

## Téma 13: Korelační analýza

### Úkol 1.: Testování nezávislosti ordinálních veličin

12 různých softwarových firem nabízí speciální programové vybavení pro vedení účetnictví. Jednotlivé programy byly posouzeny odbornou komisí složenou z počítačových odborníků a komisí složenou z profesionálních účetních. Úkolem bylo doporučit vhodný program na základě stanovení pořadí jednotlivých programů. Výsledky posouzení:

Produkt firmy číslo	1	2	3	4	5	6	7	8	9	10	11	12
Pořadí dle odborníků	6	7	1	8	4	2,5	9	12	10	2,5	5	11
Pořadí dle účetních	4	5	2	10	6	1	7	11	8	3	12	9

Vypočtete Spearmanův koeficient pořadové korelace a na hladině významnosti 0,05 testujte hypotézu, že hodnocení obou komisí jsou nezávislá.

#### Návod:

Testujeme vlastně nulovou hypotézu, že koeficient pořadové korelace je roven nule proti oboustranné alternativě.

Vytvoříme datový soubor o dvou proměnných X (hodnocení 1. komise), Y (hodnocení 2. komise) a 12 případech. Do proměnných X a Y zapíšeme zjištěná hodnocení.

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

		Spearmanovy korelace (Tabulka1)			
		ChD vynechány párově Označ. korelace jsou významné na hl. $p < ,05000$			
Dvojice proměnných		Počet plat.	Spearman R	t(N-2)	Úroveň p
X	& Y	12	0,714537	3,229806	0,009024

Spearmanův koeficient pořadové korelace nabývá hodnoty 0,7145, testová statistika se realizuje hodnotou 3,2298, odpovídající p-hodnota je 0,009024, tedy na asymptotické hladině významnosti 0,05 zamítáme hypotézu o pořadové nezávislosti hodnocení dvou komisí ve prospěch oboustranné alternativy.

Upozornění: Systém STATISTICA používá při testování hypotézy o pořadové nezávislosti veličin X, Y asymptotickou variantu testu bez ohledu na rozsah náhodného výběru. Pokud rozsah výběru nepřesáhne 20, měli bychom systém STATISTICA použít jen k výpočtu  $r_s$  a testování bychom měli provést pomocí tabelované kritické hodnoty. V našem případě pro  $n = 12$  a  $\alpha = 0,05$  je kritická hodnota 0,5804. Vidíme, že nulovou hypotézu zamítáme na hladině významnosti 0,05, protože  $0,7145 \geq 0,5804$ .

**Úkol k samostatnému řešení:** Bylo sledováno 10 žáků. Na základě psychologického vyšetření byli tito žáci seřazeni podle nervové labilita (čím byl žák labilnější, tím dostal vyšší pořadí  $R_i$ ). Kromě toho sledování žáci dostali pořadí  $Q_i$  na základě svých výsledků v matematice (nejlepší žák v matematice dostal pořadí 1). Výsledky jsou uvedeny v tabulce:

Pořadí $R_i$	1	2	3	4	5	6	7	8	9	10
Pořadí $Q_i$	9	3	8	5	4	2	10	1	7	6

Vypočtete Spearmanův koeficient pořadové korelace a na hladině významnosti 0,05 testujte hypotézu, že nervová labilita a výsledky v matematice jsou nezávislé.

Výsledek:  $r_s = -0,127$ ,  $H_0$  nezamítáme na hladině významnosti 0,05.

## Úkol 2.: Testování nezávislosti intervalových a poměrových veličin

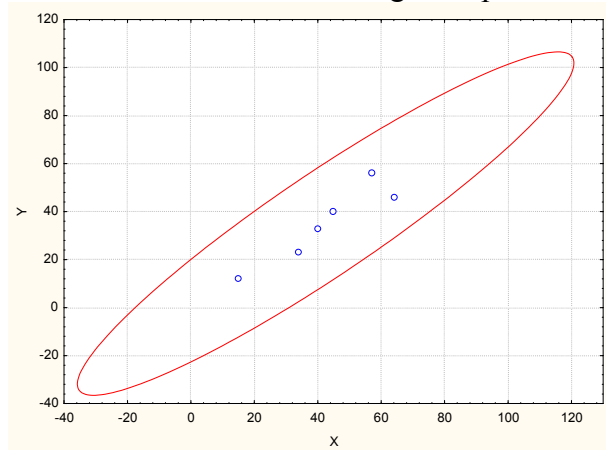
Zjišťovalo se, kolik mg kyseliny mléčné je ve 100 ml krve matek provorodiček (veličina X) a u jejich novorozenců (veličina Y) těsně po porodu. Byly získány tyto výsledky:

Číslo matky	1	2	3	4	5	6
$x_i$	40	64	34	15	57	45
$y_i$	33	46	23	12	56	40

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 měření.

**Návod:** Vytvoříme datový soubor o dvou proměnných X a Y a šesti případech. Obvyklým způsobem zobrazíme dvourozměrný tečkový diagram, s jehož pomocí posoudíme dvourozměrnou normalitu dat. Tedy:

Grafy – Bodové grafy – vypneme lineární proložení - Proměnné X, Y – OK – Detaily - Elipsa normální – OK. Ve vzniklém grafu upravíme měřítka na vodorovné a svislé ose:



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

Korelace (Tabulka3)											
Označ. korelace jsou významné na hlad. $p < ,05000$											
(Celé případy vynechány u ChD)											
Prom. X & prom. Y	Průměr	Sm.Odch.	$r(X,Y)$	$r^2$	t	p	N	Konst. záv.: Y	Směr. záv. Y	Konst. záv.: X	Směrnic záv.: X
X	42,50000	17,39828									
Y	35,00000	15,89969	0,934832	0,873912	5,265339	0,006232	6	-1,30823	0,854311	6,696994	1,022943

Ve výstupní tabulce je mj. hodnotu výběrového korelačního koeficientu  $R_{12}$  ( $r=0,9348$ ), tzn. že mezi X a Y existuje silná přímá lineární závislost), hodnota testové statistiky ( $t = 5,2653$ ) a p-hodnotu pro test hypotézy o nezávislosti ( $p=0,006232$ ),  $H_0$  tedy zamítáme na hladině významnosti 0,05. S rizikem omylu nejvýše 5% jsme tedy prokázali, že mezi oběma koncentracemi existuje závislost.

Pro testování pomocí intervalu spolehlivosti zopakujme nejprve teorii:

Nechť dvourozměrný náhodný výběr rozsahu n pochází z dvourozměrného normálního rozložení s koeficientem korelace  $\rho$ . Meze  $100(1-\alpha)\%$  asymptotického intervalu spolehlivosti pro  $\rho$  jsou:

$$d = \operatorname{tgh}\left(Z - \frac{u_{1-\alpha/2}}{\sqrt{n-3}}\right), \quad h = \operatorname{tgh}\left(Z + \frac{u_{1-\alpha/2}}{\sqrt{n-3}}\right) \text{ přičemž } \operatorname{tgh} x = \frac{e^x - e^{-x}}{e^x + e^{-x}}, \quad Z = \frac{1}{2} \ln \frac{1 + R_{12}}{1 - R_{12}}.$$

Výpočet mezí intervalu spolehlivosti: vytvoříme nový datový soubor s proměnnými DM a HM. Do Dlouhého jména proměnné DM zapíšeme příkaz

= TanH(0,5\*log((1+0,9348)/(1-0,9348))-VNormal(0,975;0;1)/sqrt(6-3))

a do Dlouhého jména proměnné HM zapíšeme příkaz

= TanH(0,5\*log((1+0,9348)/(1-0,9348))+VNormal(0,975;0;1)/sqrt(6-3))

	1	2
	DM	HM
1	0,510617	0,993014

95% interval spolehlivosti pro  $\rho$  má tedy meze 0,5106 a 0,9930, nepokrývá hodnotu 0 a tudíž hypotézu o nezávislosti veličin X, Y zamítáme na hladině významnosti 0,05.

**Poznámka:** Pokud známe výběrový koeficient korelace a rozsah výběru, můžeme test nezávislosti veličin X, Y provést pomocí Pravděpodobnostního kalkulátoru.

Statistiky – Pravděpodobnostní kalkulátor – Korelace – zadáme n a r, zaškrtneme Výpočet p z r – Výpočet.

**Úkol k samostatnému řešení:** V náhodném výběru 10 dvoučlenných domácností byl zjišťován měsíční příjem (veličina X, v tisících Kč) a vydání za potraviny (veličina Y, v tisících Kč).

$x_i$	15	21	34	35	39	42	58	64	75	90
$y_i$	3	4,5	6,5	6	7	8	9	8	9,5	10,5

Vypočtete výběrový koeficient korelace. Na hladině významnosti 0,05 testujte hypotézu o nezávislosti veličin X, Y. Sestrojte 95% asymptotický interval spolehlivosti pro  $\rho$   
Výsledek:  $r_{12} = 0,9405$ ,  $H_0$  zamítáme na hladině významnosti 0,05, s pravděpodobností aspoň 0,95 platí:  $0,7623 < \rho < 0,9862$

### Úkol 3.: Porovnání dvou korelačních koeficientů

V psychologickém výzkumu bylo vyšetřeno 426 hochů a 430 dívek. Ve skupině hochů činil výběrový koeficient korelace mezi verbální a performační složkou IQ 0,6033, ve skupině dívek činil 0,5833. Za předpokladu dvourozměrné normality dat testujte na hladině významnosti 0,05 hypotézu, že korelační koeficienty se neliší.

**Návod:** Nejprve zopakujeme teorii:

Necheť jsou dány dva nezávislé náhodné výběry o rozsazích n a  $n^*$  z dvourozměrných normálních rozložení s korelačními koeficienty  $\rho$  a  $\rho^*$ . 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  $\alpha$ , když  $U \in W$ .

### **Výpočet pomocí systému STATISTICA:**

Statistiky – Základní statistiky a tabulky – Testy rozdílů: r, %, průměry – OK – vybereme Rozdíl mezi dvěma korelačními koeficienty. Do políčka r1 napíšeme 0,6033, do políčka N1 napíšeme 426, do políčka r2 napíšeme 0,5833, do políčka N2 napíšeme 430 - Výpočet. Dostaneme p-hodnotu 0,6528, tedy nezamítáme nulovou hypotézu o shodě dvou koeficientů korelace na asymptotické hladině významnosti 0,05.