeficient dostaneme také tak, že na záložce Možnosti zaškrtneme Pearsonův & M-V chí kvadrát a Cramérovo V, na záložce Detailní výsledky vybereme Detailní 2 rozm. tabulky.

Statist.	Chí-kvadr.	sv	p
Pearsonův chí-kv.	76,83589	df=6	p=,00000
M-V chí-kvadr.	84,53528	df=6	p=,00000
Fí	,4619881		
Kontingenční koeficient	,4193947		
Cramér. V	,3266749		

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		$n_{j\cdot}$
	$y_{[1]}$	$y_{[2]}$	
$x_{[1]}$	a	b	a+b
$x_{[2]}$	c	d	c+d
$n_{\cdot k}$	a+c	b+d	n

Test nezávislosti ve čtyřpolní tabulce

Testovou statistiku pro čtyřpolní kontingenční tabulku lze zjednodušit do tvaru:

$$K = \frac{n(ad - bc)^2}{(a + b)(c + d)(a + c)(b + d)}.$$

Platí-li hypotéza o nezávislosti veličin X, Y, pak K se asymptoticky řídí rozložením $\chi^2(1)$.

Kritický obor: $W = \langle \chi^2_{1-\alpha}(1), \infty \rangle$

Nulovou hypotézu zamítáme na asymptotické hladině významnosti α , když $K \in W$.

Povšimněte si, že za platnosti hypotézy o nezávislosti $ad = bc$.

Pro čtyřpolní tabulku navrhl R. A. Fisher přesný (exaktní) test nezávislosti známý jako **Fisherův faktoriálový test**.



Sir Ronald Aylmer Fisher (1890 – 1962): Britský statistik a genetik.

(Fisherův přesný test je popsán např. v knize K. Zvára: Biostatistika, Karolinum, Praha 1998. Princip spočívá v tom, že pomocí kombinatorických úvah se vypočítají pravděpodobnosti toho, že při daných marginálních četnostech dostaneme tabulky, které se od nulové hypotézy odchyľují aspoň tak, jako daná tabulka.)

Upozornění: STATISTICA poskytuje p-hodnotu pro Fisherův přesný test. Jestliže vyjde $p \leq \alpha$, pak hypotézu o nezávislosti zamítáme na hladině významnosti α .

Příklad: V náhodném výběru 50 obézních dětí ve věku 6 – 14 let byla zjišťována obezita rodičů. Veličina X – obezita matky, veličina Y – obezita otce. Výsledky průzkumu jsou uvedeny v kontingenční tabulce:

X	Y		n _{j.}
	ano	ne	
ano	15	9	24
ne	7	19	26
n _{k.}	22	28	50

Pomocí Fisherova exaktního testu ověřte, zda lze na hladině významnosti 0,05 zamítnout hypotézu o nezávislosti náhodných veličin X a Y.

Výpočet pomocí systému STATISTICA:

Vytvoříme datový soubor o třech proměnných X, Y (varianty 0 – neobézní, 1 – obézní) a četnost a čtyřech případech:

	1 X	2 Y	3 četnost
1	obézní	obézní	15
2	obézní	neobézní	9
3	neobézní	obézní	7
4	neobézní	neobézní	19

Statistiky – Základní statistiky/tabulky – OK – Specif. Tabulky – List 1 X, List 2 Y – OK, zapneme proměnnou vah četnost – OK, Výpočet – na záložce Možnosti zaškrtneme Fisher exakt., Yates, McNemar (2x2). Dostaneme výstupní tabulku:

Statist.	Statist. : X(2) x Y(2) (obezita rodicu)		
	Chí-kvadr.	sv	p
Pearsonův chí-kv.	6,410777	df=1	p=,01134
M-V chí-kvadr.	6,548348	df=1	p=,01050
Yatesův chí-kv.	5,048207	df=1	p=,02465
Fisherův přesný, 1-str.			p=,01188
2-stranný			p=,02163
McNemarův chí-kv. (A/D)	,2647059	df=1	p=,60691
(B/C)	,0625000	df=1	p=,80259

Vidíme, že p-hodnota pro Fisherův exaktní oboustranný test je 0,02163, tedy na hladině významnosti 0,05 zamítáme hypotézu, že obezita matky a otce spolu nesouvisí.

Podíl šancí ve čtyřpolní kontingenční tabulce

Ve čtyřpolních tabulkách používáme charakteristiku $OR = \frac{ad}{bc}$, která se nazývá výběrový **podíl šancí** (odds ratio). Považujeme ho za odhad neznámého teoretického podílu šancí $op = \frac{\pi_{11}\pi_{22}}{\pi_{21}\pi_{12}}$. Můžeme si představit, že pokus se provádí za dvo-

jích 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 tedy

$$OR = \frac{ad}{bc}.$$

Jsou-li veličiny X, Y nezávislé, pak $\pi_{jk} = \pi_{j.}\pi_{.k}$, tudíž teoretický podíl šancí $op = 1$. Závislost veličin X, Y bude tím silnější, čím více se op bude lišit od 1. Avšak $op \in \langle 0, \infty \rangle$, tedy hodnoty op jsou kolem 1 rozmístěny nesymetricky. Z tohoto důvodu raději používáme logaritmus teoretického či výběrového podílu šancí.

Testování nezávislosti ve čtyřpolních tabulkách pomocí podílu šancí

Na asymptotické hladině významnosti α testujeme hypotézu H_0 : X, Y jsou stochasticky nezávislé náhodné veličiny (tj. $\ln \text{OR} = 0$) proti alternativě H_1 : X, Y nejsou stochasticky nezávislé náhodné veličiny (tj. $\ln \text{OR} \neq 0$).

Testová statistika $T_0 = \frac{\ln \text{OR}}{\sqrt{\frac{1}{a} + \frac{1}{b} + \frac{1}{c} + \frac{1}{d}}}$ se asymptoticky řídí rozložením $N(0,1)$, když nulová hypotéza platí.

Kritický obor: $W = (-\infty, -u_{1-\alpha/2}) \cup (u_{1-\alpha/2}, \infty)$.

Nulovou hypotézu tedy zamítáme na asymptotické hladině významnosti α , když se testová statistika realizuje v kritickém oboru W .

Testování nezávislosti lze provést též pomocí $100(1-\alpha)\%$ asymptotického intervalu spolehlivosti pro logaritmus podílu šancí OR , který je dán vzorcem:

$$(d, h) = \left(\ln \text{OR} - \sqrt{\frac{1}{a} + \frac{1}{b} + \frac{1}{c} + \frac{1}{d}} u_{1-\alpha/2}, \ln \text{OR} + \sqrt{\frac{1}{a} + \frac{1}{b} + \frac{1}{c} + \frac{1}{d}} u_{1-\alpha/2} \right)$$

Jestliže interval spolehlivosti neobsahuje 0, pak hypotézu o nezávislosti zamítneme na asymptotické hladině významnosti α .

Příklad (testování nezávislosti pomocí podílu šancí a pomocí statistiky K):

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:

a) Testování pomocí podílu šancí:

$OR = \frac{ad}{bc} = \frac{17 \cdot 58}{11 \cdot 39} = 2,298$. Podíl šancí nám říká, že uchazeč, který zapůsobil na komisi dobrým dojemem, má asi 2,3 x větší šanci na přijetí než uchazeč, který zapůsobil špatným dojemem.

Provedeme další pomocné výpočty:

$$\ln OR = 0,832,$$

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

Dosadíme do vzorců pro meze asymptotického intervalu spolehlivosti pro podíl šancí:

$$\ln d = \ln OR - \sqrt{\frac{1}{a} + \frac{1}{b} + \frac{1}{c} + \frac{1}{d}} u_{1-\alpha/2} = 0,832 - 0,439 \cdot 1,96 = -0,028, \ln h = \ln OR + \sqrt{\frac{1}{a} + \frac{1}{b} + \frac{1}{c} + \frac{1}{d}} u_{1-\alpha/2} = 0,832 + 0,439 \cdot 1,96 = 1,692$$

Protože interval (-0,028; 1,692) obsahuje číslo 0, na asymptotické hladině významnosti 0,05 nezamítáme hypotézu o nezávislosti dojmu u přijímací zkoušky a přijetí na fakultu.

b) Testování pomocí statistiky K:

přijetí	dojem		n _j
	dobrý	špatný	
ano	17	11	28
ne	39	58	97
n _k	56	69	125

Ověříme splnění podmínek dobré aproximace:

$$\frac{n_{1,n_1}}{n} = \frac{28 \cdot 56}{125} = 12,544, \quad \frac{n_{1,n_2}}{n} = \frac{28 \cdot 69}{125} = 15,456,$$

$$\frac{n_{2,n_1}}{n} = \frac{97 \cdot 56}{125} = 43,456, \quad \frac{n_{2,n_2}}{n} = \frac{97 \cdot 69}{125} = 53,544$$

Podmínky dobré aproximace jsou splněny.

Dosadíme do zjednodušeného vzorce pro testovou statistiku K:

$$K = \frac{n(ad - bc)^2}{(a + b)(c + d)(a + c)(b + d)} = \frac{125 \cdot (17 \cdot 58 - 11 \cdot 39)^2}{28 \cdot 97 \cdot 56 \cdot 69} = 3,6953$$

Kritický obor: $W = \langle \chi^2_{0,95}(1), \infty \rangle = \langle 3,841, \infty \rangle$.

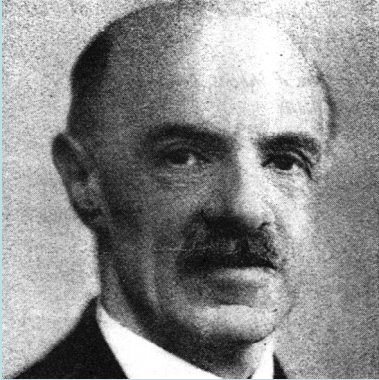
Protože testová statistika se nerealizuje k kritickému oboru, nulovou hypotézu nezamítáme na asymptotické hladině významnosti 0,05.

$$\text{Vypočteme ještě Cramérův koeficient: } V = \sqrt{\frac{K}{n(m-1)}} = \sqrt{\frac{3,6953}{125(2-1)}} = 0,1719$$

Vidíme, že mezi dojmem u přijímací zkoušky a přijetím na fakultu je pouze slabá závislost.

Závislost dvou veličin ordinálního typu

Spearmanův koeficient pořadové korelace



Charles Edward Spearman (1863 – 1945): Britský psycholog a statistik, zakladatel faktorové analýzy

Nechť X, Y jsou náhodné veličiny ordinálního typu (tj. obsahová interpretace je možná jenom u relace rovnosti a relace uspořádání).

Pořídíme dvourozměrný náhodný výběr $(X_1, Y_1), \dots, (X_n, Y_n)$ 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 , $i = 1, \dots, n$.

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

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 .

Vlastnosti Spearmanova koeficientu pořadové korelace

Pro Spearmanův koeficient pořadové korelace platí $-1 \leq r_s \leq 1$. 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.

Je-li $r_s = 1$ resp. $r_s = -1$, pak realizace $(x_i, y_i), i = 1, \dots, n$ daného náhodného výběru leží na nějaké rostoucí resp. klesající funkci.

Hodnoty r_s se nezmění, když provedeme vzestupnou transformaci původních dat.

Hodnoty r_s se vynásobí -1, když provedeme sestupnou transformaci původních dat.

Koeficient je symetrický.

Koeficient je rezistentní vůči odlehlým hodnotám.

Význam absolutní hodnoty Spearmanova koeficientu:

mezi 0 až 0,1 ... zanedbatelná pořadová závislost,

mezi 0,1 až 0,3 ... slabá pořadová závislost,

mezi 0,3 až 0,7 ... střední pořadová závislost,

mezi 0,7 až 1 ... silná pořadová závislost.

Spearmanův koeficient pořadové korelace se používá v situacích, kdy

- zkoumaná data mají ordinální charakter
- nelze předpokládat, že vztah mezi veličinami X, Y je lineární
- náhodný výběr nepochází z dvourozměrného normálního rozložení

Testování nezávislosti ordinálních veličin

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).

Jako testová statistika slouží Spearmanův koeficient pořadové korelace r_S .

Nulovou hypotézu zamítáme na hladině významnosti α ve prospěch

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

Asymptotická varianta 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: $W = (-\infty, -t_{1-\alpha/2}(n-2)) \cup (t_{1-\alpha/2}(n-2), \infty)$

Kritický obor pro levostrannou alternativu:

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

Kritický obor pro pravostrannou alternativu:

$$W = (t_{1-\alpha}(n-2), \infty).$$

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.

Příklad na testování pořadové nezávislosti

Dva lékaři hodnotili stav sedmi pacientů po téměř 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

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

Řešení:

Na hladině významnosti 0,05 testujeme 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

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

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

Výpočet pomocí systému STATISTICA

Vytvoříme datový soubor o dvou proměnných X (hodnocení 1. lékaře), Y (hodnocení 2. lékaře) a sedmi případech. Do proměnných X a Y zapíšeme zjištěná hodnocení.

	1 X	2 Y
1	4	4
2	1	2
3	6	5
4	5	6
5	3	1
6	2	3
7	7	7

Statistiky – Neparametrické statistiky – Korelace – OK – vybereme Vytvořit detailní report - Proměnné X, Y – OK – Spearmanův koef. R. Dostaneme tabulku

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

Spearmanův koeficient pořadové korelace nabývá hodnoty 0,857, testová statistika se realizuje hodnotou 3,721, odpovídající p-hodnota je 0,0137, tedy na asymptotické hladině významnosti 0,05 zamítáme hypotézu o pořadové nezávislosti hodnocení dvou lékařů ve prospěch oboustranné alternativy.

Závislost dvou veličin intervalového či poměrového typu

Pearsonův koeficient korelace



Karl Pearson (1857 – 1936): Britský statistik

Číslo

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

se nazývá Pearsonův koeficient korelace.

(Pro výpočet Pearsonova koeficientu korelace musíme znát simultánní distribuční funkci $\Phi(x, y)$ v obecném případě resp. simultánní hustotu pravděpodobnosti $\varphi(x, y)$ ve spojitém případě resp. simultánní pravděpodobnostní funkci $\pi(x, y)$ v diskrétním případě.)

Vlastnosti Pearsonova koeficientu korelace

a) $R(a_1, Y) = R(X, a_2) = R(a_1, a_2) = 0$

b) $R(a_1 + b_1X, a_2 + b_2Y) = \text{sgn}(b_1b_2) R(X, Y) = \begin{cases} R(X, Y) & \text{pro } b_1b_2 > 0 \\ -R(X, Y) & \text{pro } b_1b_2 < 0 \end{cases}$

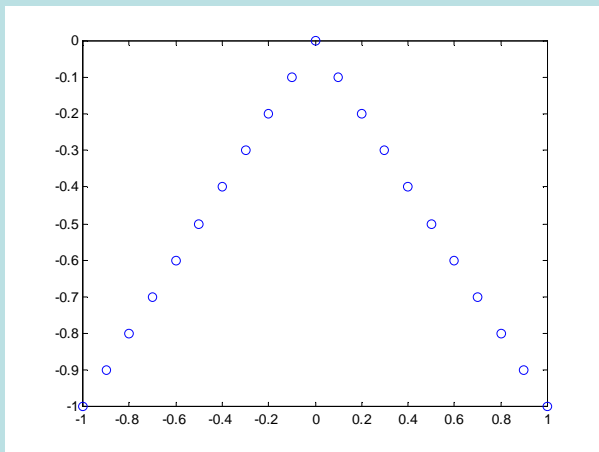
c) $R(X, X) = 1$ pro $D(X) \neq 0$, $R(X, X) = 0$ jinak

d) $R(X, Y) = R(Y, X)$

e) $|R(X, Y)| \leq 1$ a rovnost nastane tehdy a jen tehdy, když mezi veličinami X, Y existuje s pravděpodobností 1 úplná lineární závislost, tj. existují konstanty a, b tak, že pravděpodobnost $P(Y = a + bX) = 1$. Přitom $R(X, Y) = 1$, když $b > 0$ a $R(X, Y) = -1$, když $b < 0$. (Uvedená nerovnost se nazývá Cauchyova – Schwarzova – Buňakovského nerovnost.)

Z vlastností Pearsonova koeficientu korelace vyplývá, že se hodí pouze k měření těsnosti lineárního vztahu veličin X a Y . Při složitějších závislostech může dojít k paradoxní situaci, že Pearsonův koeficient korelace je nulový.

Ilustrace:



Definice nekorelovanosti

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 .)

Výběrový koeficient korelace

Nechť $(X_1, Y_1), \dots, (X_n, Y_n)$ náhodný výběr rozsahu n z dvourozměrného rozložení daného distribuční funkcí $\Phi(x,y)$. Z tohoto dvourozměrného náhodného výběru můžeme stanovit:

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

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

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

$$\text{výběrový koeficient korelace } R_{12} = \begin{cases} \frac{1}{n-1} \sum_{i=1}^n \frac{X - M_1}{S_1} \cdot \frac{Y - M_2}{S_2} = \frac{S_{12}}{S_1 S_2} & \text{pro } S_1 S_2 > 0 \\ 0 & \text{jinak} \end{cases} .$$

Vlastnosti Pearsonova koeficientu korelace se přenášejí i na výběrový koeficient korelace.

(Spearmanův koeficient pořadové korelace odpovídá Pearsonovu koeficientu korelace aplikovanému na pořadí.)

Pearsonův 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:

$$\varphi_1(x) = \int_{-\infty}^{\infty} \varphi(x, y) dy = \dots = \frac{1}{\sigma_1\sqrt{2\pi}} e^{-\frac{(x-\mu_1)^2}{2\sigma_1^2}},$$

$$\varphi_2(y) = \int_{-\infty}^{\infty} \varphi(x, y) dx = \dots = \frac{1}{\sigma_2\sqrt{2\pi}} e^{-\frac{(y-\mu_2)^2}{2\sigma_2^2}}.$$

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

Upozornění: nadále budeme předpokládat, že $(X_1, Y_1), \dots, (X_n, Y_n)$ je náhodný výběr rozsahu n z dvourozměrného normálního rozložení $N_2\left(\begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}, \begin{pmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{pmatrix}\right)$.

Předpoklad dvourozměrné normality lze orientačně ověřit pomocí dvourozměrného tečkového diagramu: tečky by měly zhruba rovnoměrně vyplnit vnitřek elipsovitého obrazce. Vrstevnice hustoty dvourozměrného normálního rozložení jsou totiž elipsy.

Do dvourozměrného tečkového diagramu můžeme ještě zakreslit $100(1-\alpha)\%$ elipsu konstantní hustoty pravděpodobnosti. Bude-li více než $100\alpha\%$ teček ležet vně této elipsy, svědčí to o porušení dvourozměrné normality. Bude-li mít hlavní osa elipsy kladnou resp. zápornou směrnici, znamená to, že mezi veličinami X a Y existuje určitý stupeň přímé resp. nepřímé lineární závislosti.

Testování hypotézy o nezávislosti

Na hladině významnosti α testujeme 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 $T_0 \sim t(n-2)$.

Kritický obor pro test H_0 proti

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

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

Příklad: Testování hypotézy o nezávislosti

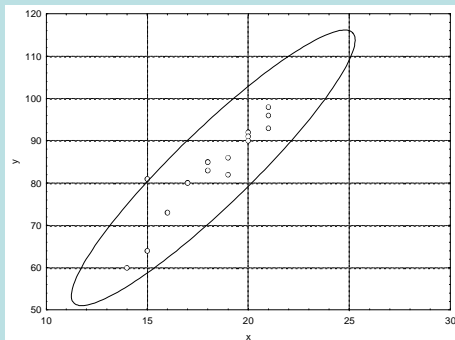
V dílně pracuje 15 dělníků. Byl u nich zjištěn počet směn odpracovaných za měsíc (náhodná veličina X) a počet zhotovených výrobků (náhodná veličina Y):

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.

Orientačně ověřte dvourozměrnou normalitu dat, vypočítejte výběrový koeficient korelace mezi X a Y a na hladině 0,01 testujte hypotézu o nezávislosti X a Y.

Řešení: Dvourozměrnou normalitu dat ověříme pomocí dvourozměrného tečkového diagramu.



Vidíme, že předpoklad dvourozměrné normality je oprávněný.

Vypočteme realizace

$$\text{výběrových průměrů: } m_1 = \frac{1}{n} \sum_{i=1}^n x_i = 18,267, m_2 = \frac{1}{n} \sum_{i=1}^n y_i = 83,6,$$

$$\text{výběrových rozptylů: } s_1^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - m_1)^2 = 5,6381, s_2^2 = \frac{1}{n-1} \sum_{i=1}^n (y_i - m_2)^2 = 121,4,$$

$$\text{výběrové kovariance: } s_{12} = \frac{1}{n-1} \sum_{i=1}^n (x_i - m_1)(y_i - m_2) = 24,2571,$$

$$\text{výběrového koeficientu korelace: } r_{12} = \frac{s_{12}}{s_1 s_2} = 0,927.$$

Realizace testové statistiky: $t_0 = \frac{r_{12} \sqrt{n-2}}{\sqrt{1-r_{12}^2}} = 8,912,$

kritický obor $W = (-\infty, -t_{0,995}(13)) \cup (t_{0,995}(13), \infty) = (-\infty, -3,012) \cup (3,012, \infty).$

Protože $t_0 \in W$, hypotézu o nezávislosti veličin X a Y zamítáme na hladině významnosti 0,01. S rizikem omylu nejvýše 1% jsme tedy prokázali, že mezi počtem směn odpracovaných za měsíc a počtem zhotovených výrobků existuje závislost.

Výpočet pomocí systému STATISTICA

Vytvoříme datový soubor o dvou proměnných X, Y a 15 případech. Dvourozměrnou normalitu dat ověříme pomocí dvou-rozměrného tečkového diagramu – viz výše.

Statistiky – Základní statistiky/tabulky – Korelační matice – OK – 1 seznam proměn. – X, Y – OK – na záložce Možnosti vybereme Zobrazit detailní tabulku výsledků – Výpočet.

Korelace (smeny a výrobky.sta) Označ. korelace jsou významné na hlad. p < ,05000 (Celé případy vynechány u ChD)											
Prom. X & prom. Y	Průměr	Sm.Odch.	r(X,Y)	r2	t	p	N	Konst. záv.: Y	Směr. záv: Y	Konst. záv.: X	Směrnic záv.: X
X	18,26667	2,37447									
X	18,26667	2,37447	1,000000	1,000000			15	0,000000	1,000000	0,000000	1,000000
X	18,26667	2,37447									
Y	83,60000	11,01817	0,927180	0,859663	8,923795	0,000001	15	5,010135	4,302365	1,562407	0,199812
Y	83,60000	11,01817									
X	18,26667	2,37447	0,927180	0,859663	8,923795	0,000001	15	1,562407	0,199812	5,010135	4,302365
Y	83,60000	11,01817									
Y	83,60000	11,01817	1,000000	1,000000			15	0,000000	1,000000	0,000000	1,000000

Výběrový koeficient korelace se realizoval hodnotou 0,92718, testová statistika nabyla hodnoty 8,924, odpovídající p-hodnota je 0,000001, tedy na hladině významnosti 0,01 zamítáme hypotézu o nezávislosti veličin X, Y.