# Cvičení 11: Jednoduchá 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 |

Vypočtěte 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.

Načteme datový soubor vedeni_ucetnictvi.sta o dvou proměnných X (hodnocení 1. komise), Y (hodnocení 2. komise) a 12 případech.

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

| | Spearmanovy korelace (veden<br>ChD vynechány párově<br>Označ. korelace jsou význam | | | |
|---|---|---|---|---|
| Dvojice prom | Poč plat | Spearr R | t(N-2 | p-hod |
| X & Y | 1. | 0,714 | 3,229 | 0,009 |

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 STATISTCA 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 $0{,}7145 \geq 0{,}5804$.

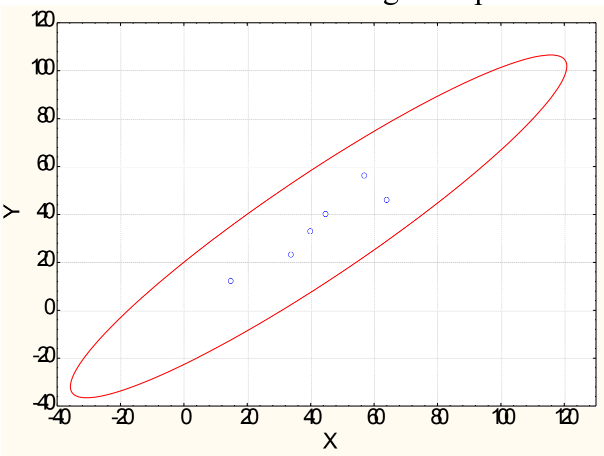
**Úkol 2.: Testování nezávislosti intervalových a poměrových veličin – oboustranná alternativa**

Zjišťovalo se, kolik mg kyseliny mléčné je ve 100 ml krve matek prvorodiček (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**: Načteme datový soubor kyselna_mlecna.sta 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 – 2 seznamy proměn. – X, Y – OK – na záložce Možnosti vybereme Zobrazit detailní tabulku výsledků – Výpočet.

Korelace (kyselina_mlecna.sta)
Označ. korelace jsou významné na hlad. p < ,05000
(Celé případy vynechány u ChD)

| Prom. prom. | Prům | Sm.Od | r(X,Y | r2 | t | p | N | Kons záv.: | Směr záv: | Kons záv.: | Směr záv.: |
|---|---|---|---|---|---|---|---|---|---|---|---|
| X | 42,50 | 17,39 | | | | | | | | | |
| Y | 35,00 | 15,89 | 0,934 | 0,873 | 5,265 | 0,006 | 6 | -1,30 | 0,854 | 6,696 | 1,022 |

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 ρ. Meze 100(1-α)% asymptotického intervalu spolehlivosti pro ρ jsou:

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

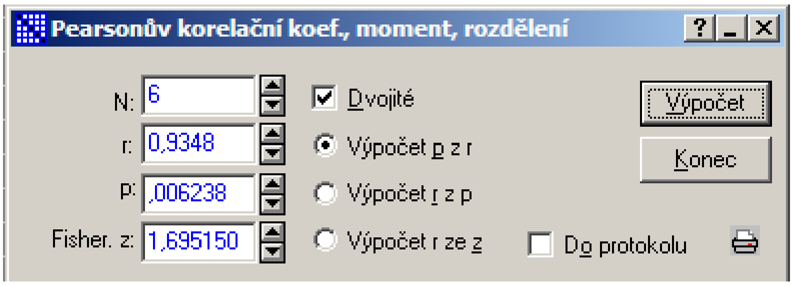
Výpočet mezí intervalu spolehlivosti: vytvoříme nový datový soubor s proměnnými DM a HM. Do Dlouhého jména proměnné DM zapíšeme příkaz
= TanH(0,5*log((1+0,9348)/(1-0,9348))-VNormal(0,975;0;1)/sqrt(6-3))
a do Dlouhého jména proměnné HM zapíšeme příkaz
= TanH(0,5*log((1+0,9348)/(1-0,9348))+VNormal(0,975;0;1)/sqrt(6-3))

| | 1<br>DM | 2<br>HM |
|---|---|---|
| 1 | 0,510 | 0,993 |

95% interval spolehlivosti pro ρ 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.

**Využití modulu Analýza síly testu**
Statistiky – Analýza síly testu – Odhad intervalu **-** Jedna korelace, t-test – OK – Pozorované R: 0,9348, N: 6, Spolehlivost: 0,95 – Výpočetní algoritmus: zaškrtneme Fisherovo Z (původní) – Vypočítat.
Zjistíme, že Dolní mez = 0,5106, Horní mez = 0,993.

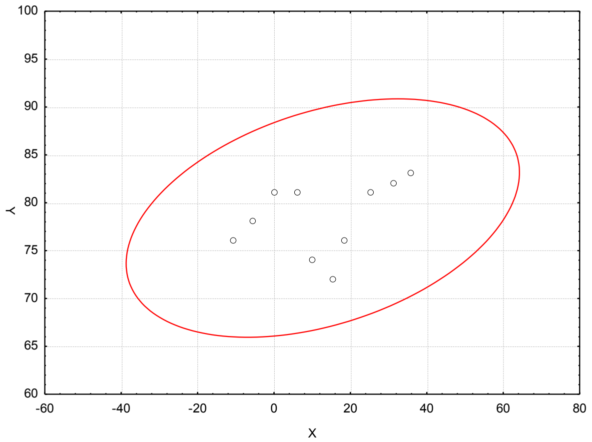
**Poznámka:** Pokud známe výběrový koeficient korelace a rozsah výběru, můžeme test nezávislosti veličin X, Y provést pomocí Pravděpodobnostního kalkulátoru**.**
Statistiky – Pravděpodobnostní kalkulátor – Korelace – zadáme n a r, zaškrtneme Výpočet p z r – Výpočet.




**Úkol 3.: Testování nezávislosti intervalových a poměrových veličin – jednostranná alternativa**
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 |

**Návod:**
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.

**Úkol 4.**: **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:
Nechť 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 $W = (-\infty, -u_{1-\alpha/2}\rangle \cup \langle u_{1-\alpha/2}, \infty)$. $H_0$ 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.

**Úkoly k samostatnému řešení**

**Příklad 1.:** Bylo sledováno 10 žáků. Na základě psychologického vyšetření byli tito žáci seřazeni podle nervové lability (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 |

Vypočtěte 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.

**Příklad 2.:** 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 |

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

**Příklad 3.:** 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.
**Výsledek:** $r_S = 0{,}282$, $H_0$ nezamítáme na hladině významnosti 0,05.

**Příklad 4.:** V následující tabulce jsou uvedeny číselné realizace a absolutní četnosti náhodného výběru $(X_1, Y_1), (X_1, Y_2), \ldots, (X_{62}, Y_{62})$ z dvourozměrného rozložení:

| x | y | | | | | | |
|---|---|---|---|---|---|---|---|
| | 1 | 3 | 5 | 7 | 9 | 11 | 13 |
| 15 | 0 | 0 | 0 | 0 | 1 | 2 | 1 |
| 25 | 0 | 0 | 0 | 5 | 4 | 2 | 0 |
| 35 | 0 | 0 | 5 | 8 | 2 | 0 | 0 |
| 45 | 0 | 5 | 6 | 4 | 0 | 0 | 0 |
| 55 | 3 | 5 | 3 | 0 | 0 | 0 | 0 |
| 65 | 4 | 2 | 0 | 0 | 0 | 0 | 0 |

Podle vzhledu dvourozměrného tečkového diagramu orientačně posuďte dvourozměrnou normalitu dat. Vypočtěte výběrový koeficient korelace a interpretujte ho. Na hladině významnosti 0,05 testujte hypotézu o nezávislosti veličin X a Y.
**Výsledek:**

Protože tečky v dvourozměrném tečkovém diagramu vytvářejí elipsovitý obrazec, lze připustit dvourozměrnou normalitu. Výběrový koeficient korelace nabývá hodnoty –0,8699, což znamená, že mezi veličinami X a Y existuje dosti silná nepřímá lineární závislost. Testová statistika se realizuje hodnotou -13,6613, odpovídající p-hodnota je velmi blízká 0, nulovou hypotézu zamítáme na hladině významnosti 0,05.