11. Testování nezávislosti náhodných veličin

11.1. Motivace: Při zpracování dat se velmi často setkáme s úkolem zjistit, zda dvě náhodné veličiny jsou stochasticky nezávislé. Testování hypotézy o nezávislosti se provádí různými způsoby podle toho, jakého typu jsou dané náhodné veličiny – zda jsou nominální, ordinální, intervalové či poměrové. Nominální náhodné veličiny umožňují obsahovou interpretaci pouze u relace rovnosti, ordinální navíc ještě u relace uspořádání, intervalové pak navíc u operace rozdílu a poměrové i u operace podílu.
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.
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. Definice (definice kontingenční tabulky)
Nechť X,Y jsou dvě nominální náhodné veličiny (tj. obsahová interpretace je možná jenom u relace rovnosti). Nechť X nabývá variant $x_{[1]}, ..., x_{[r]}$ a Y nabývá variant $y_{[1]}, ..., y_{[s]}$.
Označme:

$\pi_{jk} = P(X = x_{[j]} \wedge Y = y_{[k]})$ … simultánní pravděpodobnost dvojice variant $(x_{[j]}, y_{[k]})$

$\pi_{j.} = P(X = x_{[j]})$ … marginální pravděpodobnost varianty $x_{[j]}$

$\pi_{.k} = P(Y = y_{[k]})$ … marginální pravděpodobnost varianty $y_{[k]}$

Simultánní a marginální pravděpodobnosti zapíšeme do kontingenční tabulky:

	y	$y_{[1]}$	…	$y_{[s]}$	$\pi_{j.}$
x	π_{jk}				
$x_{[1]}$		π_{11}	…	π_{1s}	$\pi_{1.}$
…		…	…	…	…
$x_{[r]}$		π_{r1}	…	π_{rs}	$\pi_{r.}$
$\pi_{.k}$		$\pi_{.1}$	…	$\pi_{.s}$	1

Nyní pořídíme dvourozměrný náhodný výběr 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	$y_{[1]}$	...	$y_{[s]}$	$n_{j.}$
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

$n_{j.} = n_{j1} + ... + n_{js}$ je marginální absolutní četnost varianty $x_{[j]}$
$n_{.k} = n_{1k} + ... + n_{rk}$ je marginální absolutní četnost varianty $y_{[k]}$

Simultánní pravděpodobnost π_{jk} odhadneme pomocí simultánní relativní četnosti $p_{jk} = \frac{n_{jk}}{n}$, marginální pravděpodobnosti $\pi_{j.}$ a $\pi_{.k}$ odhadneme pomocí marginálních relativních četností $p_{j.} = \frac{n_{j.}}{n}$ a $p_{.k} = \frac{n_{.k}}{n}$.

11.3. Věta (věta o testové statistice K)

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 multiplikativní vztah

$\forall j = 1,..,r, \forall k = 1,..,s$: $\pi_{jk} = \pi_{j.}\,\pi_{.k}$ neboli $\frac{n_{jk}}{n} = \frac{n_{j.}}{n} \cdot \frac{n_{.k}}{n}$, tj. $n_{jk} = \frac{n_{j.} n_{.k}}{n}$. Číslo $m_{jk} = \frac{n_{j.} n_{.k}}{n}$ se nazývá teoretická četnost dvojice variant $(x_{[j]}, y_{[k]})$.

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

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

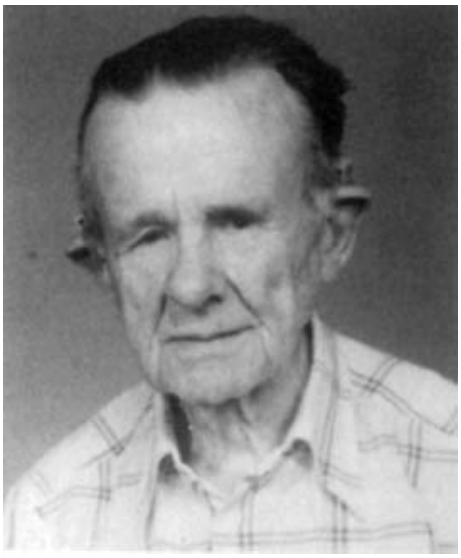
11.4. Poznámka (podmínky dobré aproximace)

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

11.5. Definice (definice Cramérova koeficientu, význam jeho hodnot)

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.
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

11.6. 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:

Nejprve vypočteme všech 12 teoretických četností:

$\frac{n_1 n_1}{n} = 30 =$, $\frac{n_1 n_2}{n} = 30 =$, $\frac{n_1 n_3}{n} = 30 =$, $\frac{n_1 n_4}{n} = 30 =$,

$\frac{n_2 n_1}{n} = 30 =$, $\frac{n_2 n_2}{n} = 30 =$, $\frac{n_2 n_3}{n} = 30 =$, $\frac{n_2 n_4}{n} = 30 =$,

$\frac{n_2 n_1}{n} = 30 =$, $\frac{n_2 n_2}{n} = 30 =$, $\frac{n_2 n_3}{n} = 30 =$, $\frac{n_2 n_4}{n} = 30 =$,

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

Nyní dosadíme do vzorce pro testovou statistiku K:

$K = \frac{(\quad)^2}{\quad} + \frac{(\quad)^2}{\quad} + \cdots + \frac{(\quad)^2}{\quad} =$, r = 3, s = 4, $\chi^2_{0,95}(6) = 12{,}6$.

Protože K ≥ 12,6, hypotézu o nezávislosti typu školy a sociální skupiny zamítáme na asymptotické hladině významnosti 0,05. Cramérův koeficient:

$V = \sqrt{\frac{\quad}{36}} =$. Hodnota Cramérova koeficientu svědčí o tom, že mezi veličinami X a Y existuje středně silná závislost.

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 četnc
1	I	univerzi	5
2	I	technick	3
3	I	ekonom	1
4	II	univerzi	3
5	II	technick	5
6	II	ekonom	2
7	III	univerzi	1
8	III	technick	2
9	III	ekonom	3
10	IV	univerzi	5
11	IV	technick	1
12	IV	ekonom	5

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 (
Četnost označených buněk > 10
Pearsonův chí-kv. : 76,8359, sv=6, p				
X	Y univerzi	Y technic	Y ekonomi	Řádk součt
I	35,0(	27,5(	27,5(	90,0(
II	38,8	30,5	30,5	100,0
III	23,3	18,3	18,3	60,0(
IV	42,7	33,6	33,6	110,0
Vs.sku	140,0	110,0	110,0	360,0

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 na záložce Možnosti zaškrtneme Pearsonův & M-V chí kvadrát a Cramérovo V a na záložce Detailní výsledky vybereme Detailní 2 rozm. tabulky.

	Statist. : X(4) x Y(3) (1		
Statist.	Chí-kva	sv	p
Pearsonův chí-k	76,83	df=	p=,00
M-V chí-kvadr.	84,53	df=	p=,00
Fi	,4619		
Kontingenční ko	,4193		
Cramér. V	,3266		

11.7. Definice (definice čtyřpolní kontingenční 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

Testová statistika K pro čtyřpolní kontingenční tabulku se dá zjednodušit do tvaru:

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

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

11.8. Věta (věta o testové statistice K pro čtyřpolní tabulky)
Testová statistika K pro čtyřpolní kontingenční tabulku se dá zjednodušit do tvaru: $K = \frac{n(ad-bc)^2}{(a+b)(c+d)(a+c)(b+d)}$
Kritický obor: $W = \langle \chi^2_{1-\alpha}, \infty \rangle$. Hypotézu o nezávislosti náhodných veličin X, Y tedy zamítáme na asymptotické hladině významnosti α, když $K \in W$.

11.9. Poznámka: U čtyřpolní KT lze rovněž použít následující podmínky dobré aproximace: $a + b > 5$, $c + d > (a + c)/3$.

11.10. Příklad: U 125 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:
Ověříme splnění podmínek dobré aproximace:
$a + b = 28 > 5$, $c + d = 97 > (a + c)/3 = 56/3 = 18{,}66$ – v pořádku
Dosadíme do zjednodušeného vzorce pro testovou statistiku K:
$K = \frac{n(ad-bc)^2}{(a+b)(c+d)(a+c)(b+d)} = \frac{125 \cdot (17 \cdot 58 - 11 \cdot 39)^2}{28 \cdot 97 \cdot 56 \cdot 69} = 5{,}595$
Kritický obor: $W = \langle \chi^2_{0{,}95}, \infty \rangle = \langle 3{,}841, \infty \rangle$.
Protože testová statistika se nerealizuje k kritickém oboru, nulovou hypotézu nezamítáme na asymptotické hladině významnosti 0,05.

11.11. Definice (definice podílu šancí)
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 $OR = \frac{ad}{bc}$.

11.12. Věta (asymptotický interval spolehlivosti pro podíl šancí a jeho využití k testování hypotézy o nezávislosti)
Asymptotický 100(1-α)% interval spolehlivosti pro skutečný podíl šancí má meze:

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

Jestliže interval spolehlivosti neobsahuje 1, pak hypotézu o nezávislosti zamítneme na asymptotické hladině významnosti α.

11.13. Příklad: Pro údaje z příkladu 11.10. vypočtěte a interpretujte podíl šancí, sestrojte 95% asymptotický interval spolehlivosti pro podíl šancí a s jeho pomocí testujte hypotézu, že přijetí na fakultu nezávisí na dojmu u přijímací zkoušky.
Řešení:
$OR = \frac{ad}{bc} = \frac{\cdot 58}{\cdot 35} = 29$. Podíl šancí nám říká, že uchazeč, který zapůsobil na komisi dobrým dojmem, má asi 2,3 x větší šanci na přijetí než uchazeč, který zapůsobil špatným dojmem. Provedeme další pomocné výpočty:

$$\ln OR = 832,$$

$$\sqrt{\frac{1}{a}+\frac{1}{b}+\frac{1}{c}+\frac{1}{d}} = \sqrt{\frac{1}{7}+\frac{1}{\cdot}+\frac{1}{\cdot}+\frac{1}{\cdot}} = 439,\quad u_{0,975} = 96$$

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

$$\ln d = \ln OR - \sqrt{\frac{1}{a}+\frac{1}{b}+\frac{1}{c}+\frac{1}{d}}\,u_{1-\alpha/2} = 33 - 43 \cdot 96 = 92$$

$$\ln h = \ln OR + \sqrt{\frac{1}{a}+\frac{1}{b}+\frac{1}{c}+\frac{1}{d}}\,u_{1-\alpha/2} = 33 + 43 \cdot 96 = 692$$

Po odlogaritmování dostaneme:

$$d = 028 = 972,\quad h = 692 = 43$$

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.

Výpočet pomocí systému STATISTICA:
Dolní a horní mez intervalu spolehlivosti pro OR zjistíme pomocí STATISTIKY. Vytvoříme datový soubor o dvou proměnných DM a HM a jed-

nom případu. Do Dlouhého jména proměnné DM napíšeme vzorec pro dolní mez:
=exp(log(2,298)-sqrt(1/17+1/11+1/39+1/58)*VNormal(0,975;0;1))
a analogicky do Do Dlouhého jména proměnné HM napíšeme vzorec pro horní mez:
=exp(log(2,298)+sqrt(1/17+1/11+1/39+1/58)*VNormal(0,975;0;1))

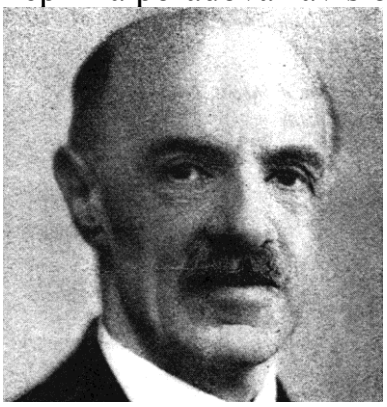
	1 DM	2 HM
1	0,972	5,431

11.14. Poznámka: Pro čtyřpolní 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.) Jestliže p-hodnota pro tento test $\leq \alpha$, pak hypotézu o nezávislosti zamítáme na hladině významnosti α.

11.15. Definice (definice Spearmanova koeficientu pořadové korelace, význam jeho hodnot)
Nechť X,Y jsou náhodné veličiny aspoň ordinálního typu. 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.

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

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.



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

11.16. Věta (věta o testování hypotézy o pořadové nezávislosti veličin X, Y)
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 $|r_S| \geq r_{S,1-\alpha}(n)$
- levostranné alternativy, když $r_S \leq - r_{S,1-2\alpha}(n)$
- pravostranné alternativy, když $r_S \geq r_{S,1-2\alpha}(n)$,

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

11.17. Věta (asymptotická varianta testu)

Pro $n > 20$ lze použít testovou statistiku $T_0 = r_S\sqrt{n-2}\,/\sqrt{1-r_S^2}$, 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.

Pro $n > 30$ lze použít testovou statistiku $r_S\sqrt{n-1}$. Platí-li H_0, pak $r_S\sqrt{n-1} \approx N(0, 1)$. Nulovou hypotézu tedy zamítáme na asymptotické hladině významnosti α ve prospěch
oboustranné alternativy, když $r_S\sqrt{n-1} \in (-\infty, u_{1-\alpha/2}) \cup (u_{1-\alpha/2}, \infty)$,
levostranné alternativy, když $r_S\sqrt{n-1} \in (-\infty, u_{1-\alpha})$,
pravostranné alternativy, když $r_S\sqrt{n-1} \in (u_{1-\alpha}, \infty)$

11.18. 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:

$$r_S = 1 - \frac{6}{7^2 - 1}\left[\quad + \quad + \quad + \quad + \quad + \quad + \quad \right] = 0{,}857.$$

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

	Spearmanovy korelace (dva le ChD vynechány párově Označ. korelace jsou významr			
Dvojice prom	Poč platné	Spearm R	t(N-2	Úroveň
X & Y	7	0,857	3,721	0,013

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.

11.19. Definice (definice Pearsonova koeficientu korelace)

Nechť (X, Y) je náhodný vektor, přičemž náhodné veličiny X, Y jsou aspoň intervalového typu. Číslo

$$R_{XY} = \left(\frac{(X - EX)(Y - EY)}{\sqrt{D(X)} \cdot \sqrt{D(Y)}} \right) = \frac{C_{XY}}{\sqrt{D(X)}\sqrt{D(Y)}} \text{ pro } D(X), D(Y) > 0$$

se nazývá Pearsonův koeficient korelace.

(Pro výpočet Pearsonova koeficentu korelace musíme znát simultánní distribuční funkci $\Phi(x,y)$ v obecném případě resp. simultánní hustotu pravděpodobnosti

φ(x,y) ve spojitém případě resp. simultánní pravděpodobnostní funkci π(x,y) v diskrétním případě.)

11.20. Věta (věta o vlastnostech koeficientu korelace)

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

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

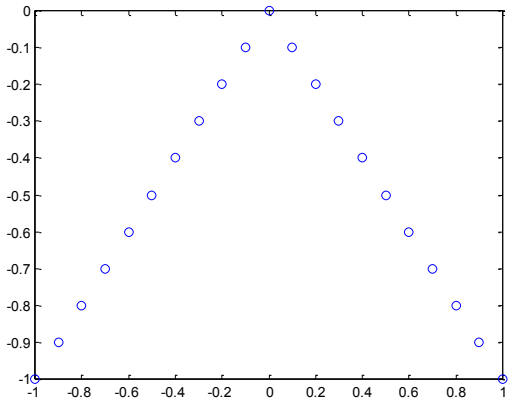
c) $R(X, X) = 1$ pro $D(X) \neq 0$, $R(X, X) = 0$ jinak

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:



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

11.21. Definice (definice výběrového koeficientu korelace)
Nechť $(X_1, Y_1), ..., (X_n, Y_n)$ 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 $R_{12} = \begin{cases} \frac{1}{n-1}\sum_{i=1}^{n} \frac{(X_i - M_1)(Y_i - M_2)}{S_1 \cdot S_2} = \frac{S_{12}}{S_1 S_2} & \text{pro } S_1 S_2 > 0 \\ 0 \text{ jinak} \end{cases}$.

11.22. Poznámka: Vlastnosti Pearsonova koeficientu korelace uvedené v 11.20. se přenášejí i na výběrový koeficient korelace.

11.23. Věta (věta o koeficientu 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:

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

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

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 $(X_1, Y_1), ..., (X_n, Y_n)$ je náhodný výběr rozsahu n z dvourozměrného normálního rozložení

$N_2\left(\begin{pmatrix}\mu_1\\ \mu_2\end{pmatrix}, \begin{pmatrix}\sigma_1^2 & \rho\sigma_1\sigma_2\\ \rho\sigma_1\sigma_2 & \sigma_2^2\end{pmatrix}\right)$.

11.24. Věta (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: $T_0 = \frac{R\sqrt{n-2}}{\sqrt{1-R^2}}$.

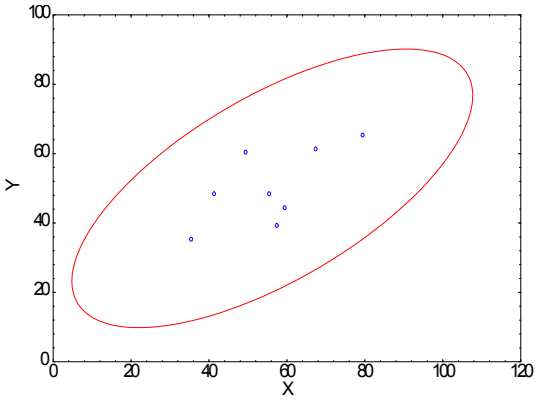
Platí-li nulová hypotéza, pak $T_0 \sim t(n-2)$.
Kritický obor pro test H_0 proti
- oboustranné alternativě: $W = (-\infty, t_{1-\alpha/2}(n-2)) \cup (t_{1-\alpha/2}(n-2), \infty)$,
- levostranné alternativě: $W = (-\infty, t_{1-\alpha}(n-2))$,
- pravostranné alternativě: $W = (t_{1-\alpha}(n-2), \infty)$.

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

11.25. 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, protože vrstevnice hustoty dvourozměrného normálního rozložení jsou elipsy.



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

Výpočet pomocí systému STATISTICA

a) Vytvoříme datový soubor o dvou proměnných X, Y a 8 případech. Dvourozměrnou normalitu dat ověříme pomocí dvourozměrného 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 (dva testy.sta)
Označ. korelace jsou významné na hlad. p < ,05000
(Celé případy vynechány u ChD)

Prom.	Prům	Sm.Od	r(X,Y	r2	t	p	N	Kons záv.:	Směr záv:	Kons záv.:	Směr záv.:
X	56,25	13,99									
X	56,25	13,99	1,000	1,000			8	0,000	1,000	0,000	1,000
X	56,25	13,99									
Y	50,00	10,92	0,666	0,444	2,191	0,070	8	20,71	0,520	13,54	0,854
Y	50,00	10,92									
X	56,25	13,99	0,666	0,444	2,191	0,070	8	13,54	0,854	20,71	0,520
Y	50,00	10,92									
Y	50,00	10,92	1,000	1,000			8	0,000	1,000	0,000	1,000

Výběrový koeficient korelace se realizoval hodnotou 0,6668, testová statistika nabyla hodnoty 2,1917, odpovídající p-hodnota pro oboustranný test je 0,0709, tedy pro jednostranný test je 0,035045. Na hladině významnosti 0,05 zamítáme hypotézu o nezávislosti veličin X, Y ve prospěch pravostranné alternativy.

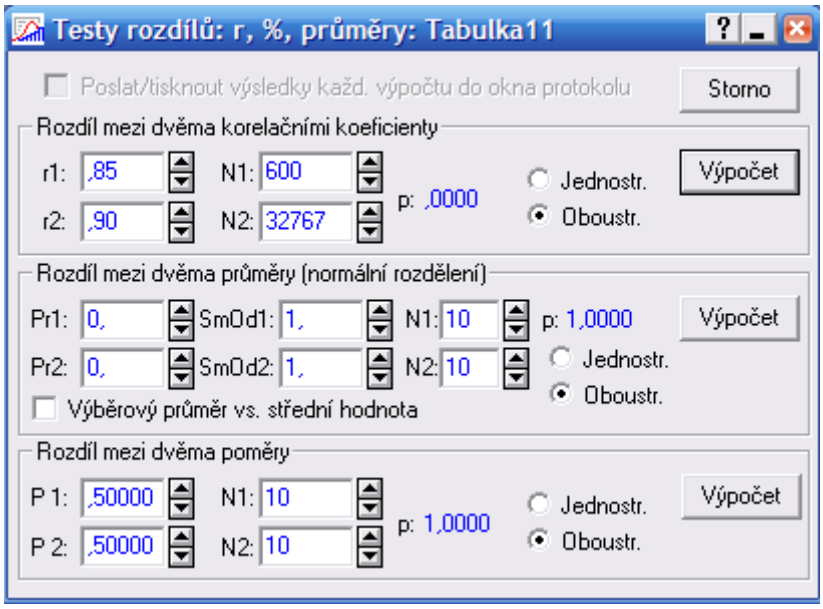
b) Můžeme využít toho, že již známe r_{12}. Statistiky – Pravděpodobnostní kalkulátor – Korelace – vyplníme n = 8, r = 0,6668, odškrtneme Dvojité, zaškrtneme Výpočet p z r – Výpočet. V okénku p se objeví hodnota 0,035455, tedy na hla-

dině významnosti 0,05 zamítáme hypotézu o nezávislosti veličin X a Y ve prospěch pravostranné alternativy.

11.26. Věta (test o 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 $U = \left(\frac{1}{2}\ln\frac{1+r}{1-r} - \frac{1}{2}\ln\frac{1+c}{1-c}\right)\sqrt{n-3}$, která má za platnosti H_0 pro $n \geq 10$ asymptoticky rozložení N(0,1), přičemž $Z = \frac{1}{2}\ln\frac{1+r}{1-r}$ je tzv. Fisherova Z-transformace. Kritický obor pro test H_0 proti oboustranné alternativě tedy je $W = (-\infty, u_{1-\alpha/2}) \cup (u_{1-\alpha/2}, \infty)$. H_0 zamítáme na asymptotické hladině významnosti α, když $U \in W$.

11.27. 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:
$Z = \frac{1}{2}\ln\frac{1+0{,}85}{1-0{,}85} = 1{,}256$, $U = (1{,}256 - \frac{1}{2}\ln\frac{1+0{,}9}{1-0{,}9})\sqrt{600-3} = -2{,}297$, $u_{0,975} = 1{,}96$, $W = (-\infty, -1{,}96) \cup (1{,}96, \infty)$. Protože $U \in W$, H_0 zamítáme na asymptotické hladině významnosti 0,05.

Výpočet pomocí systému STATISTICA (pouze přibližný):
Statistiky – Základní statistiky a tabulky – Testy rozdílů: r, %, průměry – OK – vybereme Rozdíl mezi dvěma korelačními koeficienty. Do políčka r1 napíšeme 0,85, do políčka N1 napíšeme 600, do políčka r2 napíšeme 0,9, do políčka N2 napíšeme 32767 (větší hodnotu systém neumožní) - Výpočet. Dostaneme p-hodnotu 0,0000, tedy zamítáme nulovou hypotézu na asymptotické hladině významnosti 0,05.



Upozornění: Pokud bychom chtěli pomocí systému STATISTICA provést přesnější test s využitím statistiky U, můžeme vypočítat Fisherovu Z- transformaci pomocí Pravděpodobnostního kalkulátoru – Korelace, kde zadáme realizaci výběrového koeficientu korelace, rozsah výběru. Zajímá nás Fisher z.

11.28. Věta (test o porovnání dvou koeficientů korelace)
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 $W = (-\infty, u_{1-\alpha/2}) \cup (u_{1-\alpha/2}, \infty)$. H_0 zamítáme na asymptotické hladině významnosti α, když $U \in W$.

11.29. 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:

$Z_{} = \frac{1}{2}\ln\frac{1+}{1-} = 0,7752$ $Z^{*} = \frac{1}{2}\ln\frac{1+}{1-} = 0,388$, $U = \frac{0,7752 - 0,3884}{\sqrt{\frac{1}{10-} + \frac{1}{10-}}} = 0,924$, $u_{0,975}$ = 1,96, $W = (-\infty; -1,96) \cup (1,96; \infty)$. Protože $U \in$, H_0 zamítáme na asymptotické hladině významnosti 0,05.

Výpočet pomocí systému STATISTICA:
Statistiky – Základní statistiky a tabulky – Testy rozdílů: r, %, průměry – OK – vybereme Rozdíl mezi dvěma korelačními koeficienty. Do políčka r1 napíšeme 0,65, do políčka N1 napíšeme 100, do políčka r2 napíšeme 0,37, do políčka N2 napíšeme 142 - Výpočet. Dostaneme p-hodnotu 0,0038, tedy zamítáme nulovou hypotézu na asymptotické hladině významnosti 0,05.

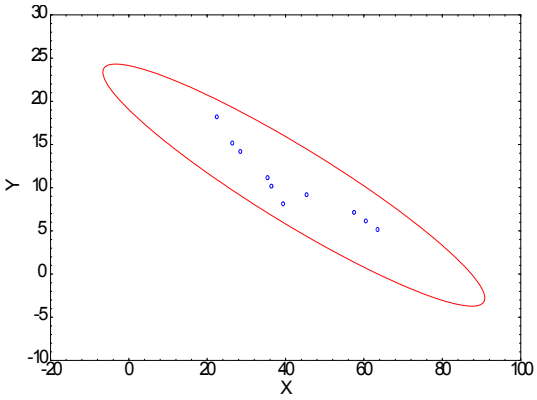
11.30. Věta (věta o asymptotickém intervalu spolehlivosti pro koeficient korelace)
Nechť dvourozměrný náhodný výběr rozsahu n pochází z dvourozměrného normálního rozložení s koeficientem korelace ρ. Meze 100(1-α)% asymptotického intervalu spolehlivosti pro ρ jsou:

$d = \tanh\left(Z - \frac{u_{1-\alpha/2}}{\sqrt{n-3}}\right)$, $h = \tanh\left(Z + \frac{u_{1-\alpha/2}}{\sqrt{n-3}}\right)$ přičemž $\tanh x = \frac{e^{x} - }{}$, $Z = \frac{1}{2}\ln\frac{1+}{1-}$.

11.31. 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 $1 \in W$, zamítáme na hladině významnosti 0,05 hypotézu o nezávislosti veličin X a Y.

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