Téma 11: Analýza závislosti dvou ordinálních, intervalových a poměrových veličin

Úkol 1.: **Testování nezávislosti ordinálních veličin**
Dva lékaři hodnotili stav sedmi pacientů po témž chirurgickém zákroku tak, aby nejvyšší pořadí měl nejtěžší případ.

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

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

Návod:
Testujeme vlastně nulovou hypotézu, že Spearmanův koeficient pořadové korelace je roven nule proti oboustranné alternativě. Popis testu je v odstavci 11.3.
Načteme datový soubor dva_lekari.sta o dvou proměnných a 7 případech. Proměnná X obsahuje hodnocení 1. lékaře, proměnná Y hodnocení 2. lékaře.
Statistiky – Neparametrické statistiky – Korelace – OK – vybereme Vytvořit detailní report - Proměnné X, Y – OK – Spearmanův koef. R.

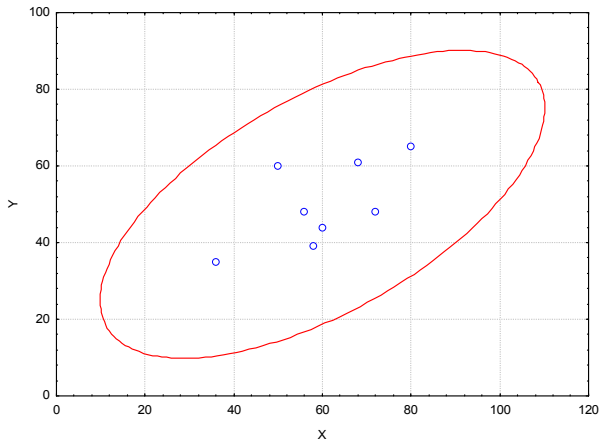
	Spearmanovy korelace (dva lekari.sta) ChD vynechány párově Označ. korelace jsou významné na hl. p <,05000			
Dvojice proměnných	Počet plat.	Spearman R	t(N-2)	Úroveň p
X & Y	7	0,857143	3,721042	0,013697

Komentář: Ve výstupní tabulce najdeme Spearmanův koeficient, hodnotu asymptotické testové statistiky a odpovídající p-hodnotu. Nulovou hypotézu zamítáme na asymptotické hladině významnosti 0,05, protože p-hodnota = 0,013697 ≤ 0,05. Podmínky pro použitá asymptotické varianty testu však nejsou splněny, protože počet pozorování je příliš malý. Korektní postup by spočíval v tom, že pomocí systému STATISTICA bychom vypočetli hodnotu r_S a tu porovnali s kritickou hodnotou $r_{S,0,95}(7) = 0,745$. Protože 0,857 ≥ 0,745, nulovou hypotézu zamítáme na hladině významnosti 0,05. Pořadová závislost mezi hodnoceními obou lékařů je vysoká, $r_S = 0,857$.

Úkol 2.: **Testování nezávislosti intervalových a poměrových veličin**
Máme k dispozici výsledky testů ze dvou předmětů zjištěné u osmi náhodně vybraných studentů určitého oboru.
Počet bodů z 1. testu: 80 50 36 58 72 60 56 68
Počet bodů z 2. testu: 65 60 35 39 48 44 48 61
Nakreslete dvourozměrný tečkový diagram, vypočtěte výběrový korelační koeficient, sestrojte 95% asymptotický interval spolehlivosti pro korelační koeficient a na hladině významnosti 0,05 testujte hypotézu o nezávislosti výsledků obou testů.

Návod:
Načteme datový soubor dva_testy.sta. Proměnná X obsahuje počet bodů z 1. testu, proměnná Y počet bodů z 2. testu. Obvyklým způsobem zobrazíme dvourozměrný tečkový diagram, s jehož pomocí posoudíme dvourozměrnou normalitu dat.

Tedy: Grafy – Bodové grafy – Proměnné X, Y Variables - OK- Detaily - Proložení vypnuto – Elipsa Normální – OK. Po vytvoření grafu upravíme měřítka na obou osách, minimum pro obě osy je 0, maximum pro osu x je 120, pro osu y je 100.



Komentář: Vidíme, že body vytvářejí elipsovitý obrazec a leží uvnitř 95% elipsy konstantní hustoty pravděpodobnosti. Předpoklad dvourozměrné normality je oprávněný. Dále vidíme, že mezi počty bodů z 1. a 2. testu bude existovat určitý stupeň přímé lineární závislosti, tzn., že u studentů, kteří měli vysoký resp. nízký počet bodů v 1. testu, lze očekávat vysoký resp. nízký počet bodů ve 2. testu.

Testování hypotézy o nezávislosti: Statistiky – Základní statistiky/tabulky – Korelační matice – OK – 2 seznamy X,Y, OK – Možnosti – Zobrazit detailní tabulku výsledků – Výpočet.

	Korelace (dva_testy.sta) Označ. korelace jsou významné na hlad. p < ,05000 (Celé případy vynechány u ChD)				
Prom. X & prom. Y	r(X,Y)	r2	t	p	N
X					
Y	0,626377	0,392348	1,968267	0,096582	8

Komentář: Ve výstupní tabulce je hodnota výběrového korelačního koeficientu R_{12} (r = 0,6264, tzn. že mezi X a Y existuje nepříliš silná přímá lineární závislost) a p-hodnotu pro test hypotézy o nezávislosti (p = 0,097, H_0 tedy nelze zamítnout na hladině významnosti 0,05.

Pro testování pomocí intervalu spolehlivosti použijeme Větu 11.4.7.2. Výběrový koeficient korelace R_{12} se podle vzorce (11.22) transformuje na veličinu $Z = \frac{1}{2}\ln\frac{1+R_{12}}{1-R_{12}}$, která má i při malém rozsahu výběru přibližně normální rozložení se střední hodnotou $E(Z) = \frac{1}{2}\ln\frac{1+\rho}{1-\rho} + \frac{\rho}{2(n-1)}$ (2. sčítanec lze při větším n zanedbat) a rozptylem $D(Z) = \frac{1}{n-3}$. Standardizací veličiny Z dostaneme veličinu $U = \frac{Z - E(Z)}{\sqrt{D(Z)}}$, která má asymptoticky rozložení N(0,1). Tudíž 100(1-α)% asymptotický interval spolehlivosti pro $\frac{1}{2}\ln\frac{1+\rho}{1-\rho}$ bude mít meze

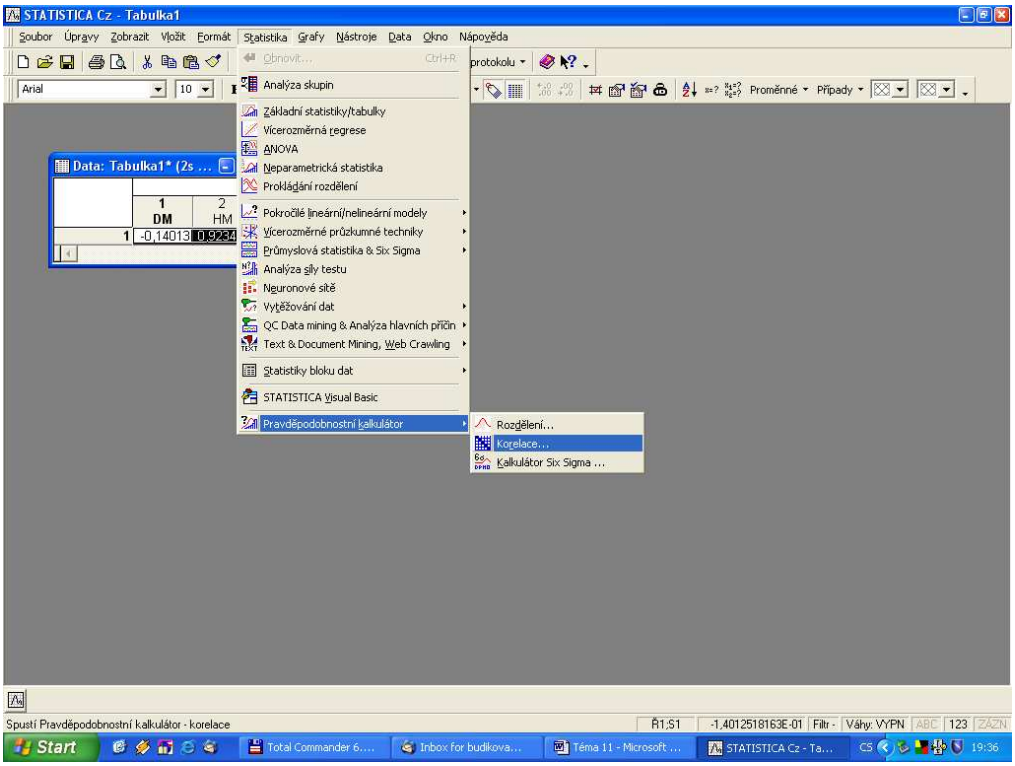
$Z \pm \frac{u_{1-\alpha/2}}{\sqrt{n-3}}$. Jelikož Z = arctgh R_{12}, dostáváme R_{12} = tgh Z a meze intervalu spolehlivosti pro ρ můžeme psát ve tvaru $\operatorname{tgh}\left(Z \pm \frac{u_{1-\alpha/2}}{\sqrt{n-3}} \right)$ - viz vzorec (11.25.)

Ve STATISTICE vypočteme meze 100(1-α)% asymptotického intervalu spolehlivosti pro koeficient korelace ρ tak, že otevřeme nový datový soubor se dvěma proměnnými (pojmenujeme je DM a HM) a jedním případem.
Do Long name proměnné DM zapíšeme příkaz
= TanH(0,5*log((1+0,6264)/(1-0,6264))-VNormal(0,975;0;1)/sqrt(8-3))
a do Long name proměnné HM zapíšeme příkaz
= TanH(0,5*log((1+0,6264)/(1-0,6264))+VNormal(0,975;0;1)/sqrt(8-3))

	1 DM	2 HM
1	-0,14013	0,923454

95% asymptotický interval spolehlivosti pro ρ má tedy meze –0,14013 a 0,923454, pokrývá hodnotu 0 a tudíž hypotézu o nezávislosti veličin X, Y nelze zamítnout na asymptotické hladině významnosti 0,05.

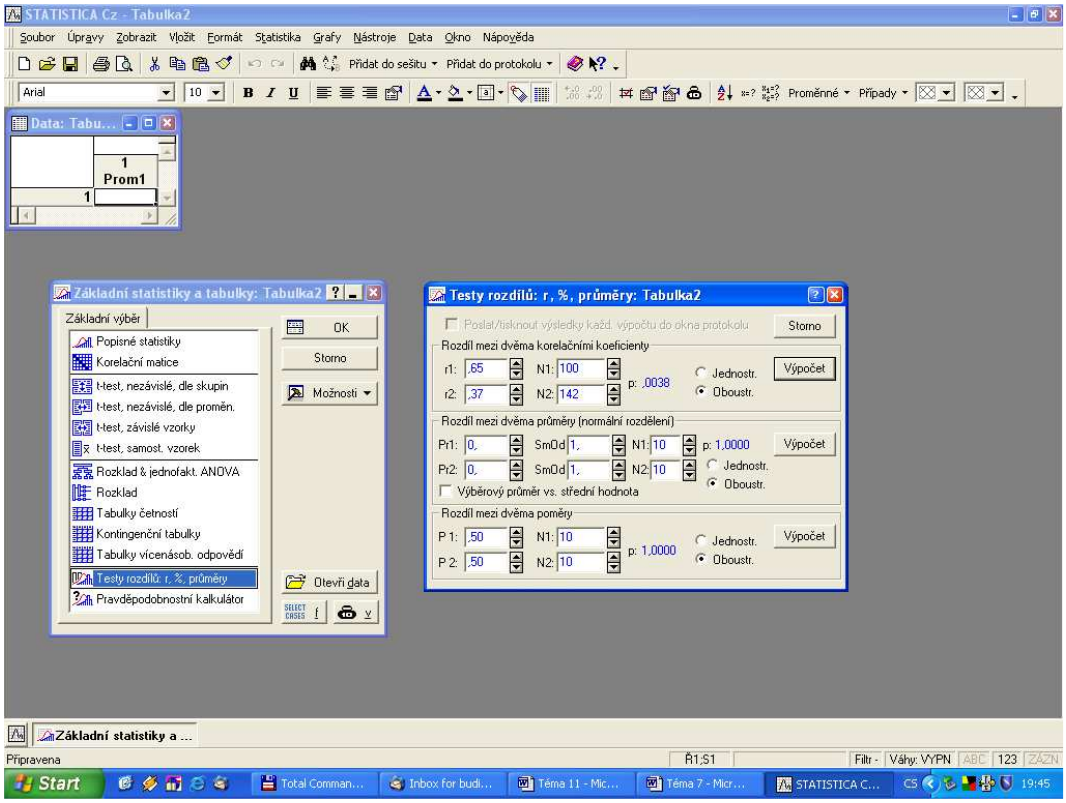
Poznámka: Pokud známe výběrový koeficient korelace a rozsah výběru, můžeme test nezávislosti veličin X, Y provést pomocí Pravděpodobnostního kalkulátoru.
Statistiky – Pravděpodobnostní kalkulátor – Korelace – zadáme n a r, zaškrtneme Výpočet p z r – Výpočet.



Zadáme N = 8, r = 0,6264, OK. Dostaneme p = 0,096566 a navíc ještě Fisher. Z = 0,735469

Úkol 6.: **Porovnání dvou korelačních koeficientů**
Lékařský výzkum se zabýval sledováním koncentrací látek A a B v moči pacientů trpících určitou ledvinovou chorobou. U 100 zdravých osob činil výběrový koeficient korelace mezi koncentracemi obou látek 0,65 a u 142 osob trpících zmíněnou chorobou byl 0,37. Za předpokladu dvourozměrné normality dat testujte na hladině významnosti 0,05 hypotézu, že korelační koeficienty se neliší.

Návod:
Ve STATISTICE je test shody dvou koeficientů korelace (viz Věta 11.4.6.1.) implementován. 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.



Příklady k samostatnému řešení

Příklad 1.: U určitého výrobku hodnotil expert dvě vlastnosti na desetibodové stupnici tak, že nula je nejhorší a desítka nejlepší hodnocení. Máte k dispozici výsledky hodnocení 11 náhodně vybraných výrobků:

1. vlastnost	3,1	2,8	4,4	5,8	5,1	4,3	4,7	2,9	5,3	5,4	5,9
2. vlastnost	7,2	6,5	6,9	8,4	76	4,4	3,8	7,1	4,3	4,7	8,9

Vypočtěte Spearmanův koeficient pořadové korelace a na hladině významnosti 0,05 testujte hypotézu, že hodnocení obou vlastností jsou pořadově nezávislá.

Řešení:

Načteme datový soubor hodnoceni_experta.sta. Proměnná X obsahuje bodové hodnocení první vlastnosti, proměnná Y bodové hodnocení druhé vlastnosti. Vypočteme Spearmanův koeficient pořadové korelace.

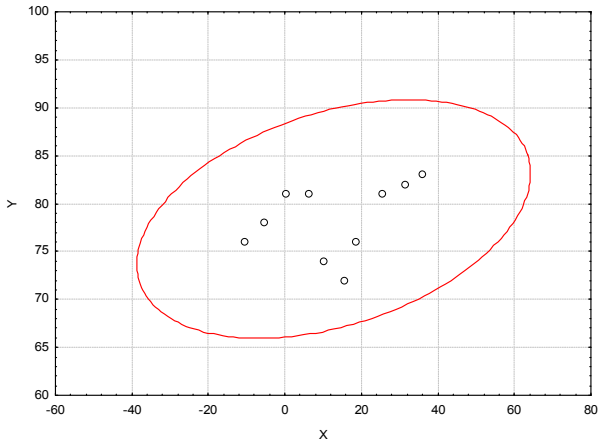
	Spearmanovy korelace (hodnoceni_experta.sta) ChD vynechány párově Označ. korelace jsou významné na hl. p <,05000			
Dvojice proměnných	Počet plat.	Spearman R	t(N-2)	Úroveň p
X & Y	11	0,281818	0,881170	0,401145

Vidíme, že r_S nabývá hodnoty 0,282. Vztah mezi bodovými hodnoceními je slabý. Nyní na hladině významnosti 0,05 testujeme nulovou hypotézu, že Spearmanův koeficient pořadové korelace je roven nule proti oboustranné alternativě. Asymptotická testová statistika se realizuje hodnotou 0,88117, odpovídající p-hodnota je 0,4011, tedy na asympotické hladině významnosti 0,05 nelze zamítnout hypotézu, že bodová hodnocení jsou pořadově nezávislá.

Příklad 2.: Při průzkumu příčin dopravních nehod bylo provedeno měření diastolického tlaku 10 skupin řidičů autobusů při různých teplotách vnějšího ovzduší. Data znázorněte graficky, posuďte jejich dvourozměrnou normalitu, vypočtěte realizaci výběrového koeficientu korelace a na hladině významnosti 0,05 testujte hypotézu, že teplota ovzduší neovlivňuje krevní tlak řidičů proti alternativě, že že mezi teplotou a tlakem existuje kladná korelace.

Teplota ovzduší (ve ° C):	-10,5	-5,4	0,2	6,4	10,2	15,6	18,5	25, 5	28,9	31,5	35,8
průměrný tlak (v mm Hg):	76	78	81	81	74	72	76	81	82	83	84

Řešení:
Načteme datový soubor ridici_autobusu.sta. Proměnná X obsahuje teploty, proměnná Y tlaky. Vytvoříme dvourozměrný tečkový diagram s 95% elipsou konstantní hustoty pravděpodobnosti:



Komentář: Vzhled diagramu svědčí o dvourozměrné normalitě dat.

Číselná realizace výběrového koeficientu korelace: $r_{12} = 0,3823$ svědčí o existenci poměrně slabé přímé lineární závislosti mezi vnější teplotou a diastolickým krevním tlakem řidičů autobusů – s rostoucí teplotou poněkud roste krevní tlak.
Na hladině významnosti 0,05 testujeme hypotézu $H_0 : \rho = 0$ proti pravostranné alternativě $H_1 : \rho > 0$. Pomocí Pravděpodobnostního kalkultátoru zjistíme p-hodnotu pro tuto jednostrannou alternativu: p = 0,1378. Na hladině významnosti 0,05 tedy nezamítáme hypotézu, že vztah mezi teplotou a tlakem je pouze náhodný.