# Cvičení 10: Hodnocení kontingenčních tabulek

**Úkol 1.: Testování hypotézy o nezávislosti, měření síly závislosti**
V roce 1950 zkoumali Yule a Kendall barvu očí a vlasů u 6800 mužů.

| Barva očí | Barva vlasů | | | |
|---|---|---|---|---|
| | světlá | kaštanová | černá | rezavá |
| modrá | 1768 | 807 | 180 | 47 |
| šedá nebo zelená | 946 | 1387 | 746 | 53 |
| hnědá | 115 | 438 | 288 | 16 |

Na asymptotické hladině významnosti 0,05 testujte hypotézu o nezávislosti barvy očí a barvy vlasů. Vypočtěte Cramérův koeficient. Simultánní četnosti znázorněte graficky.

**Návod:**
Testujeme hypotézu $H_0$: X, Y jsou stochasticky nezávislé náhodné veličiny proti
$H_1$: X, Y nejsou stochasticky nezávislé náhodné veličiny. Testová statistika má tvar:

$$K = \sum_{j=1}^{r}\sum_{k=1}^{s} \frac{\left(n_{jk} - \frac{n_{j.}n_{.k}}{n}\right)^2}{\frac{n_{j.}n_{.k}}{n}}$$. Platí-li $H_0$, pak K se asymptoticky řídí rozložením $\chi^2((r-1)(s-1))$,

kde r, s jsou počty variant jednotlivých proměnných.
Hypotézu o nezávislosti veličin X, Y tedy zamítáme na asymptotické hladině významnosti α, když $K \geq \chi^2_{1-\alpha}((r-1)(s-1))$.
V našem případě zjistíme, že K = 1088,15, r = 3 , s = 4, $\chi^2_{1-\alpha}((r-1)(s-1) = \chi^2_{0,95}(6) =$ 12,592 a protože hodnota testové statistiky K = 1088,15 ≥ 12,592, zamítáme nulovou hypotézu na asymptotické hladině významnosti 0,05.

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.
Otevřeme datový soubor oci_vlasy.sta o 12 případech a třech proměnných (OCI, VLASY, CETNOST).
Před provedním testu je zapotřebí ověřit podmínky dobré aproximace:
Statistiky – Základní statistiky/tabulky – Kontingenční tabulky - Specif. tabulky – List 1 OCI, List 2 VLASY, OK, Váhy - CETNOST, Stav zapnuto, OK – na záložce Možnosti zaškrtneme Očekávané četnosti – Výpočet.

| Souhrnná tab.: Očekávané četnosti (oci_vlasy.sta)<br>Četnost označených buněk > 10<br>Pearsonův chí-kv. : 1088,15, sv=6, p=0,00000 | | | | | |
|---|---|---|---|---|---|
| OCI | VLASY světlá | VLASY kaštanová | VLASY černá | VLASY rezavá | Řádk. součty |
| modrá | 1167,259 | 1085,976 | 500,902 | 47,8622 | 2802,000 |
| šedá nebo zelená | 1304,731 | 1213,875 | 559,895 | 53,4990 | 3132,000 |
| hnědá | 357,010 | 332,149 | 153,202 | 14,6388 | 857,000 |
| Vš.skup. | 2829,000 | 2632,000 | 1214,000 | 116,0000 | 6791,000 |

Podmínky dobré aproximace jsou splněny. Všechny teoretické četnosti jsou větší než 5. Nyní budeme testovat hypotézu o nezávislosti proměnných OCI, VLASY.

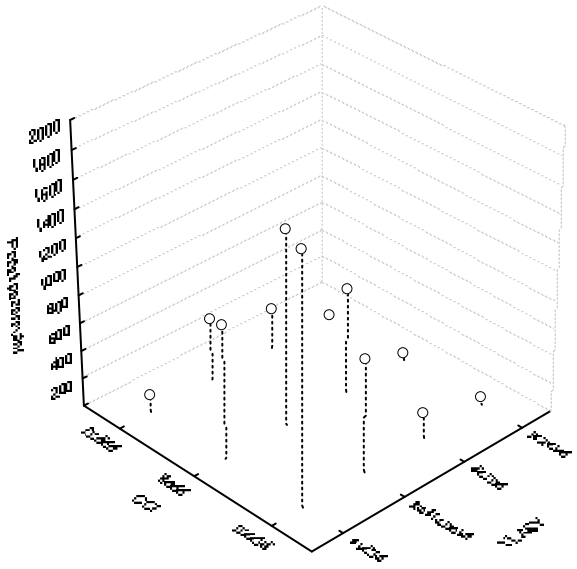
Návrat do Výsledky; kontingenční tabulky – na záložce Detaily zaškrtneme Pearsonův & M-L Chi - kvadrát, Phi & Cramerovo V – Detailní výsledky – Detailní 2 rozm. tabulky.

| Statist. | Chí-kvadr. | sv | p |
|---|---|---|---|
| Pearsonův chí-kv. | 1088,149 | df=6 | p=0,0000 |
| M-V chí-kvadr. | 1155,669 | df=6 | p=0,0000 |
| Fí | ,4002923 | | |
| Kontingenční koeficient | ,3716246 | | |
| Cramér. V | ,2830494 | | |

Ve výstupní tabulce najdeme mj. hodnotu testové stastistiky (Pearsonův chí-kv = 1088,149) s počtem stupňů volnosti (sv = 6) a odpovídající p-hodnotou (p = 0,0000), dále Cramérův koeficient (V = 0,283). Protože p-hodnota je mnohem menší než 0,05, nulovou hypotézu o nezávislosti barvy očí a barvy vlasů zamítáme na asymptotické hladině významnosti 0,05. Cramérův koeficient svědčí o slabé závislosti barvy očí a vlasů.

Pro grafické znázornění četností se vrátíme do Výsledky; kontingenční tabulky – Detailní výsledky – 3D histogramy. Po vytvoření grafu 2 krát poklepeme levým tlačítkem myši na pozadí grafu:
Rozvržení grafu – Typ Šipky – OK. Graf lze natáčet pomocí volby Zorný bod.

**Dvourozměrné rozdělení:**
**OCI x VLASY**



**Úkol 2.: Fisherův faktoriálový test**
100 náhodně vybraných mužů a žen bylo dotázáno, zda dávají přednost nealkoholickému nápoji A či B. Údaje jsou uvedeny ve čtyřpolní kontingenční tabulce.

| preferovaný nápoj | pohlaví | |
|---|---|---|
| | muž | žena |
| A | 20 | 30 |
| B | 30 | 20 |

Na hladině významnosti 0,05 testujte pomocí Fisherova faktoriálového testu hypotézu, že preferovaný typ nápoje nezáleží na pohlaví respondenta.

**Návod:** Vytvoříme nový datový soubor o třech proměnných NAPOJ, POHLAVI, CETNOST a čtyřech případech. Do proměnné NAPOJ napíšeme dvakrát pod sebe 1 (nápoj A) a dvakrát pod sebe 2 (nápoj B). Do proměnné POHLAVI napíšeme jedničku (1 – muž) a dvojku (2 – žena) a znovu jedničku a dvojku. D proměnné CETNOST napíšeme uvedené četnosti. Statistiky – Základní statistiky/tabulky – Kontingenční tabulky - Specif. tabulky – List 1 NAPOJ, List 2 POHLAVI, OK, Váhy - CETNOST, Stav zapnuto, OK – na záložce Možnosti zaškrtneme Fisher exakt, Yates, McNemar (2x2) – Detailní výsledky – Detailní 2-rozm. tabulky.

| Statist. | Statist. : POHLAVI(2) x NAPOJ(2) (kap11_2) | | |
|---|---|---|---|
| | Chí-kvadr. | sv | p |
| Pearsonův chí-kv. | 4,000000 | df=1 | p=,04550 |
| M-V chí-kvadr. | 4,027103 | df=1 | p=,04478 |
| Yatesův chí-kv. | 3,240000 | df=1 | p=,07186 |
| Fisherův přesný, 1-str. | | | p=,03567 |
| 2-stranný | | | p=,07134 |
| McNemarův chí-kv. (A/D) | ,0250000 | df=1 | p=,87437 |
| (B/C) | ,0166667 | df=1 | p=,89728 |

Ve výstupní tabulce je mimo jiné uvedena p-hodnota pro oboustranný a jednostranný test. V našem případě se jedná o oboustranný test (nevíme, zda muži více preferují nápoj A či nápoj B než ženy), zajímáme se tedy o Fisherův přesný, 2-str. Ta je 0,07134. Protože p-hodnota je větší než 0,05, nezamítáme na hladině významnosti 0,05 hypotézu, že preferovaný typ nápoje nezáleží na pohlaví respondenta.

**Úkol 3.: Podíl šancí**

Pro údaje z úkolu 2 vypočtěte podíl šancí a sestrojte 95% asymptotický interval spolehlivosti pro podíl šancí. Pomocí tohoto intervalu spolehlivosti testujte na asymptotické hladině významnosti 0,05 hypotézu, že preferovaný typ nápoje nezáleží na pohlaví respondenta.

**Návod:** Nejprve zopakujme teorii:

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 $OR = \frac{ad}{bc}$. Považujeme ho za odhad skutečného podílu šancí $o\rho$. Pomocí 100(1-α)% asymptotického intervalu spolehlivosti pro logaritmus skutečného podílu šancí $\ln o\rho$ lze na asymptotické hladině významnosti α testovat hypotézu o nezávislosti nominálních veličin X a Y. Asymptotický 100(1-α)% interval spolehlivosti pro přirozený logaritmus skutečného podílu šancí má meze:

$\ln OR \pm \sqrt{\frac{1}{a}+\frac{1}{b}+\frac{1}{c}+\frac{1}{d}}u_{1-\alpha/2}$. Jestliže interval spolehlivosti nezahrne 0, pak hypotézu o nezávislosti zamítneme na asymptotické hladině významnosti α.

V našem případě podíl šancí vypočteme ručně. $OR = \frac{ac}{bd} = \frac{20 \cdot 20}{30 \cdot 30} = \frac{4}{9} = 0,\overline{4}$. Dolní a horní mez intervalu spolehlivosti pro OR zjistíme pomocí STATISTIKY. Vytvoříme datový soubor o dvou proměnných DM a HM a dvou případech. Do Dlouhého jména proměnné DM napíšeme vzorec pro dolní mez:

=log(4/9)-sqrt(1/20+1/30+1/30+1/20)*VNormal(0,975;0;1)

a analogicky do Do Dlouhého jména proměnné HM napíšeme vzorec pro horní mez:
=log(4/9)+sqrt(1/20+1/30+1/30+1/20)*VNormal(0,975;0;1)

| | 1<br>DM | 2<br>HM |
|---|---|---|
| 1 | -1,61108 | -0,01078 |

Výsledek: -1,61108 < ln oρ < -0,01078 s pravděpodobností přibližně 0,95. Protože tento interval spolehlivosti neobsahuje 0**,** na asymptotické hladině významnosti 0,05 zamítáme hypotézu, že preferovaný typ nápoje nezáleží na pohlaví respondenta.
Tento výsledek je v rozporu s výsledkem, ke kterému dospěl Fisherův přesný test. Je to způsobeno tím, že test pomocí asymptotického intervalu spolehlivosti je pouze přibližný.

**Úkol 4.**: **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

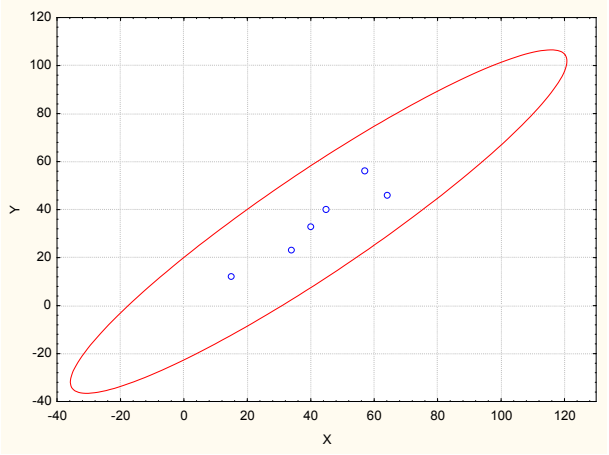
| | Spearmanovy korelace (vedeni_ucetnictvi.sta)<br>ChD vynechány párově<br>Označ. korelace jsou významné na hl. p <,05000 | | | |
|---|---|---|---|---|
| Dvojice proměnných | Počet plat. | Spearman R | t(N-2) | p-hodn. |
| X & Y | 12 | 0,714537 | 3,229806 | 0,009024 |

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 STATISTCA použít jen k výpočtu $r_S$ a testování bychom měli provést pomocí tabelované kritické hodnoty. V našem případě pro n = 12 a α = 0,05 je kritická hodnota 0,5804. Vidíme, že nulovou hypotézu zamítáme na hladině významnosti 0,05, protože 0,7145 ≥ 0,5804.

**Úkol 5.**: **Testování nezávislosti intervalových a poměrových veličin**
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 kyselina_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.

| Prom. X & prom. Y | Korelace (kyselina_mlecna.sta)<br>Označ. korelace jsou významné na hlad. p < ,05000<br>(Celé případy vynechány u ChD)<br>Průměr | Sm.Odch. | r(X,Y) | r2 | 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.

**Úkoly k samostatnému řešení:**
**Příklad 1.:** 18 mužů onemocnělo určitou chorobou. Někteří z nich se léčili, jiní ne. Někteří se uzdravili, jiní zemřeli. Údaje jsou uvedeny ve čtyřpolní kontingenční tabulce.

| přežití | léčení | |
|---|---|---|
| | ano | ne |
| ano | 5 | 3 |
| ne | 6 | 4 |

Vypočtěte a interpretujte podíl šancí. Pomocí intervalu spolehlivosti pro podíl šancí testujte na asymptotické hladině významnosti 0,05 hypotézu, že přežití nezávisí na léčení proti tvrzení, že léčení zvyšuje šance na přežití.

**Výsledek:** $OR = 1,\bar{1}$, nulovou hypotézu nezamítáme asymptotické hladině významnosti 0,05, protože levostranný 95% asymptotický interval spolehlivosti pro logaritmus podílu šancí je $(-1,80498; \infty)$.

**Příklad 2.:** 200 respondentů, z nichž bylo 73 žen, hodnotilo úroveň jistého časopisu. 34 žen ji hodnotilo kladně, stejně jako 47 mužů. Ostatní respondenti se o úrovni časopisu vyjádřili záporně. Na hladině významnosti 0,05 testujte pomocí Fisherova přesného testu, že hodnocení úrovně časopisu nezávisí na pohlaví respondenta. Vypočtěte Cramérův koeficient.

**Výsledek:** Sestavíme čtyřpolní kontingenční tabulku simultánních absolutních četností:

| hodnocení časopisu | pohlaví respondenta | | $n_{j.}$ |
|---|---|---|---|
| | muž | žena | |
| kladné | 47 | 34 | 81 |
| záporné | 80 | 39 | 119 |
| $n_{.k}$ | 127 | 73 | 200 |

Kladné hodnocení časopisu pozorujeme u 37% mužů a u 46,6 % žen.
Další výsledky máme v tabulce:

| | Statist. : hodnoceni(2) x pohlavi(2) (Tabulka13 | | |
|---|---|---|---|
| Statist. | Chí-kvadr. | sv | p |
| Pearsonův chí-kv. | 1,760835 | df=1 | p=,18452 |
| M-V chí-kvadr. | 1,752654 | df=1 | p=,18555 |
| Yatesův chí-kv. | 1,386184 | df=1 | p=,23905 |
| Fisherův přesný, 1-str. | | | p=,11967 |
| 2-stranný | | | p=,23131 |
| McNemarův chí-kv. (A/D) | 17,76316 | df=1 | p=,00003 |
| (B/C) | ,5697674 | df=1 | p=,45035 |
| Fí pro tabulky 2 x 2 | ,0938306 | | |
| Tetrachorická korelace | ,1507792 | | |
| Kontingenční koeficient | ,0934202 | | |

Fisherův přesný test poskytl p-hodnotu 0,23131, tedy na hladině významnosti 0,05 nezamítáme hypotézu o nezávislosti hodnocení úrovně časopisu na pohlaví respondenta. Cramérův koeficient je 0,0938, což svědčí o zanedbatelné závislosti mezi sledovanými veličinami.

**Příklad 3.:** Zajímá nás, zda má lokalita v ČR vliv na objem exportu do sousedních zemí. Sledujeme lokality: Ostrava, Brno, Plzeň, Praha a země: Slovensko, Rakousko, Německo, Polsko, USA). Máme k dispozici tato data:

| Odkud: | Kam: | | | | |
|---|---|---|---|---|---|
| | Slovensko | Rakousko | Německo | Polsko | USA |
| Ostrava | 350 | 216 | 189 | 626 | 46 |
| Brno | 387 | 489 | 274 | 126 | 115 |
| Plzeň | 52 | 83 | 264 | 132 | 51 |
| Praha | 484 | 594 | 737 | 447 | 141 |

Na asymptotické hladině významnosti 0,05 testujte hypotézu, že lokalita a země jsou nezávislé náhodné veličiny. (Data jsou uložena v souboru export.sta). Vypočtěte též Cramérův koeficient a interpretujte ho.
**Výsledek:** Podmínky dobré aproximace jsou splněny. Testová statistika K nabývá hodnoty 821,59, odpovídající p-hodnota je velmi blízká nule, tedy na asymptotické hladině významnosti 0,05 považujeme za prokázanou závislost objemu exportu na lokalitě v České republice.
Cramérův koeficient nabývá hodnoty 0,223, tedy mezi sledovanými proměnnými existuje slabá závislost.

**Příklad 4.**: 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 5.:** 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.
**Výsledek:** $r_{12} = 0{,}9405$, $H_0$ zamítáme na hladině významnosti 0,05