Korelačná analýza v systéme STATISTICA

RNDr. Marie Budíková, Dr.
Bc. Ester Železňáková

Přírodovědecká fakulta Masarykovy univerzity
Ústav matematiky a statistiky

Obsah

Úvod

Táto bakalárska práca vznikla s cieľom vytvoriť interaktívnu študijnú oporu k predmetom Výpočetná štatistika a Aplikovaná štatistika na uľahčenie riešenia úloh spojených s testami nezávislosti medzi dvoma náhodnými veličinami, dostupná taktiež v elektronickej podobe s videotutorálmi a interaktívnymi tutoriálmi na webovej adrese https://is.muni.cz/do/sci/UMS/el/korelacni_analyza_statistica/index.html.

Závislosťami dvoch náhodných veličín sa zaoberá dvojrozmerná štatistika v podobe jednoduchej korelačnej analýzy. Celej tejto problematike sa venujeme v práci, ktorá je rozdelená do štyroch kapitol, z ktorých úvodná kapitola hovorí najskôr o základných typoch závislosti medzi náhodnými veličinami. V druhej kapitole prechádzame k pojmu korelácia a zoznamujeme sa s rôznymi mierami tejto závislosti. Od tohto teoretického základu sa odráža tretia kapitola zameraná na testy nezávislosti náhodných veličín. Posledná štvrtá kapitola tvorí praktickú časť práce, kde aplikujeme jednotlivé testy nezávislosti na konkrétnych príkladoch. Každé riešenie pozostáva z ručného výpočtu a výpočtu za pomoci systému STATISTICA. V prílohách uvádzame tri zdrojové kódy nami vytvorených makier. Tieto makrá boli vytvorené a použité pre testy, ktoré nie sú v systéme implementované.

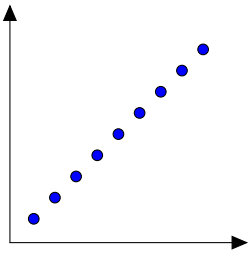
Kapitola 1

Závislosť náhodných veličín

V praktických situáciách pri pozorovaní náhodných veličín častokrát uvažujeme nad viacerými veličinami súčasne a skúmame, akým spôsobom sa ovplyvňujú. Zaujíma nás, aký je vzťah (závislosť) medzi danými veličinami, ktorý môže siahať od nezávislosti až po úplnú závislosť.

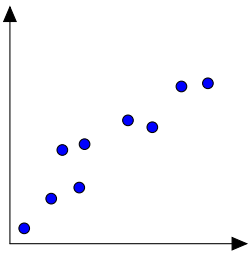
Táto kapitola sa zaoberá základnými typmi závislosti medzi náhodnými veličinami. Pri jej tvorbe sme vychádzali z literárneho zdroja [6].

- **Funkčná (deterministická, pevná) závislosť**
  popisuje vzťah, ktorý sa prejavuje s istotou, teda s pravdepodobnosťou rovnou jednej. Napr.: nech X je veľkosť strany štvorca a Y je plocha štvorca, potom platí: $Y = X^2$. Ide o vzťah daný funkčným predpisom $Y = f(X)$, čo znamená, že každej realizácii náhodnej veličiny X odpovedá (je priradená) práve jedna realizácia náhodnej veličiny Y. Funkčnú závislosť preto nazývame aj pevnou závislosťou.



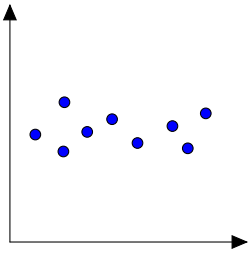
Obrázok 1.1: *Funkčná závislosť*

- **Stochastická (štatistická, voľná) závislosť**
  je vzťah, v ktorom jedna náhodná veličina ovplyvňuje s určitou pravdepodobnosťou druhú náhodnú veličinu. Napr.: nech X je nadmorská výška a Y je teplota vzduchu, potom má teplota vzduchu so stúpajúcou nadmorskou výškou tendenciu klesať. Ide o vzťah, kedy každej realizácii náhodnej veličiny X môže odpovedať viac realizácií náhodnej veličiny Y. Jedná sa o voľnú závislosť medzi kvantitatívnymi náhodnými veličinami.



Obrázok 1.2: *Stochastická závislosť*

- **Stochastická nezávislosť**
  je opakom stochastickej závislosti, kedy sa náhodné veličiny navzájom neovplyvňujú. Ako jednoznačný príklad uveďme X ako hustotu vody a Y ako objem vody. Hustota vody nezávisí od jej objemu. Veličiny X, Y sú nezávislé, ak $F(x,y)$[1] $= F_1(x).F_2(y)$ pre $\forall(x,y) \in \mathbb{R}^2$.



Obrázok 1.3: *Stochastická nezávislosť*

Dané typy závislosti sú doprevádzané obrázkami na ukážku toho, ako môžu vypadať jednotlivé vzťahy graficky.

Závislosti, ktoré obvykle sledujeme v oblasti spoločenských a biologických vied, nemajú čisto funkčne deterministický charakter, preto je pre ich analýzu nutné použitie štatistických metód. Problematikou závislostí náhodných veličín sa zaoberajú dva obory štatistiky: korelačná a regresná analýza. V nasledujúcej kapitole sa budeme venovať výhradne korelačnej analýze, presnejšie jednoduchej korelačnej analýze, ktorá bude obsahom celej práce.

[1] Združená distribučná funkcia náhodných veličín sa v tomto prípade rovná súčinu marginálnych distribučných funkcií. Definíciu distribučnej a marginálnej funkcie viď [1, str. 37].

Kapitola 2

Hodnotenie závislosti dvoch náhodných veličín

Hodnotením závislosti dvoch náhodných veličín sa zaoberá jednoduchá korelačná analýza, ktorá kladie dôraz viac na intenzitu vzájomného vzťahu než na skúmanie veličín v smere príčina – následok (regresia). Závislosti, ktoré skúma sú predovšetkým lineárne, kde korelácia, z latinského correlátio [cor – spolu + relatio – vzťah], je mierou lineárneho vzťahu. Dôležitou skutočnosťou je, že korelácia nie je kauzalita. Veličiny, ktoré spolu korelujú, sú pravdepodobne navzájom závislé, ale nemožno z toho usúdiť, že by sa podmieňovali.

Úlohou korelačnej analýzy je koreláciu identifikovať, kvantifikovať a štatisticky otestovať. Nevyhnutnou súčasťou je logický rozbor problému, a to z hľadiska významu samotnej korelácie, ktorá môže byť skreslená alebo nemusí vôbec existovať. Pri použití dvojrozmerných metód sa často vyskytuje falošná korelácia, čo je neexistujúca korelácia medzi premennými X a Y, ktorá sa dôsledkom iných (nezohľadnených) premenných zdá ako silná. Hodnotu korelácie posúva taktiež nedostatok homogenity vo vzorke a formálny vzťah medzi veličinami (korelácia percentuálnych charakteristík, ktoré sa doplňujú do 100 %).

V tejto kapitole sa zoznámime s rôznymi koeficientami, na základe ktorých budeme posudzovať existenciu a silu závislostí. Hlavnými literárnymi zdrojmi sú [4], [3], [5].

2.1 Kovariancia

Lineárny vzťah môže byť priamy, tj. s rastúcimi (resp. klesajúcimi) hodnotami jednej veličiny rastú (resp. klesajú) hodnoty druhej veličiny a naopak; alebo nepriamy, tj. s rastúcimi hodnotami jednej veličiny klesajú hodnoty druhej a naopak.

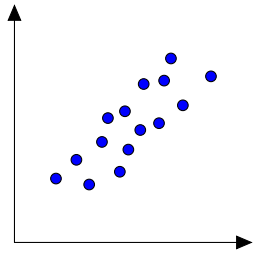
O tom, či sú dve náhodné veličiny X a Y vo vzájomnom lineárnom vzťahu (priamom, nepriamom), sa môžeme presvedčiť na základe kovariancie $C(X,Y)$ definovanej ako stredná hodnota súčinu centrovaných náhodných veličín:

$$C(X,Y) = E([X - E(X)][Y - E(Y)]) = E(XY) - E(X)E(Y). \tag{2.1}$$

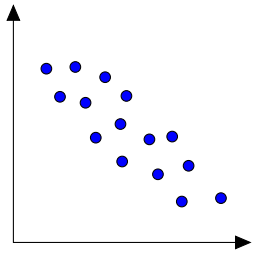
- Ak $C(X,Y) > 0$, *medzi X a Y existuje priamy lineárny vzťah.*
- Ak $C(X,Y) < 0$, *medzi X a Y existuje nepriamy lineárny vzťah.*
- Ak $C(X,Y) = 0$, *medzi X a Y neexistuje lineárny vzťah (nekorelovanosť).*

Z hodnoty kovariancie sme schopní určiť smer lineárneho vzťahu, ale nie jeho silu. Dôvodom je, že kovariancia závisí od jednotiek, v ktorých sú náhodné veličiny merané a nadobúda preto hodnoty z intervalu $(-\infty, \infty)$.

Dané závislosti môžeme orientačne posúdiť z bodového grafu, v ktorom je každá dvojica hodnôt znázornená jedným bodom v rovine. V nasledujúcich troch obrázkoch znázorníme korelačný vzťah medzi veličinami odpovedajúci hodnote kovariancie.

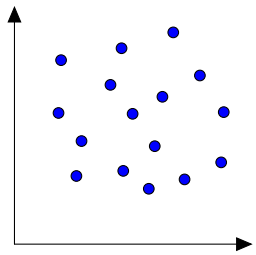


Obrázok 2.1: *Kladná kovariancia*

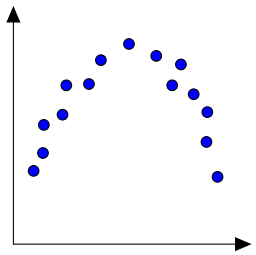


Obrázok 2.2: *Záporná kovariancia*

Z obrázkov 2.1 a 2.2 je zrejmé, že sa jedná o priamy, resp. nepriamy lineárny vzťah, kedže sa dané hodnoty menia rovnakým, resp. opačným smerom, pričom priebeh tejto závislosti je možné schematicky popísať priamkou. Obrázok 2.3 zachytáva dve odlišné situácie, pritom obe odrážajú takmer nulovú hodnotu kovariancie. Je to prostým dôsledkom toho, že kovariancia hovorí o existencii lineárneho vzťahu, ktorý sa ani v jednom z grafov nevyskytuje. Preto nulovosť kovariancie nemusí hovoriť nič o obecnom vzťahu pozorovaných veličín, ktorý môže byť aj nelineárny.



(a) *lineárna nezávislosť*



(b) *nelineárna závislosť*

Obrázok 2.3: *Takmer nulová kovariancia*

2.2 Korelačné koeficienty

Zhrnutím vlastností kovariancie z predchádzajúcej časti dospievame k záveru, že kovariancia nie je najvhodnejšou mierou závislosti. Používa sa väčšinou ako pomocný nástroj pri meraní intenzity lineárneho vzťahu.

Najpoužívanejšou charakteristikou intenzity lineárneho vzťahu medzi dvoma náhodnými veličinami je korelačný koeficient. Existuje rada korelačných koeficientov, ktoré aplikujeme v závislosti od typu náhodných veličín. V našom prípade sa obmedzíme na dva

korelačné koeficienty [1], konkrétne Pearsonov a Spearmanov korelačný koeficient, určujúce závislosti intervalových, pomerových alebo ordinálnych veličín.

2.2.1 Pearsonov koeficient korelácie

Kovariancia náhodných veličín X a Y bola definovaná ako stredná hodnota súčinu centrovaných veličín X a Y.

Pearsonov koeficient korelácie definujeme ako strednú hodnotu súčinu štandardizovaných náhodných veličín:

$$R(X,Y) = E\left(\frac{X-E(X)}{\sqrt{Var(X)}} \cdot \frac{Y-E(Y)}{\sqrt{Var(Y)}}\right) = \frac{C(X,Y)}{\sqrt{Var(X)} \cdot \sqrt{Var(Y)}} \tag{2.2}$$

pre $\sqrt{Var(X)}$, $\sqrt{Var(Y)} > 0$, inak 0.

Štandardizáciou náhodných veličín sme takto získali novú bezrozmernú mieru lineárnej závislosti, ktorá nadobúda hodnoty z intervalu $\langle -1;1 \rangle$. Znamienko korelačného koeficienta závisí od kovariancie, podľa ktorej hodnotu koeficienta interpretujeme:

- $R(X,Y) > 0 \Leftrightarrow$ *priama lineárna závislosť,*
- $R(X,Y) < 0 \Leftrightarrow$ *nepriama lineárna závislosť,*
- $R(X,Y) = 0 \Leftrightarrow$ *lineárna nezávislosť (nekorelovanosť).*

Pre zjednodušenie budeme koeficient $R(X,Y)$ označovať ako R. Čím viac sa $|R|$ blíži k 1, tým je lineárna závislosť silnejšia a naopak, čím viac sa $|R|$ blíži k 0, tým je lineárna závislosť slabšia. Svoje extrémne hodnoty (tj. $-1,1$) nadobúda vtedy, keď platí $P(Y = a+bX) = 1$, pre $a,b \in \mathbb{R}\,(b \neq 0)$ a všetky realizácie veličín X, Y ležia na priamke. V takomto prípade ide o funkčnú priamu ($b > 0$), popr. nepriamu ($b < 0$) lineárnu závislosť.

Vzhľadom k možnosti existencie nelineárneho vzťahu medzi X a Y nie je možné pri nulovom koeficiente R (tj. $C(X,Y) = 0$) označiť veličiny X a Y za nezávislé. Nulová hodnota koeficienta je nutnou podmienkou nezávislosti, nie však podmienkou postačujúcou. Nekorelovanosť odpovedá nezávislosti v prípade, že veličiny X a Y pochádzajú z dvojrozmerného normálneho rozdelenia (dôkaz viď nasledujúca kapitola, sekcia 3.1).

Predstavu o význame hodnôt koeficienta korelácie získame rovnako ako pri kovariancii: z bodových grafov. Podľa charakteru rozloženia bodov v grafe môžeme odhadnúť, aká silná, resp. žiadna závislosť medzi veličinami existuje.

Kedže koeficient korelácie meria silu lineárneho vzťahu, potom tesnosť bodov okolo pomyselnej priamky indikuje jeho veľkosť, pričom znamienko je oplyvnené smerom tejto závislosti. Ak sa vrátime k obrázkom z prechádzajúcej sekcie, silnejšiu závislosť zachytávajú obrázky 2.1, 2.2, kde ležia body blízko pomyselnej priamky, o čom svedčia aj konkrétne hodnoty korelačného koeficienta $0,81$ a $-0,85$. Z obrázku 2.3 by sme očakávali nulovú hodnotu korelačného koeficienta, v oboch prípadoch, kedže na obrázku 2.3a, sú body rozptýlené do „kruhu“ a na 2.3b je zjavná nelineárna závislosť. Hodnoty korelačných

[1] O iných korelačných koeficientoch sa môžeme dočítať v literatúre [11], [2].

koeficientov sú $-0,06$ a $0,12$. Prakticky tieto výsledky zodpovedajú nášmu očakávaniu, koeficient vzťahujúci k obrázku 2.3a je takmer nulový a o niečo vyššia hodnota koeficienta vzťahujúceho k obrázku 2.3b je spôsobená jeho použitím na veličiny závisle nelineárne.

Výberové charakteristiky

Výpočet korelačného koeficienta je podmienený znalosťou konkrétneho simultánneho rozdelenia náhodného vektora $(X,Y)'$, čo sa prakticky stáva veľmi zriedka. V praxi sa preto odkazujeme na náhodný výber[2] $(\mathbb{X},\mathbb{Y}) = ((X_1,Y_1)',\ldots,(X_n,Y_n)')$ rozsahu n z dvojrozmerného rozdelenia s distribučnou funkciou $F(x,y)$, kedy môžeme charakteristiky náhodných veličín odhadnúť pomocou výberových charakteristík (štatistík). Výberové charakteristiky sú funkcie daného náhodného výberu, ktoré nezávisia od neznámeho parametra rozdelenia, ktorým sa náhodný výber riadi.

Definujme nasledovné štatistiky:

1. **výberový priemer**

$$M = \frac{1}{n}\sum_{i=1}^{n} X_i,$$

2. **výberový rozptyl**

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - M)^2 = \frac{1}{n-1}\left(\sum_{i=1}^{n} X_i^2 - nM^2\right),$$

2.1 **výberová smerodajná odchýlka**

$$S = \sqrt{S^2},$$

3. **výberová kovariancia**

$$S_{12} = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - M_1)(Y_i - M_2) = \frac{1}{n-1}\left(\sum_{i=1}^{n} X_iY_i - nM_1M_2\right),$$

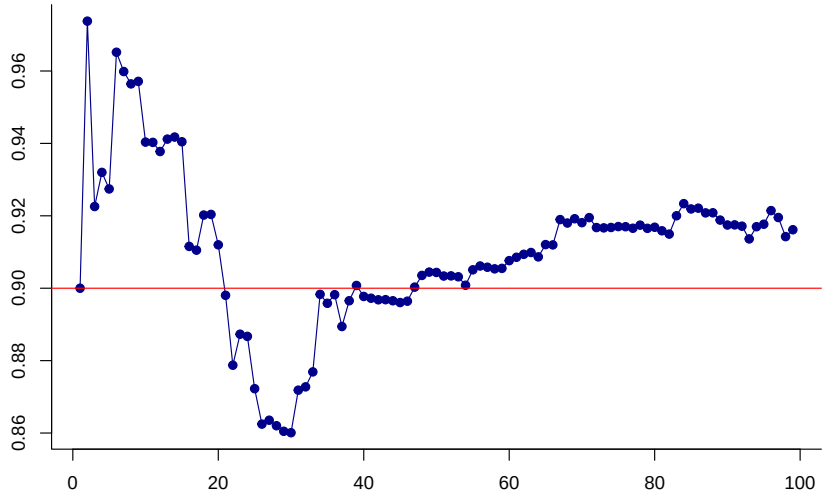
kde $M_1 = \frac{1}{n}\sum_{i=1}^{n} X_i$ a $M_2 = \frac{1}{n}\sum_{i=1}^{n} Y_i$.

Výberový korelačný koeficient je potom definovaný ako pomer výberovej kovariancie k súčinu výberových smerodajných odchýliek veličín X, Y:

$$R_{12} = \frac{S_{12}}{S_1S_2} = \frac{1}{n-1}\sum_{i=1}^{n}\frac{X_i - M_1}{S_1}\cdot\frac{Y_i - M_2}{S_2}, \tag{2.3}$$

pre $S_1S_2 \neq 0$.

[2] Dvojrozmerným náhodným výberom rozsahu n sa rozumie postupnosť n nezávislých, rovnako rozdelených náhodných vektorov. Presnú definíciu viď [4, str.52]



Obrázok 2.4: *Ilustrácia vychýlenosti*

Výberový korelačný koeficient ako bodový odhad Pearsonovho koeficienta korelácie preberá všetky jeho vlastnosti. Bohužiaľ sa nejedná o nestranný, ale vychýlený bodový odhad. Vychýlenosť je však zanedbateľná pri rozsahu výberu $n > 30$. Túto skutočnosť zachytáva obrázok 2.4, pre ktorý sme vygenerovali 100 náhodných vektorov z dvojrozmerného normálneho rozdelenia $N_2\left(\binom{0}{0}, \binom{1 \quad 0,9}{0,9 \quad 1}\right)$ a následne vypočítali korelačné koeficienty medzi zložkami tohto náhodného výberu v závislosti od rozsahu. Ako môžeme vidieť, realizácia výberového korelačného koeficienta ozn. r sa pri výbere väčšieho rozsahu pohybuje okolo hodnoty $0,9$.

2.2.2 Spearmanov koeficient korelácie

Na určenie závislosti medzi dvoma náhodnými veličinami X, Y sa používa často tzv. poradová korelácia. Spočíva v nahradení realizácií náhodných veličín poradovými číslami, tj. pôvodné hodnoty x_i a y_i, pre $i = 1,2,\ldots,n$, nahradzujeme ich poradovými číslami R_i a Q_i. Pokiaľ sa nejaká pôvodná hodnota opakuje viackrát [3], priradíme každej takejto hodnote rovnaké poradové číslo určené ako aritmetický priemer poradových čísel, ktoré by tieto hodnoty mali, keby boli rôzne a nasledovali za sebou.

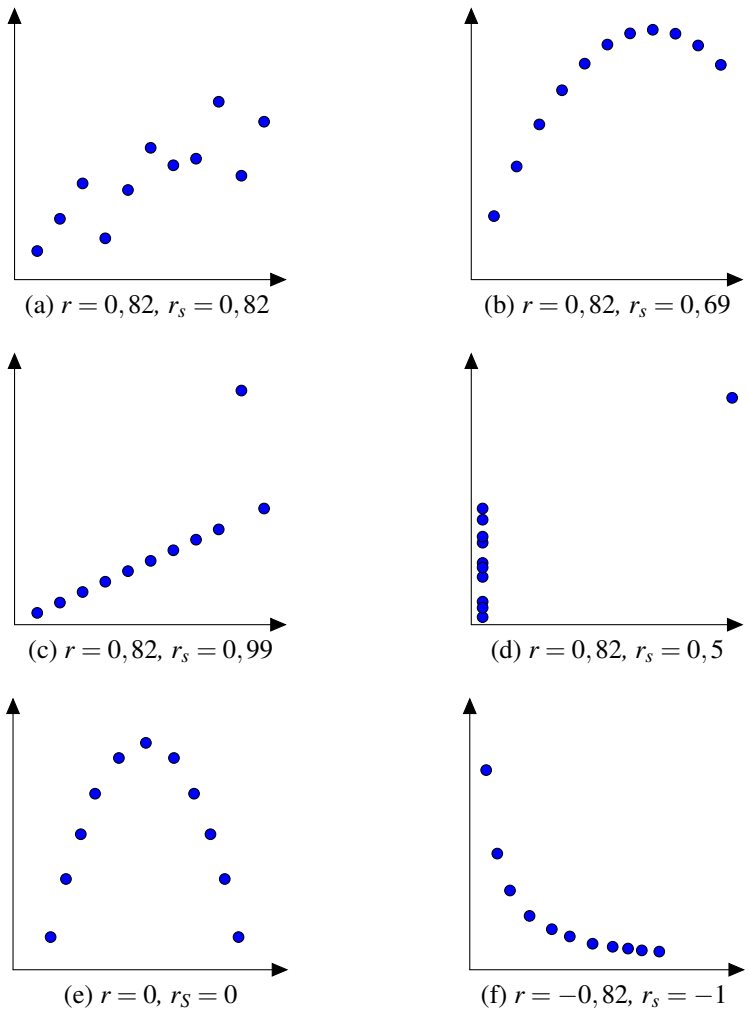
Ide o neparametrickú metódu popisujúcu monotónnu závislosť, nielen lineárnu, ale obecne rastúcu alebo klesajúcu. Neparametrickou charakteristikou závislosti náhodných veličín X, Y je Spearmanov[4] koeficient poradovej korelácie daný vzťahom

$$r_s = 1 - \frac{6}{n(n^2-1)} \sum_{i=1}^{n} (R_i - Q_i)^2. \tag{2.4}$$

[3] Príliš veľký výskyt rovnakých hodnôt spôsobuje skreslenie Spearmanovho koeficienta, a preto je vhodné použiť korigovaný Spearmanov koeficient definovaný v [11, str. 396].

[4] Spearmanov koeficient poradovej korelácie je odhad jeho teoretickej hodnoty R_s.

Spearmanov koeficient poradovej korelácie je totožný s výberovým koeficientom korelácie aplikovaným na poradie zložiek náhodného výberu; dôkaz viď [2, str. 560]. Nadobúda obdobne hodnoty z intervalu $\langle -1;1\rangle$. Hodnoty blízke 0 ukazujú na slabšiu poradovú závislosť premenných, hodnoty blízke 1 či -1 na tesnejšiu poradovú závislosť (priamu, nepriamu). Rovnosť $r_s = 1$ platí, ak $R_i - Q_i = 0$, tj. pri zhodných poradiach a naopak, rovnosť $r_s = -1$ je splnená, ak sú poradia presne opačné. Potom realizácia daného náhodného výberu (x_i, y_i), $i = 1,2,\ldots,n$, leží na nejakej monotónnej krivke, ako môžeme vidieť na obrázku 2.5f.



(a) $r = 0,82$, $r_s = 0,82$ (b) $r = 0,82$, $r_s = 0,69$

(c) $r = 0,82$, $r_s = 0,99$ (d) $r = 0,82$, $r_s = 0,5$

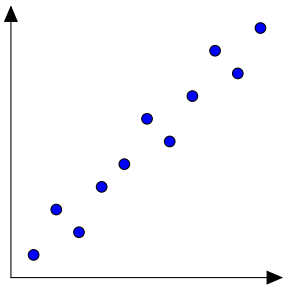
(e) $r = 0$, $r_S = 0$ (f) $r = -0,82$, $r_s = -1$

Obrázok 2.5: *Rôzne bodové konfigurácie a k nim dopočítané hodnoty Pearsonovho výberového (r) a Spearmanovho (r_s) korelačného koeficienta*

Obrázok 2.5 zobrazuje rôzne bodové konfigurácie na uľahčenie interpretácie hodnôt Spearmanovho a tiež Pearsonovho výberového koeficienta korelácie, pričom kladie dôraz na porovnanie toho, ako sa oba koeficienty chovajú v závislosti od druhu a sily daných vzťahov medzi veličinami. Ak ide o nelineárnu monotónnu závislosť, koeficient $|r_s|$ je väčší než koeficient $|r|$. Graficky a číselne na 2.5f. To, že je Spearmanov koeficient založený na poradových číslach náhodných veličín, umožňuje jeho rezistentnosť voči odľahlým

hodnotám, na rozdiel od Pearsonovho koeficienta, ktorý je na odľahlé hodnoty citlivý, viď obrázok 2.5c. Závislosť zobrazená na obrázku 2.5e je závislosť kvadratická, na ktorú ani jeden z koeficientov nemá dosah.

Meranie závislosti dvoch poradí je vlastne špeciálnym prípadom merania lineárnej závislosti dvoch náhodných veličín, čo sa odráža aj v tom, že Spearmanov koeficient je rovný Pearsonovmu výberovému koeficientu veličín X a Y, kedy obe nadobúdajú hodnoty $1, 2, \ldots, n$. Tento prípad ilustruje obrázok 2.6.



Obrázok 2.6: *Zhodné korelačné koeficinety* ($r = r_s = 0,97$)

Z vlastností Spearmanovho koeficienta vyplýva, že sa používa v situáciách, kedy skúmané náhodné veličiny majú ordinálny charakter a nepredpokladáme medzi nimi čisto lineárny vzťah ani typ rozdelenia.

Kapitola 3

Štatistická indukcia – teoretická časť

Dôležitou súčasťou analýzy dát je testovanie štatistických hypotéz. Pri skúmaní závislosti dvoch náhodných veličín je obsahom testovanej nulovej hypotézy tvrdenie, že náhodné veličiny sú nezávislé oproti alternatívnej hypotéze, že náhodné veličiny sú závislé. V tejto kapitole sa zameriame na testy nezávislosti pomocou už známych korelačných koeficientov a taktiež na testy modifikovaných hypotéz o korelačných koeficientoch.

Ako každé testovanie, aj testovanie nezávislosti musí spĺňať určité predpoklady. Predpoklady testovania nezávislosti pri použití Pearsonovho (výberového) korelačného koeficienta sú, že náhodný výber pochádza z dvojrozmerného normálneho rozdelenia pre veličiny intervalového či pomerového typu, medzi ktorými možno predpokladať linearitu. Pri porušení týchto predpokladov, alebo v prípade veličín čisto ordinálneho typu, používame Spearmanov korelačný koeficient. Pri tvorbe kapitoly sme čerpali z [11], [9], [4], [3].

3.1 Predpoklad normality

Uvažujme normálne rozdelený náhodný vektor $(X,Y)'$ so simultánnou hustotou

$$f(x,y) = \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}} e^{-\frac{1}{2(1-\rho^2)}\left[\left(\frac{x-\mu_1}{\sigma_1}\right)^2 - 2\rho\frac{x-\mu_1}{\sigma_1}\frac{y-\mu_2}{\sigma_2} + \left(\frac{y-\mu_2}{\sigma_2}\right)^2\right]},$$

kde $\mu_1 = E(X)$, $\mu_2 = E(Y)$, $\sigma_1^2 = Var(X)$, $\sigma_2^2 = Var(Y)$ a $\rho = R(X,Y)$.

Marginálne hustoty sú

$$f_1(x) = \frac{1}{\sigma_1\sqrt{2\pi}} e^{-\frac{(x-\mu_1)^2}{2\sigma_1^2}}, \; f_2(y) = \frac{1}{\sigma_2\sqrt{2\pi}} e^{-\frac{(y-\mu_2)^2}{2\sigma_2^2}}.$$

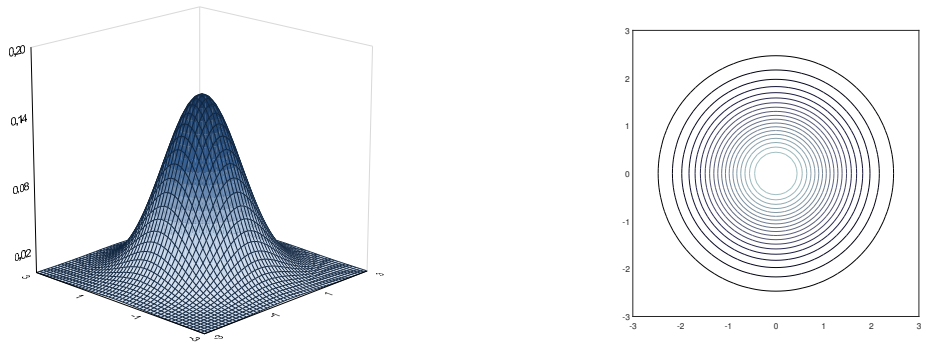
Z predchádzajúcich kapitol vieme, že veličiny X a Y sú nekorelované, ak sa ich korelačný koeficient R, v literatúre často označovaný ako ρ, rovná nule a nezávislé, pokiaľ platí $F(x,y) = F_1(x)F_2(y)$ pre $\forall(x,y) \in \mathbb{R}^2$, v obecnom prípade. Dosadením nulovej hodnoty korelačného koeficienta do vzorca simultánnej hustoty normálneho rozdelenia potom dostávame

$$f(x,y) = \frac{1}{2\pi\sigma_1\sigma_2} e^{-\frac{1}{2}\left[\left(\frac{x-\mu_1}{\sigma_1}\right)^2 + \left(\frac{y-\mu_2}{\sigma_2}\right)^2\right]} = \frac{1}{\sqrt{2\pi}\sigma_1} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma_1}\right)^2} \cdot \frac{1}{\sqrt{2\pi}\sigma_2} e^{-\frac{1}{2}\left(\frac{y-\mu}{\sigma_2}\right)^2}.$$

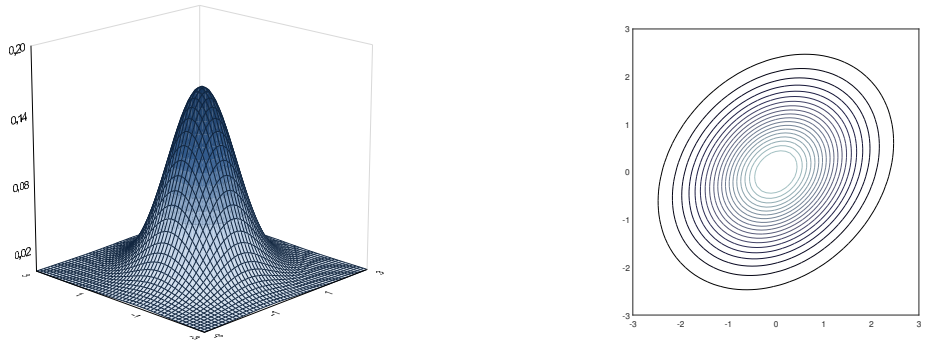
Je zrejmé, že v tomto prípade platí $f(x,y) = f_1(x)f_2(y)$ pre $\forall(x,y) \in \mathbb{R}^2$. Normalita náhodných veličín teda zaisťuje, že nekorelovanosť odpovedá nezávislosti, a preto sa pri testovaní nezávislosti môžeme opierať o Pearsonov korelačný koeficient. To, ako sa tento koeficient (resp. jeho odhad) aplikuje pri testovaní, si ukážeme v nasledujúcej časti.

Dvojrozmernú normalitu vieme orientačne posúdiť z bodového grafu. V prípade dostatočne veľkého počtu pozorovaní by mali body grafu vytvoriť obrazec elipsy, pretože vrstevnice hustoty dvojrozmerného normálneho rozdelenia sú elipsy.

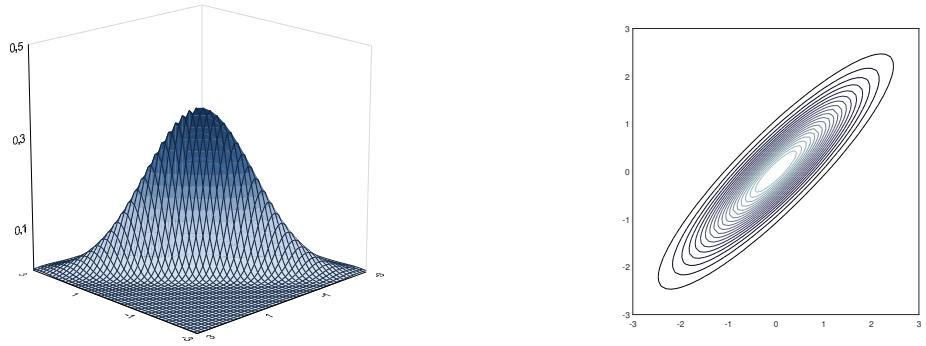
Na ukážku toho si zobrazíme tri rôzne grafy hustoty dvojrozmerného normálneho rozdelenia s nulovými strednými hodnotami a jednotkovými rozptylmi, ktoré sa líšia v hodnote parametra korelačného koeficienta, a k nim odpovedajúce vrstevnice. V prípade nulového korelačného koeficienta ide o štandardizované dvojrozmerné normálne rozdelenie, kedy vrstevnice hustoty tohto rozdelenia tvoria kružnice, viď obrázok 3.1. S rastúcou, resp. klesajúcou absolútnou hodnotou koeficienta majú vrstevnice elipsovitejší tvar ako môžeme vidieť na obrázkoch 3.2, 3.3.



Obrázok 3.1: *Graf hustoty a vrstevnice hustoty štandardizovaného dvojrozmerného normálneho rozdelenia*



Obrázok 3.2: *Graf hustoty a vrstevnice hustoty dvojrozmerného normálneho rozdelenia s parametrami* $\mu_1 = 0$, $\mu_2 = 0$, $\sigma_1^2 = 1$, $\sigma_2^2 = 1$, $\rho = 0,25$



Obrázok 3.3: *Graf hustoty a vrstevnice hustoty dvojrozmerného normálneho rozdelenia s parametrami* $\mu_1 = 0$, $\mu_2 = 0$, $\sigma_1^2 = 1$, $\sigma_2^2 = 1$, $\rho = 0,9$

3.2 Test významnosti korelačného koeficienta

Uvažujme náhodný výber $(\mathbb{X}, \mathbb{Y})$ z dvojrozmerného normálneho rozdelenia rozsahu n, s korelačným koeficientom R a výberovým korelačným koeficientom R_{12}.

Výberový korelačný koeficient R_{12} interpretujeme ako bodový odhad korelačného koeficienta R. Kedže sa jedná o výberovú charakteristiku, jeho nenulová hodnota nemusí znamenať, že aj skutočný korelačný koeficient je nenulový. V praxi sa preto vyžaduje otestovať, či je zistená hodnota výberového korelačného koeficienta štatisticky významná pre tvrdenie, že medzi skúmanými veličinami X, Y skutočne existuje závislosť.

Odpoveď na túto otázku môžeme získať z testu významnosti korelačného koeficienta, ktorý uskutočníme pomocou kritického oboru pre štatistiku T, viď (3.1), alebo intervalu, ktorý je na tejto štatistike založený.

Testujeme nulovú hypotézu

$$H_0 : R = 0,$$

tj. hypotézu o nezávislosti náhodných veličín X a Y, oproti niektorej alternatívnej hypotéze:

- **obojstranná alternatíva:**

$$H_1 : R \neq 0, \text{ tj. veličiny } X \text{ a } Y \text{ sú stochasticky závislé,}$$

- **pravostranná alternatíva:**

$$H_1 : R > 0, \text{ tj. veličiny } X \text{ a } Y \text{ sú kladne korelované,}$$

- **ľavostranná alternatíva:**

$$H_1 : R < 0, \text{ tj. veličiny } X \text{ a } Y \text{ sú záporne korelované.}$$

Ako testovacie kritérium používame štatistiku

$$T = \frac{R_{12}\sqrt{n-2}}{\sqrt{1-R_{12}^2}}. \tag{3.1}$$

Štatistika T sa pri platnosti nulovej hypotézy riadi Studentovým rozdelením s $(n-2)$ stupňami voľnosti.

Na zvolenej hladine významnosti α zamietame nulovú hypotézu vtedy, keď sa štatistika T realizuje v kritickom obore W. Kritický obor má tvar

$$W = \left(-\infty, -t_{1-\alpha/2}(n-2)\right\rangle \cup \left\langle t_{1-\alpha/2}(n-2), \infty\right) \textit{ pre obojstrannú alternatívu,}$$
$$W = \left(-\infty, -t_{1-\alpha}(n-2)\right\rangle \textit{ pre ľavostrannú alternatívu,}$$
$$W = \left\langle t_{1-\alpha}(n-2), \infty\right) \textit{ pre pravostrannú alternatívu.}$$

V prípade, že je rozsah výberu $n > 30$, je možné Studentovo rozdelenie aproximovať štandardizovaným normálnym rozdelením $N(0,1)$ a kritický obor upraviť nahradením kvantilov Studentovho rozdelenia kvantilmi štandardizovaného normálneho rozdelenia.

Test hypotézy o nezávislosti môžeme uskutočniť pomocou nasledujúcich intervalov, ktoré sa odvádzajú na základe skutočnosti, že testovacia štatistika T sa za platnosti nulovej hypotézy riadi Studentovaým rozdelením s $(n-2)$ stupňami voľnosti.

$$P\left(t_{\alpha/2}(n-2) < T < t_{1-\alpha/2}(n-2)\right) = 1-\alpha,$$

odkiaľ úpravou nerovností a nahradením odhadu R_{12} koeficientom R dostaneme

$$P\left(\underbrace{\frac{-t_{1-\alpha/2}(n-2)}{\sqrt{t^2_{1-\alpha/2}(n-2)+n-2}}}_{D} < R < \underbrace{\frac{t_{1-\alpha/2}(n-2)}{\sqrt{t^2_{1-\alpha/2}(n-2)+n-2}}}_{H}\right) = 1-\alpha,$$

kde D je dolná a H je horná hranica intervalu, ktorý má tak tvar:

$$(D,H) = \left(\frac{-t_{1-\alpha/2}(n-2)}{\sqrt{t^2_{1-\alpha/2}(n-2)+n-2}}, \frac{t_{1-\alpha/2}(n-2)}{\sqrt{t^2_{1-\alpha/2}(n-2)+n-2}}\right)$$

pre obojstrannú alternatívu,

$$(D,\infty) = \left(\frac{-t_{1-\alpha}(n-2)}{\sqrt{t^2_{1-\alpha}(n-2)+n-2}}, \infty\right)$$

pre pravostrannú alternatívu,

$$(-\infty,H) = \left(-\infty, \frac{t_{1-\alpha}(n-2)}{\sqrt{t^2_{1-\alpha}(n-2)+n-2}}\right)$$

pre ľavostrannú alternatívu.

Ak hodnota výberového korelačného koeficienta nespadá do daného intervalu (D,H), resp. $(-\infty,H)$ alebo (D,∞), nulovú hypotézu o nezávislosti veličín X a Y zamietame na hladine významnosti α v prospech alternatívnej hypotézy.

Využitím štatistického systému pri testovaní hypotézy o korelačnom koeficiente zamietame nulovú hypotézu, pokiaľ je p-hodnota daného testu menšia alebo rovná ako zvolená hladina významnosti α. Toto kritérium platí pre všetky nasledujúce testy.

3.3 Interval spoľahlivosti pre korelačný koeficient

Pri náhodnom výbere z dvojrozmerného normálneho rozdelenia a pre hodnoty koeficientu R blízke nule[1] je možné považovať výberové rozdelenie koeficientu R_{12} za približne normálne. Pre obecný prípad $-1 < R < 1$ sa však výberové rozdelenie s rastúcim R stále viac zošikmuje, a preto je konštrukcia asymptotického intervalu spoľahlivosti pre R založená na Fisherovej Z-transformácii výberového korelačného koeficienta.

Transformovaná náhodná veličina

$$Z = \frac{1}{2}\ln\frac{1+R_{12}}{1-R_{12}} \tag{3.2}$$

má približne normálne rozdelenie so strednou hodnotou $E(Z) = \frac{1}{2}\ln\frac{1+R}{1-R} + \frac{R}{2(n-1)}$ so zanedbateľným druhým sčítancom pri väčšom n a rozptylom $Var(Z) = \frac{1}{n-3}$.

Štandardizáciou veličiny Z dostaneme veličinu

$$U = \frac{Z - E(Z)}{\sqrt{Var(Z)}}, \tag{3.3}$$

ktorá ma asymptoticky štandardizované normálne rozdelenie $N(0,1)$.

Interval spoľahlivosti pre strednú hodnotu veličiny Z, $E(Z)^*$[2] $= \frac{1}{2}\ln\frac{1+R}{1-R}$, vyjadríme z

$$P(u_{\alpha/2} < U < u_{1-\alpha/2}) = 1-\alpha,$$

úpravou nerovností do tvaru

$$P\Bigg(\underbrace{Z - \frac{u_{1-\alpha/2}}{\sqrt{n-3}}}_{D} < E(Z)^* < \underbrace{Z + \frac{u_{1-\alpha/2}}{\sqrt{n-3}}}_{H}\Bigg) = 1-\alpha.$$

Interval spoľahlivosti pre korelačný koeficient R potom dostaneme spätnou trasformáciou.

[1] Ak pre korelačný koeficient platí $|R| < 0,5$ a pre rozsah výberu platí $n > 100$, potom môžeme rozdelenie výberového korelačného koeficienta aproximovať normálnym rozdelením. V tomto prípade sa používa pre koeficient R $(1-\alpha)100\%$ interval spoľahlivosti, uvedený napr. v [9, str.173] alebo v [4, str.134].

[2] Strednú hodnotu berieme so zanedbaným druhým sčítancom.

Keďže $Z = \operatorname{arctgh} R_{12}$, potom $R_{12} = \operatorname{tgh} Z$ a hranice intervalu spoľahlivosti môžeme písať v tvare

$$(D,H) = \left(\operatorname{tgh}\left(Z - \frac{u_{1-\alpha/2}}{\sqrt{n-3}}\right), \operatorname{tgh}\left(Z + \frac{u_{1-\alpha/2}}{\sqrt{n-3}}\right)\right)$$

pre obojstrannú alternatívu,

$$(D,1) = \left(\operatorname{tgh}\left(Z - \frac{u_{1-\alpha}}{\sqrt{n-3}}\right), 1\right)$$

pre pravostrannú alternatívu,

$$(-1,H) = \left(-1, \operatorname{tgh}\left(Z + \frac{u_{1-\alpha}}{\sqrt{n-3}}\right)\right)$$

pre ľavostrannú alternatívu, pričom $\operatorname{tgh} x = \frac{e^x - e^{-x}}{e^x + e^{-x}}$.

Tento interval spoľahlivosti môžeme použiť pri testovaní nulovej hypotézy o danej hodnote koeficienta korelácie, viď sekcia (3.4). Nulovú hypotézu zamietame na asymptotickej hladine významnosti α v prospech alternatívnej hypotézy v prípade, že sa daná hodnota v intervale spoľahlivosti nenachádza.

3.4 Test hypotézy o danej hodnote korelačného koeficienta

Tento test umožňuje overiť zhodu korelačného koeficienta s danou konštantou c, pre ktorú platí $c \in (-1,1)$. Testujeme nulovú hypotézu

$$H_0 : R = c$$

oproti niektorej z alternatívnych hypotéz

$$H_1 : R \neq c,$$
$$H_1 : R > c,$$
$$H_1 : R < c.$$

Testovacia štatistika je

$$U = \left(Z - \frac{1}{2}\ln\frac{1+c}{1-c} - \frac{c}{2(n-1)}\right)\sqrt{n-3}, \tag{3.4}$$

kde Z je Fisherova Z-transformácia daná vzťahom (3.2).

Štatistika U má za platnosti nulovej hypotézy pre $n \geq 10$ asymptoticky rozdelenie $N(0,1)$. Ak vypočítaná hodnota testovacej štatistiky padne do kritického oboru

$$W = \left(-\infty, -u_{1-\alpha/2}\right\rangle \cup \left\langle u_{1-\alpha/2}, \infty\right) \textit{ pre obojstrannú alternatívu,}$$
$$W = \left(-\infty, -u_{1-\alpha}\right\rangle \textit{ pre ľavostrannú alternatívu,}$$
$$W = \left\langle u_{1-\alpha}, \infty\right) \textit{ pre pravostrannú alternatívu,}$$

nulovú hypotézu zamietame na asymptotickej hladine významnosti α.

3.5 Test zhody dvoch korelačných koeficientov

Zoberme do úvahy situáciu, že máme k dispozícii dva nezávislé náhodné výbery rozsahov n a n^* z dvojrozmerných normálnych rozdelení s koeficientami korelácie R a R^*. Zaujíma nás, či sa ich korelačné koeficienty štatisticky významne líšia, tj. či oba výbery pochádzajú z rovnakého základného súboru. Testujeme preto nulovú hypotézu

$$H_0 \colon R = R^*$$

oproti niektorej z alternatívnych hypotéz

$$H_1 \colon R \neq R^*,$$
$$H_1 \colon R > R^*,$$
$$H_1 \colon R < R^*.$$

Ak označíme výberové korelačné koeficienty výberov R_{12} a R^*_{12}, potom ich Fisherove Z-transformácie sú

$$Z = \frac{1}{2}\ln\frac{1+R_{12}}{1-R_{12}} \quad \text{a} \quad Z^* = \frac{1}{2}\ln\frac{1+R^*_{12}}{1-R^*_{12}}. \tag{3.5}$$

Testovacie kritérium je štatistika

$$U = \frac{Z - Z^*}{\sqrt{\frac{1}{n-3} + \frac{1}{n^*-3}}}, \tag{3.6}$$

ktorej rozdelenie, pri platnosti H_0 a podmienke $n \geq 10$, je asymptoticky $N(0,1)$. Kritický obor pre test hypotézy H_0 oproti H_1 je

$$W = \left(-\infty, -u_{1-\alpha/2}\right\rangle \cup \left\langle u_{1-\alpha/2}, \infty\right) \textit{ pre obojstrannú alternatívu,}$$
$$W = \left(-\infty, -u_{1-\alpha}\right\rangle \textit{ pre ľavostrannú alternatívu,}$$
$$W = \left\langle u_{1-\alpha}, \infty\right) \textit{ pre pravostrannú alternatívu.}$$

Ak realizácia štatistiky U leží v kritickom obore W, potom H_0 zamietame na asymptotickej hladine významnosti α.

3.6 Test zhody *k* korelačných koeficientov

Tento test je zovšeobecnením predchádzajúceho testu, ktorý umožňuje testovať zhodu $k > 2$ korelačných koeficientov.

Budeme predpokladať, že máme k dispozícii k navzájom nezávislých náhodných výberov z dvojrozmerných normálnych rozdelení s korelačnými koeficientami $R_1, R_2, \ldots, R_k$ a rozsahmi $n_1, n_2, \ldots, n_k$, pre ktoré platí $n = n_1 + n_2 + \cdots + n_k$. Odpovedajúce výberové korelačné koeficienty označíme ako $R^1_{12}, R^2_{12}, \ldots, R^k_{12}$.

Testujeme nulovú hypotézu

$$H_0 \colon R_1 = R_2 = \cdots = R_k$$

oproti alternatívnej hypotéze

$$H_1: \text{„aspoň dva z } k \text{ koeficientov } R_1, R_2, \cdots, R_k \text{ sú rozdielne“}.$$

Testovacie kritérium je štatistika

$$\chi^2 = \sum_{i=1}^{k} (n_i - 3)(Z_i - b)^2, \tag{3.7}$$

kde $b = \frac{1}{n-3k} \cdot \sum_{i=1}^{k} (n_i - 3) Z_i$ a Z_i je Fisherova Z-transformácia výberového korelačného koeficienta R_{12}^i daná vzťahom (3.2), pre $i = 1, 2, \ldots, k$.

Za platnosti nulovej hypotézy má štatistika χ^2 asymptoticky χ^2 rozdelenie s $(k-1)$ stupňami voľnosti a kritický obor

$$W = \left\langle \chi^2_{1-\alpha}(k-1), \infty \right),$$

pre ktorý zamietame nulovú hypotézu, ak $\chi^2 \in W$.

V prípade, že nulovú hypotézu o zhode k korelačných koeficientov zamietneme, využijeme Tukeyov test, podľa ktorého zistíme, ktoré korelačné koeficienty sa od seba štatisticky významne líšia. Pre každú dvojicu $i < j$ $(i, j = 1, \ldots, k)$ určíme nerovnosť

$$|Z_i - Z_j| \geq q_{k,\infty}(\alpha) \cdot \sqrt{\frac{1}{2}\left(\frac{1}{n_i - 3} + \frac{1}{n_j - 3}\right)}, \tag{3.8}$$

kde $q_{k,\infty}(\alpha)$ je tabelovaná hodnota uvedená v štatistických tabuľkách.

Štatisticky významne sa od seba líšia tie koeficienty, pre ktoré daná nerovnosť platí.

3.7 Test významnosti Spearmanovho korelačného koeficienta

Konečne sa dostávame k problému, kedy nemôžeme apriorne prijať predpoklad normality. V takomto prípade sa pri testovaní nezávislosti náhodných veličín X a Y používa metóda určená pre náhodné veličiny ordinálneho typu. Táto neparametrická metóda je založená na Spearmanovom korelačnom koeficiente definovanom v kapitole 2, ktorý spočíva v nahradení konkrétnych hodnôt náhodných veličín ich poradovými číslami.

Obdobne testujeme nulovú hypotézu o nulovej hodnote teoretického Spearmanovho koeficientu korelácie

$$H_0: R_s = 0,$$

tj. hypotézu, že veličiny X a Y sú poradovo nezávislé oproti niektorej z alternatívnych hypotéz

$$H_1: R_s \neq 0 \text{ , } \textit{tj. veličiny X a Y sú poradovo závislé,}$$
$$H_1: R_s > 0 \text{ , } \textit{tj. medzi X a Y existuje priama poradová závislosť,}$$
$$H_1: R_s < 0 \text{ , } \textit{tj. medzi X a Y existuje nepriama poradová závislosť.}$$

Pri výbere malého rozsahu $(n < 20)$ slúži ako testovacie kritérium Spearmanov koeficient korelácie r_s vypočítaný podľa vzťahu (2.4), pre ktorý sa kritický obor W zostavuje použitím špeciálnych tabelovaných kritických hodnôt[3] štatistiky r_s nasledovne:

$$W = \left(-1, -r_{s,1-\alpha/2}(n)\right\rangle \cup \left\langle r_{s,1-\alpha/2}(n), 1\right) \textit{ pre obojstrannú alternatívu,}$$
$$W = \left(-1, -r_{s,1-\alpha}(n)\right\rangle \textit{ pre ľavostrannú alternatívu,}$$
$$W = \left\langle r_{s,1-\alpha}(n), 1\right) \textit{ pre pravostrannú alternatívu.}$$

Hypotézu o poradovej nezávislosti zamietame na hladine významnosti α, ak $r_s \in W$.

V závislosti od rozsahu daného výberu existujú asymptotické varianty tohto testu. Pre $n > 20$ používame testovaciu štatistiku (3.1) nahradením výberového korelačného koeficienta Spearmanovým koeficientom korelácie, tj.

$$T = \frac{r_s\sqrt{n-2}}{\sqrt{1-r_s^2}}, \tag{3.9}$$

ktorá má za platnosti nulovej hypotézy asymptoticky Studentovo rozdelenie s $(n-2)$ stupňami voľnosti.

Na hladine významnosti α zamietame nulovú hypotézu v prípade, že sa štatistika T realizuje v kritickom obore W totožnom kritickému oboru štatistiky (3.1):

$$W = \left(-\infty, -t_{1-\alpha/2}(n-2)\right\rangle \cup \left\langle t_{1-\alpha/2}(n-2), \infty\right) \textit{ pre obojstrannú alternatívu,}$$
$$W = \left(-\infty, -t_{1-\alpha}(n-2)\right\rangle \textit{ pre ľavostrannú alternatívu,}$$
$$W = \left\langle t_{1-\alpha}(n-2), \infty\right) \textit{ pre pravostrannú alternatívu.}$$

Pre $n > 30$ používame testovaciu štatistiku

$$U = r_s\sqrt{n-1}, \tag{3.10}$$

ktorá má za planosti nulovej hypotézy asymptoticky štandardizované normálne rozdelenie $N(0,1)$. Kritický obor má obdobne tvar

$$W = \left(-\infty, -u_{1-\alpha/2}\right\rangle \cup \left\langle u_{1-\alpha/2}, \infty\right) \textit{ pre obojstrannú alternatívu,}$$
$$W = \left(-\infty, -u_{1-\alpha}\right\rangle \textit{ pre ľavostrannú alternatívu,}$$
$$W = \left\langle u_{1-\alpha}, \infty\right) \textit{ pre pravostrannú alternatívu.}$$

Štatistický systém STATISTICA využíva pri testovaní štatistiku (3.9), bez ohľadu na rozsah výberu.

[3] Tabelované kritické hodnoty nájdeme v štatistických tabuľkách.

Kapitola 4

Štatistická indukcia – praktická časť

Doposiaľ získané znalosti z predchádzajúcich kapitol tvoria nutný teoretický základ na nasledujúce použitie v praxi. Obsahom tejto kapitoly bude riešenie praktických príkladov, na ktorých si predovšetkým ukážeme aplikáciu testov uvedených v kapitole 3. Riešenie jednotlivých príkladov bude zahŕňať ručný výpočet a výpočet za pomoci systému STATISTICA. Literárne zdroje použité pri tvorbe príkladov uvádzame vždy samostatne za každým zadaním príkladu.

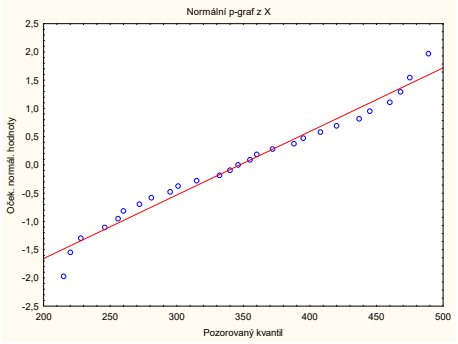
Poznámka. Pri testovaní hypotéz budeme používať hladinu významnosti $\alpha = 0,05$, pokiaľ nebude uvedené inak.

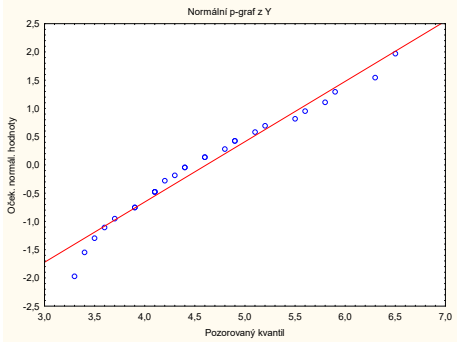
Príklad 1 (Test významnosti korelačného koeficienta, 3.2.)**.**
Sú dané údaje o 27 vybraných pozemkoch, na ktorých poľnohospodárske závody pestujú v určitej oblasti ozimný jačmeň. Nadmorskú výšku pozemku (v metroch) označíme X, hektárový výnos jačmeňa (v t/ha) Y. Hodnota výberového korelačného koeficienta medzi danými veličinami je −0,93. Testujte na hladine významnosti 1 %, či skutočne medzi nimi nepriama závislosť existuje. [12, str. 758]
(Dáta dostupné na `https://goo.gl/z2sk2q`; KORELACE – Environment – E702.)

Riešenie (*Ručný výpočet.*)**.**
Pred testovaním závislosti sa uistíme, či daný náhodný výber pochádza z dvojrozmerného normálneho rozdelenia. Postupujeme tak, že najskôr otestujeme normalitu náhodných veličín X, Y a následne zobrazíme dvojrozmerný bodový diagram s elipsou 95% konštantnej hustoty pravdepodobnosti, na základe ktorého posúdime dvojrozmernú normalitu dát. Toto overenie uskutočníme využitím systému STATISTICA.

Vytvoríme dátový súbor s 2 premennými a 27 prípadmi. Pomenujeme ich X, Y a do tabuľky zapíšeme odpovedajúce hodnoty. (Tabuľku neuvádzame z dôvodu rozsiahleho počtu pozorovaní.)

Grafické overenie jednorozmernej normality: *Grafy – 2D grafy – Normální pravděpodobnostní grafy – odškrtneme Neurčovat prům. pozici svázaných pozorování – Proměnné; X, Y; OK – OK*.



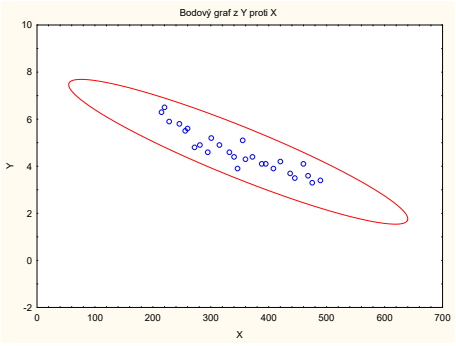


Testy hypotézy o normálnom rozdelení náhodných veličín: *Statistiky – Základní statistiky/tabulky – Tabulky četností; OK – záložka Normalita; zaškrtneme Lilieforsův test a Shapiro-Wilkův W test – Proměnné; X, Y; OK – Testy normality*.

	Testy normality				
Proměnná	N	max D	Lilliefors p	W	p
X	27	0,0803	p > .20	0,9557	0,2929
Y	27	0,1142	p > .20	0,9574	0,3221

Normálne pravdepodobnostné grafy svedčia v prospech normálneho rozdelenia náhodných veličín, pretože body ležia takmer na alebo v blízkosti danej priamky, o čom svedčia aj výsledky testov hypotézy o normalite; p-hodnoty Lilieforsovej varianty Kolmogorovho-Smirnovho testu pre obe veličiny väčšie ako $0,2$ a p-hodnoty Shapiro-Wilksovho testu $0,2929$ a $0,3221$. Tieto hodnoty sú väčšie ako daná hladina významnosti $\alpha = 0,01$, a preto nulovú hypotézu o normalite náhodných veličín X, Y nezamietame.

Dvojrozmerný bodový diagram: *Grafy – Bodové grafy – odškrtneme Typ proložení – Proměnné; X, Y; OK – záložka Detaily; Elipsa: normální, Koeficient 0,99 – OK. (Na zobrazenie celej elipsy je potrebné zväčšiť merítko: klikneme pravým tlačitkom na graf – možnosti grafu – osa; merítko a upravíme minimum a maximum podľa potreby)*



Vykreslením dvojrozmerného bodového diagramu vidíme, že všetky body ležia vo vnútri elipsy, čo považujeme za splnený predpoklad dvojrozmernej normality.

Na hladine významnosti 1 % testujeme nulovú hypotézu

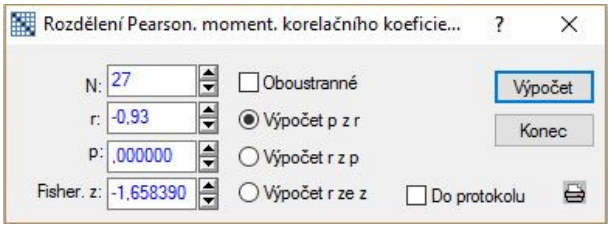
$$H_0 : R = 0 \quad \text{oproti} \quad H_1 : R < 0.$$

So známou hodnotou výberového korelačného koeficienta $r = -0,93$ z výberu rozsahu $n = 27$ spočítame testovaciu štatistiku T:

$$t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}} = \frac{-0,93 \cdot 5}{\sqrt{1-(-0,93)^2}} = -12,65.$$

V štatistických tabuľkách nájdeme kvantil Studentovho rozdelenia pre spoľahlivosť $1-\alpha$ a stupne voľnosti $n-2$, tj. $t_{0,99}(25) = 2,485$ a zostavíme ľavostranný kritický obor $W = (-\infty; -2,485\rangle$. Realizácia testovacej štatistiky $-12,65$ leží v kritickom obore W, a preto na hladine významnosti 1 % zamietame nulovú hypotézu v prospech alternatívnej hypotézy, že medzi veličinami nepriama závislosť skutočne existuje. Znamená to, že s klesajúcou nadmorskou výškou rastie hektárový výnos jačmeňa.

Riešenie (*Výpočet za pomoci systému STATISTICA.*)**.**
Do úvahy zoberieme to, že máme k dispozícii hodnotu výberového korelačného koeficienta a môžeme tak jednoducho spočítať p-hodnotu testu: *Statistiky – Pravdpodobnostní kalkulátor – Korelace; vyplníme N: 27, r: −0,93; odškrtneme Oboustranné a zaškrtneme Výpočet p z r – Výpočet.*



V tabuľke v políčku p sa zobrazila hodnota $0,00000$, čo značí, že je p-hodnota testu veľmi malá a teda menšia ako hladina významnosti $0,01$, podľa čoho zamietame na hladine významnosti 1 % nulovú hypotézu v prospech alternatívnej hypotézy.

Poznámka. Výpočet p-hodnoty jednostranného testu z p-hodnoty obojstranného testu ozn. p prebieha v závislosti od znamienka výberového korelačného koeficienta. Ak je $r < 0$, potom sa p-hodnota ľavostranného testu rovná $p/2$, resp. $1 - p/2$ pre pravostranný test. Ak je $r > 0$, potom sú p-hodnoty jednostranných testov presne opačné, tj. $1 - p/2$ pre ľavovostranný test a $p/2$ pre pravostranný test. V našom prípade sa teda jedná o p-hodnotu ľavostranného testu.

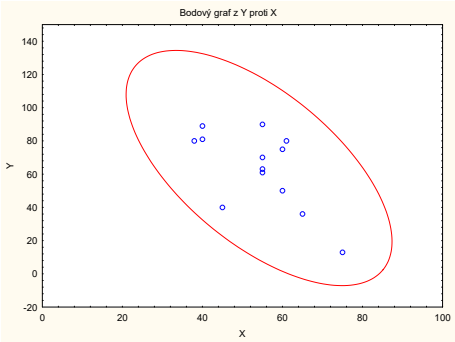
Ďalšou možnosťou riešenia je výpočet z tabuľky dát, kedy nám STATISTICA poskytne, okrem iného, aj hodnotu testovacej štatistiky. Tento spôsob si ukážeme v nasledujúcom príklade.

Príklad 2 (Test významnosti korelačného koeficienta, 3.2.)**.**
Z uvedených údajov vyjadrite závislosť požadovaného množstva výrobku (v kusoch), ktoré je spotrebiteľ ochotný nakúpiť pri daných cenách (v korunách) a pomocou vhodného testu overte, či sa jedná o štatisticky významnú závislosť. [10, str. 293]

Požadované množstvo	*45*	*55*	*38*	*40*	*40*	*55*	*60*	*60*	*75*	*65*	*55*	*55*	*61*
Cena	*40*	*63*	*80*	*89*	*81*	*61*	*50*	*75*	*13*	*36*	*70*	*90*	*80*

Riešenie (*Ručný výpočet.*)**.**
Postup overenia dvojrozmernej normality je analogický ako v predchádzajúcom príklade, naďalej sa ale obmedzíme na testy normality a dvojrozmerný bodový diagram vynechaním normálne pravdepodobnostných grafov.

	Testy normality				
Proměnná	N	max D	Lilliefors p	W	p
X	13	0,2234	p <,10	0,9314	0,3558
Y	13	0,1490	p >.20	0,9140	0,2078



Pretože sú p-hodnoty oboch testov jednorozmernej normality väčšie ako hladina významnosti $0,05$ a v dvojrozmernom bodovom diagrame vypĺňajú body vnútro elipsy, hodnotíme predpoklad dvojrozmernej normality za oprávnený.

Označme požadované množstvo ako veličinu X a cenu ako Y. K dispozícii máme výber rozsahu $n = 13$. Závislosť medzi veličinami vyjadríme pomocou výberového korelačného koefcienta R_{12}. Na jeho výpočet spočítame najprv realizácie výberových priemerov M_1 a M_2, výberových smerodajných odchýliek S_1 a S_2 a výberovej kovariancie S_{12}:

$$m_1 = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{1}{13}(45+55+\cdots+61) = \frac{704}{13} = 54,154,$$

$$m_2 = \frac{1}{n}\sum_{i=1}^{n} y_i = \frac{1}{13}(40+63+\cdots+80) = \frac{828}{13} = 63,692,$$

$$s_1 = \sqrt{\frac{1}{n-1}\left(\sum_{i=1}^{n} x_i^2 - nm_1^2\right)} = \sqrt{\frac{1}{12}\left(45^2+55^2+\cdots+61^2 - 13\frac{704^2}{13^2}\right)} = 10,862,$$

$$s_2 = \sqrt{\frac{1}{n-1}\left(\sum_{i=1}^{n} y_i^2 - nm_2^2\right)} = \sqrt{\frac{1}{12}\left(40^2+63^2+\cdots+80^2 - 13\frac{828^2}{13^2}\right)} = 23,139,$$

$$s_{12} = \frac{1}{n-1}\left(\sum_{i=1}^{n} x_i y_i - nm_1m_2\right) = \frac{1}{12}\left(45\cdot 40+55\cdot 63+\cdots+61\cdot 80 - 13\frac{704}{13}\frac{828}{13}\right),$$

$s_{12} = -157,032.$

Dosadením konkrétnych hodnôt a jednoduchým výpočtom dostávame

$$r = \frac{s_{12}}{s_1 s_2} = \frac{-1507,032}{10,862 \cdot 23,139} = -0,625.$$

Hodnota výberového korelačného koeficienta svedčí o stredne silnej, nepriamej závislosti medzi požadovaným množstvom a cenou výrobku. To znamená, že s rastom ceny výrobku dochádza k poklesu jeho požadovaného množstva.

Tvrdenie o štatistickej významnosti danej závislosti overíme pomocou testu o korelačnom koeficiente. Testujeme nulovú hypotézu

$$H_0 : R = 0 \quad \text{oproti} \quad H_1 : R \neq 0$$

za pomoci testovacej štatistiky T; $t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}} \frac{-0,625\sqrt{11}}{\sqrt{1-(-0,625)^2}} = -2,654.$
Kritický obor tvaru $W = \left(-\infty; -t_{0,975}(11)\right\rangle \cup \left\langle t_{0,975}(11); \infty\right) = (-\infty; -2,201\rangle \cup \langle 2,201; \infty)$ môžeme zjednodušiť porovnaním absolútnej hodnoty testovacej štatistiky s kvantilom Studentovho rozdelenia, tj. $|t| \geq t_{0,975}(11)$. Pretože platí $|-2,654| > 2,201$ ($\Rightarrow t \in W$), zamietame nulovú hypotézu na hladine významnosti 5 % a preukázali sme, že medzi požadovaným množstvom a cenou výrobku existuje štatisticky významná závislosť.

Riešenie (*Výpočet za pomoci systému STATISTICA.*)**.**
Vytvoríme dátový súbor s 2 premennými a 13 prípadmi. Pomenujeme ich X (požadované množstvo), Y (cena) a do tabuľky zapíšeme odpovedajúce hodnoty.

	1 X	2 Y
1	45	40
2	55	63
3	38	80
4	40	89
5	40	81
6	55	61
7	60	50
8	60	75
9	75	13
10	65	36
11	55	70
12	55	90
13	61	80

Postup: *Statistiky – Základní statistiky/tabulky – Korelační matice; OK – 2 seznamy (obd. matice); X, Y; OK – záložka Možnosti; zaklikneme Zobrazit detailní tabulku výsledků – Výpočet.*

	Korelace Označ. korelace jsou významné na hlad. p <,05000 (Celé případy vynechány u ChD)										
Prom. X & prom. Y	Průměr	Sm.Odch.	r(X,Y)	r2	t	p	N	Konst. záv: Y	Směr. záv. Y	Konst. záv: X	Směrnic záv: X
X	54,1538	10,8616									
Y	63,6923	23,1387	-0,6248	0,3904	-2,6542	0,0224	13	135,7748	-1,3311	72,8348	-0,2933

Výberový korelačný koeficient nadobúda hodnotu $-0,6248$ a testovacia štatistika pre test o korelačnom koeficiente hodnotu $-2,6542$. Pri odpovedajúcej p-hodnote testu $0,0224$ a hladine významnosti $\alpha = 0,05$ zamietame nulovú hypotézu o nezávislosti veličín, pretože platí $0,0224 < 0,05$.

Príklad 3 (Test významnosti korelačného koeficienta, 3.2.)**.**
Bol vykonaný hypotetický experiment zameraný na vzťah medzi pracovnou spokojnosťou a pracovným výkonom. Vzorka 42 náhodne vybraných zamestnancov hodnotila svoju vlastnú úroveň spokojnosti s prácou. Táto miera spokojnosti koreluje s hodnotením orgánov dohľadu výkonu, pričom sa dá predpokladať, že sa tieto veličiny riadia dvojrozmerným normálnym rozdelením. Otázkou je, či existuje vzťah medzi týmito dvoma mierami v celej populácii. Pomocu intervalu spoľahlivosti pre korelačný koeficient overte hypotézu o nezávislosti náhodných veličín oproti alternatíve, že sú náhodné veličiny závislé, ak ich výberový korelačný koeficient činí 0,27. [Vlastný zdroj]

Riešenie (*Ručný výpočet.*)**.**
Testujeme nulovú hypotézu

$$H_0 : R = 0 \quad \text{oproti} \quad H_1 : R \neq 0.$$

Poznáme hodnotu koeficientu $r = 0,27$ a taktiež rozsah výberu $n = 42$, ktorý je dostatočne veľký na to, aby sme pri testovaní hypotézy o nezávislosti pomocou intervalu, uvedeného v 3.2 nahradili kvantil Studentovho rozdelenia kvantilom štandardizovaného normálneho rozdelenia, tj. $t_{0,975}(40) \to u_{0,975} = 1,96$. Zostavíme obojstranný interval a pozrieme sa, či obsahuje hodnotu výberového korelačného koeficienta.

$$(D,H) = \left(-\frac{u_{0,975}}{\sqrt{u_{0,975} + n - 2}}, \frac{u_{0,975}}{\sqrt{u_{0,975} + n - 2}} \right) = \left(-\frac{1,96}{\sqrt{1,96 + 40}}, \frac{1,96}{\sqrt{1,96 + 40}} \right)$$
$$= (-0,303; 0,303)$$

Keďže hodnota $0,27 \in (-0,301; 0,301)$, nezamietame na hladine významnosti 5 % nulovú hypotézu o nezávislosti náhodných veličín.

Riešenie (*Výpočet za pomoci systému STATISTICA.*)**.**
Nakoľko nie je výpočet daného intervalu implementovaný v systéme STATISTICA, využijeme ju ako inteligentnú kalkulačku.

Vytvoríme dátový súbor s 2 premennými a 1 prípadom. Prvú premennú pomenujeme D (dolná hranica) a do dlhého mena napíšeme:
$= -VNormal(0,975;0;1)/Sqrt(VNormal(0,975;0;1) + 42 - 2)$.
Druhú premennú pomenujeme H (horná hranica) a do dlhého mena napíšeme:
$= VNormal(0,975;0;1)/Sqrt(VNormal(0,975;0;1) + 42 - 2)$.

	1 D	2 H
1	-0,3026	0,3026

Získali sme hranice intervalu, ktorý má tak tvar $(-0,3026;0,3026)$. Pretože interval pokrýva hodnotu výberového korelačného koeficienta $0,27$, nezamietame nulovú hypotézu na hladine významnosti $5\,\%$.

Príklad 4 (Interval spoľahlivosti pre korelačný koeficient, 3.3.)**.**
Zo základného súboru všetkých pracovníkov v určitej profesii bolo vybraných 80. Zisťujeme závislosť medzi výškou príjmu (v tis. Kč) a dĺžkou praxe (v rokoch). Pre výberový korelačný koeficient platí $r = 0,72$. *Za predpokladu dvojrozmernej normality určite* 95% *obojstranný interval spoľahlivosti pre korelačný koeficient* R. [7, str. 101]

Riešenie (*Ručný výpočet.*)**.**
Je známy rozsah $n = 80$ a koeficient $r = 0,72$. Na konštrukciu intervalu spoľahlivosti pre korelačný koeficient použijeme Fisherovu Z-transformáciu výberového korelačného koeficienta, teda

$$z = \frac{1}{2}\ln\left(\frac{1+r}{1-r}\right) = \frac{1}{2}\ln\left(\frac{1+0,72}{1-0,72}\right) = 0,908.$$

Zo štatistických tabuliek poznáme hodnotu kvantilu štandardizovaného normálneho rozdelenia $u_{0,975} = 1,96$. Spočítame najprv interval spoľahlivosti pre $\frac{1}{2}\ln\left(\frac{1+R}{1-R}\right)$:

$$\begin{aligned} P\left(z - \frac{u_{1-\alpha/2}}{\sqrt{n-3}} < \frac{1}{2}\ln\left(\frac{1+R}{1-R}\right) < z + \frac{u_{1-\alpha/2}}{\sqrt{n-3}}\right) &= 1-\alpha, \\ P\left(0,908 - \frac{u_{0,975}}{\sqrt{77}} < \frac{1}{2}\ln\left(\frac{1+R}{1-R}\right) < 0,908 + \frac{u_{0,975}}{\sqrt{77}}\right) &= 0,95, \\ P\left(0,685 < \frac{1}{2}\ln\left(\frac{1+R}{1-R}\right) < 1,131\right) &= 0,95. \end{aligned}$$

Podľa vzťahu $Z = \operatorname{arctgh} R_{12} \Rightarrow R_{12} = \operatorname{tgh} Z$, pričom $\operatorname{tgh} x = \frac{e^x - e^{-x}}{e^x + e^{-x}}$, prevedieme interval spoľahlivosti do mierky korelačného koeficienta:

$$\begin{aligned} P\left(\operatorname{tgh} 0,685 < R < \operatorname{tgh} 1,131\right) &= 0,95, \\ P\left(\frac{e^{0,685} - e^{-0,685}}{e^{0,685} + e^{-0,685}} < R < \frac{e^{1,131} - e^{-1,131}}{e^{1,131} + e^{-1,131}}\right) &= 0,95. \end{aligned}$$

Po vyčíslení dostávame interval $(0,59;0,81)$, ktorý obsahuje hodnotu korelačného koeficienta so spoľahlivosťou $95\,\%$.

Riešenie (*Výpočet za pomoci systému STATISTICA.*)**.**
1.spôsob: Vytvoríme dátový súbor s 2 premennými a 1 prípadom, ktoré nazveme D a H.
Do dlhého mena premennej D napíšeme:
$= TanH(0,5 * log((1+0,72)/(1-0,72)) - VNormal(0,975;0;1)/sqrt(77))$.
Do dlhého mena premennej H napíšeme:
$= TanH(0,5 * log((1+0,72)/(1-0,72)) + VNormal(0,975;0;1)/sqrt(77))$.

	1 D	2 H
1	0,5943	0,8114

Z výstupnej tabuľky dostávame hranice 95% intervalu spoľahlivosti pre korelačný koeficient. Daný interval je teda v tvare $(0,5943;0,8114)$.

2.spôsob: Využijeme modul „Analýza síly testu“: *Statistiky – Analýza síly testu – Odhad intervalu – Jedna korelace,t-test; OK – Vyplníme Pozorované R: 0,72; Velik.vzorku (N): 80; Spolehlivost: 0,95; zaškrtneme Fisherovo Z (původ.) – Vypočítat.*

	Odhad intervalu Jedna korelace, t-test
	Hodnota
Pozorovaný korel. koef. R	0,7200
Korelace dle nulové hypotézy (Ró0)	0,0000
Oboustranná p-hodnota	0,0000
Velikost vz. ve skup. (N)	80,0000
Interval spolehlivosti	0,9500
Meze spolehlivosti (Fisher. Z původní):	
Ró:	
Dolní mez	0,5943
Horní mez	0,8114

Získavame totožné výsledky, hranice intervalu spoľahlivosti $0,5943$ a $0,8114$.

Príklad 5 (Test hypotézy o danej hodnote korelačného koeficienta, 3.4.)**.**
U 17 rozlične degradovaných vzoriek bavlny bola stanovená relatívna viskozita, a to v roztoku ethylendiaminového komplexu s meďou (CUEN) a v alkalickom roztoku hydroxidu tetraamonmeďnatého (CUOXAN). Z týchto hodnôt boli výpočítané stupne polymerácie DP_1 v roztoku CUEN a DP_2 v roztoku CUOXAN. Predpokladajme, že pokiaľ je korelačný koeficient medzi polymeračnými stupňami nevýznamne menší než hodnota 0,85, existuje medzi výsledkami z oboch rozpúšťadiel významný lineárny vzťah. Testujte hypotézu o danej hodnote korelačného koeficienta oproti alternatíve, že je jeho hodnota menšia, ak sa hodnota výberového korelačného koeficienta rovná 0,614. [13, str. 577]

Riešenie (*Ručný výpočet.*)**.**
Je daná konštanta $c = 0,85$, koeficient $r = 0,614$ a rozsah $n = 17$. Testujeme nulovú hypotézu

$$H_0 : R = 0,85 \quad \text{oproti} \quad H_1 : R < 0,85.$$

Z hodnoty výberového korelačného koeficienta spočítame jeho Fisherovu Z-transformáciu a následne testovaciu štatistiku U:

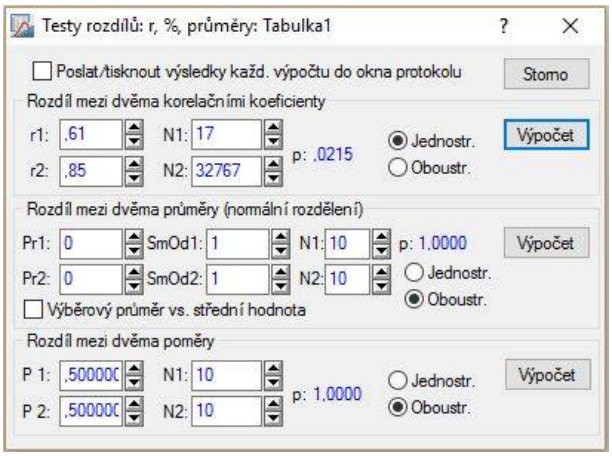
$$z = \frac{1}{2}\ln\left(\frac{1+r}{1-r}\right) = \frac{1}{2}\ln\left(\frac{1+0,614}{1-0,614}\right) = 0,715,$$

$$u = \left(z - \frac{1}{2}\ln\frac{1+c}{1-c} - \frac{c}{2(n-1)}\right)\sqrt{n-3} = \left(0,715 - \frac{1}{2}\ln\frac{1,85}{0,15} - \frac{0,85}{2\cdot 16}\right)\sqrt{14} = -2,124.$$

S kvantilom $u_{0,95} = 1,645$ zostavíme ľavostranný kritický obor $W = (-\infty; -1,645\rangle$. Hodnota testovacej štatistiky $u \in W$, a preto na hladine významnosti 5 % zamietame nulovú hypotézu v prospech alternatívy, čo znamená, že závislosť medzi stupňami polymerácie nie je tak silná, ako sa predpokladalo.

Riešenie (*Výpočet za pomoci systému STATISTICA.*)**.**
Približný výpočet dostaneme postupom: *Statistiky – Základní statistiky/tabulky – Testy rozdílů: r,%,průmery; OK – Rozdíl medzi dvěma korelačními koeficienty; do políčka r1 zapíšeme hodnotu koeficienta 0,614, rozsah 17 do N1, do políčka r2 zapíšeme danú konštantu 0,85 a do N2 maximálnu hodnotu 32767; zaklikneme Jednostr. – Výpočet.*

Poznámka. Políčko N2 by malo obsahovať hodnotu „nekonečno“, v STATISTICE preto nastavujeme maximálnu možnú hodnotu (32767).



Získavame p-hodnotu testu rovnú $0,0215$. Na hladine významnosti 5 % zamietame nulovú hypotézu, pretože platí $0,0215 < 0,05$.

Pre prípad presného výpočtu založenom na testovacej štatistike U sme vytvorili makro *testhodnotykoef.svb*. Zdrojový kód uvádzame v prílohách (str. 36).

Postupujeme: *Soubor – Otevřít; vyberieme testhodnotykoef.svb – klávesou F5 spustíme makro – tabuľka Koeficient r; 0,614; OK – tabuľka Rozsah výberu; 17; OK – tabuľka Konstanta c; 0,85 – tabuľka Hladina významnosti; 0,05; OK – tabuľka Alternativna hypoteza; vyberieme Lavostranna altenativa; OK.*

	Test hypotezy o dane hodnote koeficientu R			
	1 n	2 U	3 u_1-alpha	4 p-value
Zhmuti	17	-2,1230	1,6449	0,0169

Z výstupnej tabuľky dostávame okrem rozsahu predovšetkým hodnotu testovacej štatistiky $-2,1230$, hodnotu kvantilu kritického oboru $1,6449$ a p-hodnotu $0,0169 \rightarrow 0,0169 < 0,05$ a nulovú hypotézu zamietame na hladine významnosti 5 %.

Príklad 6 (Test zhody dvoch korelačných koeficientov, 3.5.)**.**
V psychologickom výskume bolo vyšetrených 12 chlapcov a 15 dievčat. V skupine chlapcov činil výberový korelačný koeficient medzi verbálnou a performačnou zložkou IQ 0,6033, v skupine dievčat činil 0,5833. Za predpokladu dvojrozmernej normality dát testujte hypotézu, že korelačné koeficienty sa nelíšia. [3, str. 232]

Riešenie (*Ručný Výpočet.*)**.**
Rozsah výberu pre chlapcov označíme $n = 12$, pre dievčatá $n^* = 15$ a k nim odpovedajúce výberové korelačné koeficienty $r = 0,6033$ a $r^* = 0,5833$. Testujeme nulovú hypotézu

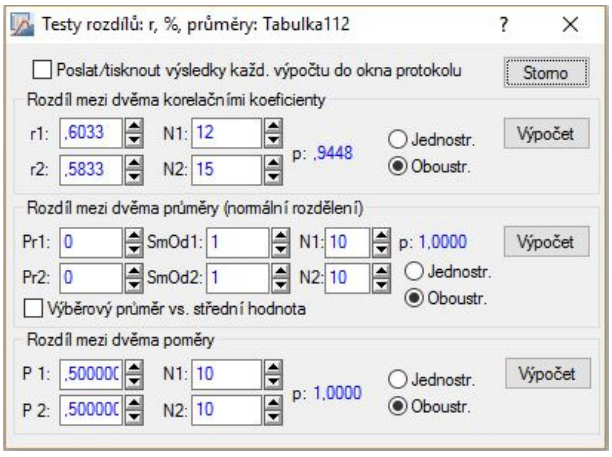
$$H_0 : R = R^* \quad \text{oproti} \quad H_1 : R \neq R^*.$$

Opäť počítame Fisherove Z-transformácie výberových korelačných koeficientov a testovaciu štatistiku U:

$$z = \frac{1}{2}\ln\left(\frac{1+r}{1-r}\right) = \frac{1}{2}\ln\left(\frac{1+0,6033}{1-0,6033}\right) = 0,698,$$
$$z^* = \frac{1}{2}\ln\left(\frac{1+r^*}{1-r^*}\right) = \frac{1}{2}\ln\left(\frac{1+0,5833}{1-0,5833}\right) = 0,667,$$
$$u = \frac{z-z^*}{\sqrt{\frac{1}{n-3}+\frac{1}{n^*-3}}} = \frac{0,698-0,667}{\sqrt{\frac{1}{9}+\frac{1}{12}}} = 0,07.$$

S hodnotou kvantilu štandardizovaného normálneho rozdelenia $u_{0,975} = 1,96$ má kritický obor tvar $W = (-\infty;-1,96\rangle \cup \langle 1,96;\infty)$. Keďže realizácia testovacej štatistiky $0,07 \notin W$, nulovú hypotézu o rovnosti korelačných koeficientov nezamietame na asymptotickej hladine významnosti 5 %.

Riešenie (*Výpočet za pomoci systému STATISTICA.*)**.**
Za pomoci systému spočítame p-hodnotu testu, podľa postupu z predchádzajúceho príkladu: *Statistiky – Základní statistiky/tabulky – Testy rozdílů: r, %, průmery; OK – Rozdíl medzi dvěma korelačními koeficienty; r1=0,6033, N1=12, r2=0,5833 a N2=15; zaklikneme Oboustr. – Výpočet.*



Vzhľadom na vysokú p-hodnotu 0,9448, väčšiu než $\alpha = 0,05$, nezamietame nulovú hypotézu o zhode korelačných koeficientov na asymptotickej hladine významnosti 5 %.

Príklad 7 (Test zhody k korelačných koeficientov, 3.6.)**.**
Lekársky výskum sa zaoberal sledovaním koncentrácií látok A a B v moči pacientov trpiacich obličkovými ochoreniami. U 100 zdravých osôb činil výberový korelačný koeficient medzi koncentráciami oboch látok 0,65. U 142 osôb trpiacich ochorením N činil tento výberový korelačný koeficient 0,37 a u 175 osôb trpiacich ochorením M bol 0,55. Zistite, či je závislosť medzi koncentráciami látok A a B vo všetkých troch skupinách osôb rovnaká. [1, str. 234]

Riešenie (*Ručný výpočet.*)**.**
K dipozícii máme tri výbery rozsahov $n_1 = 100$, $n_2 = 142$, $n_3 = 175$ s výberovými korelačnými koeficientami $r^1 = 0,65$, $r^2 = 0,37$, $r^3 = 0,55$. Na vyriešenie problému použijeme test o zhode k (v našom prípade $k = 3$) korelačných koeficientov:

$$H_0 \colon R_1 = R_2 = R_3 \quad \text{oproti} \quad H_1 \colon \text{„aspoň dva koeficienty sú rozdielne“.}$$

Pre každý z výberov vypočítame Fisherovu Z-transformáciu:

$$z_1 = \frac{1}{2}\ln\left(\frac{1+r^1}{1-r^1}\right) = \frac{1}{2}\ln\left(\frac{1+0,65}{1-0,65}\right) = 0,775,$$
$$z_2 = \frac{1}{2}\ln\left(\frac{1+r^2}{1-r^2}\right) = \frac{1}{2}\ln\left(\frac{1+0,37}{1-0,37}\right) = 0,388,$$
$$z_3 = \frac{1}{2}\ln\left(\frac{1+r^3}{1-r^3}\right) = \frac{1}{2}\ln\left(\frac{1+0,55}{1-0,55}\right) = 0,618.$$

Platí $n = n_1 + n_2 + n_3 = 417$. Spočítame koeficient b a následne testovaciu štatistiku χ^2:

$$b = \frac{1}{n-3k}\sum_{i=1}^{k}(n_i - 3)z_i = \frac{1}{417-9}(97 \cdot 0,775 + 139 \cdot 0,388 + 172 \cdot 0,55) = 0,577,$$

$$\chi^2 = \sum_{i=1}^{k}(n_i - 3)(Z_i - b)^2 = (97 \cdot 0,198^2 + 139 \cdot (-0,189)^2 + 172 \cdot 0,041^2) = 9,052.$$

Pre $\alpha = 0,05$ a kvantil $\chi^2_{0,95}(2) = 5,99$ je kritický obor $W = \langle 5,99; \infty)$. Pretože štatistika $\chi^2 \in W$, na asymptotickej hladine významnosti 5 % zamietame nulovú hypotézu, že závislosť medzi koncentráciami látok A, B je vo všetkých troch skupinách rovnaká. Tukeyovým testom preto zistíme, medzi ktorými skupinami osôb existuje štatisticky významný rozdiel. S tabelovanou hodnotou $q_{k,\infty}(\alpha) = q_{3,\infty}(0,05) = 3,31$ spočítame nerovnosti

$$|z_i - z_j| \geq q_{k,\infty}(\alpha) \cdot \sqrt{\frac{1}{2}\left(\frac{1}{n_i - 3} + \frac{1}{n_j - 3}\right)};$$

$$|0,775 - 0,388| \geq 3,31 \cdot \sqrt{\frac{1}{2}\left(\frac{1}{97} + \frac{1}{139}\right)} \to |0,387| \geq 0,310,$$

$$|0,775 - 0,618| \geq 3,31 \cdot \sqrt{\frac{1}{2}\left(\frac{1}{97} + \frac{1}{172}\right)} \to |0,157| \ngeq 0,297,$$

$$|0,388 - 0,618| \geq 3,31 \cdot \sqrt{\frac{1}{2}\left(\frac{1}{139} + \frac{1}{172}\right)} \to |-0,230| \ngeq 0,267.$$

Nerovnosť platí iba v prvom prípade, čím sme dokázali, že medzi zdravými osobami a osobami trpiacimi chorobou N existuje na hladine $0,05$ štatisticky významný rozdiel, pričom ostatné rozdiely sú štatisticky nevýznamné.

Riešenie (*Výpočet za pomoci systému STATISTICA.*)**.**
Systém STATISTICA nemá implementované testy pre tento typ úlohy, preto sme vytvorili makrá, *testzhodykkoef.svb* a *Tukeytest.svb*, odpovedajúce testu zhody k korelačných koeficientov a Tukeyovmu testu. Zdrojové kódy uvádzame v prílohách (str. 39, 42).

Pred spustením makier vytvoríme dátovú tabuľku s 2 premennými a 3 prípadmi. Premenné nazveme K pre koeficienty, N pre rozsahy a zapíšeme odpovedajúce hodnoty.

	1 K	2 N
1	0,65	100
2	0,37	142
3	0,55	175

Test zhody k korelačných koeficientov: *Soubor – Otevřít; vyberieme testzhodykkoef.svb – klávesou F5 spustíme makro – tabuľka Výber premenných; z prvého zoznamu vyberieme K, z druhého N; OK – tabuľka Hladina významnosti; 0,05; OK.*

	Test shody k korelacnich koeficientu			
	1 k	2 ChiStat	3 Chi 1-alpha (k-1)	4 p-value
Zhmuti	3	9,0518	0,1026	0,0108

Vo výslednej tabuľke nájdeme počet korelačných koeficientov 3, hodnotu testovacej štatistiky $9,0518$, hodnotu kvantilu kritického oboru $5,9915$ a p-hodnotu testu $0,0108$. P-hodnota je menšia ako hladina významnosti $0,05$, podľa čoho zamietame nulovú hypotézu na hladine významnosti $5\,\%$ v prospech alternatívy.

Tukeyov test: *Soubor – Otevřít; vyberieme Tukeytest.svb – klávesou F5 spustíme makro – tabuľka Výber premenných; z prvého zoznamu vyberieme K, z druhého N; OK – tabuľka Tabelovana hodnota; 3,31; OK.*

	Tukeyuv test: Z(i) - Z(j) >= Crit		
i / j	1	2	3
1	0,0000	0,3869	0,1569
2	0,3869	0,0000	0,2300
3	0,1569	0,2300	0,0000

Výsledná tabuľka Tukeyovho testu zobrazuje maticu rozdielov Fisherových Z-transformácií, z ktorých sú zvýraznené tie hodnoty, pre ktoré platí nerovnosť testu. Podľa indexov i, j vidíme, že štatisticky významný rozdiel je medzi Fisherovými transformáciami z_1 a z_2, tj. medzi skupinami zdravých osôb a osôb trpiacimi chorobou N.

Poznámka. Makro *Tukeytest.svb* vytvára taktiež tabuľku kritických hodnôt testu označenú `tabC`, ktorej výpis je potlačený zakomentovaným príkazom `tabC.Visible=True`. V prípade, že chce čitateľ tabuľku zobraziť, stačí v zdrojovom kóde tento príkaz odkomentovať, tj. `'tabC.Visible=True` nahradíme `tabC.Visible=True`.

Príklad 8 (Test významnosti Spearmanovho korelačného koeficienta, 3.7.)**.**
Bolo sledovaných 10 poslucháčov 2. ročníka VŠE. Na základe psychologického vyšetrenia boli títo poslucháči zoradení podľa nervovej lability (čím bol poslucháč labilnejší, tým dostal vyššie poradie R_i). Okrem toho sledovaní poslucháči dostali poradie Q_i na základe výsledkov v štatistike (najlepší poslucháč dostal poradie 1). Výsledky sú uvedené v tabuľke. Zistite, či je nervová labilita nezávislá od výsledkov v štatistike. [8, str. 90]

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

Riešenie (*Ručný výpočet.*)**.**
Keďže máme k dispozícii poradia 10 poslucháčov, budeme testovať poradovú nezávislosť s nulovou hypotézou

$$H_0 : R_s = 0 \quad \text{oproti} \quad H_1 : R_s \neq 0$$

a ako testovacie kritérium použijeme Spearmanov korelačný koeficient r_s. Na jeho výpočet určíme najprv rozdiely $R_i - Q_i$:

$R_i - Q_i$	-8	-1	-5	-1	1	4	-3	7	2	4 ,

$$r_s = 1 - \frac{6}{n(n^2-1)} \sum_{i=1}^{n} (R_i - Q_i)^2 = 1 - \frac{6}{10(10^2-1)} ((-8)^2 + (-1)^2 + \cdots + 4^2) = -0,127.$$

Pri hladine významnosti $\alpha = 0,05$, rozsahu $n = 10$ a tabuľkovej hodnote $r_{s,0,975}(10) = 0,6364$ má kritický obor tvar $W = \langle -1; -0,6364 \rangle \cup \langle 0,6364; 1 \rangle$. Hodnota Spearmanovho koeficienta korelácie $-0,127 \notin W$, a preto nulovú hypotézu o (poradovej) nezávislosti nervovej lability poslucháčov a výsledkov v štatistike nezamietame na hladine významnosti 5 %.

Riešenie (*Výpočet za pomoci systému STATISTICA.*)**.**
Vytvoríme dátový súbor s 2 premennými R_i, Q_i a 10 prípadmi, kde zapíšeme odpovedajúce poradia.

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

Spearmanov koeficient korelácie: *Statistiky – Neparametrická statistika – Korelace (Spearman, . . .); OK – Proměnné; Ri, Qi; OK – Spearman.R.*

	Spearmanovy korelace ChD vynechány párově Označ. korelace jsou významné na hl. p <,05000	
Proměnná	R_i	Q_i
R_i	1,0000	-0,1273
Q_i	-0,1273	1,0000

Hodnotu Spearmanovho koeficienta korelácie porovnáme s kritickou tabelovanou hodnotou $r_{s,0,975}(10) = 0,6364 \rightarrow |-0,1273| < 0,6364 \rightarrow$ nulovú hypotézu o (poradovej) nezávislosti nezamietame.

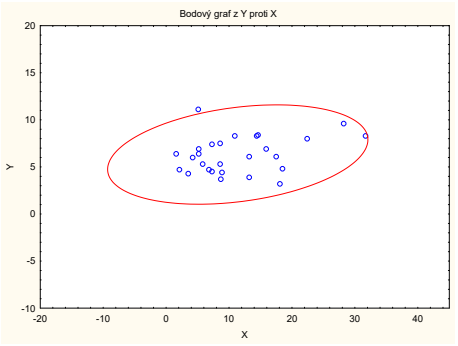
Príklad 9 (Test významnosti Spearmanovho korelačného koeficienta, 3.7.)**.**
V tabuľke je zaznamenaný percentuálny podiel ľudí s pracovnou zmluvou na dobu určitú (veličina X) a percentuálny podiel nezamestnaných ľudí (veličina Y) v krajinách Európskej únie. Pomocou vhodného koeficienta určite a otestujte závislosť medzi pozorovanými veličinami X, Y. [11, str. 400]

Štát	x_i	y_i
Belgicko	*8,6*	*7,5*
Bulharsko	*5,2*	*6,9*
Česká republika	*8,6*	*5,3*
Dánsko	*8,7*	*3,7*
Nemecko	*14,6*	*8,4*
Estónsko	*2,1*	*4,7*
Írsko	*7,3*	*4,5*
Grécko	*10,9*	*8,3*
Španielsko	*31,7*	*8,3*
Francúzsko	*14,4*	*8,3*
Taliansko	*13,2*	*6,1*
Cyprus	*13,2*	*3,9*
Lotyšsko	*4,2*	*6*
Litva	*3,5*	*4,3*
Luxembursko	*6,8*	*4,7*
Maďarsko	*7,3*	*7,4*
Malta	*5,2*	*6,4*
Holandsko	*18,1*	*3,2*
Rakúsko	*8,9*	*4,4*
Poľsko	*28,2*	*9,6*
Portugalsko	*22,4*	*8*
Rumunsko	*1,6*	*6,4*
Slovinsko	*18,5*	*4,8*
Slovensko	*5,10*	*11,10*
Fínsko	*15,9*	*6,9*
Švédsko	*17,5*	*6,1*
Anglicko	*5,8*	*5,3*

Riešenie (*Ručný výpočet.*)**.**
V tomto prípade typ dát nerozhoduje o postupe riešenia úlohy, rozhodneme tak ale podľa splneného resp. nesplneného predpokladu dvojrozmernej normality.

Podľa tabuľky testov jednorozmernej normality zamietame normalitu náhodnej veličiny X. Porušenie normality naznačuje aj dvojrozmerný bodový diagram, a preto v rámci korektnosti použijeme Spearmanov korelačný koeficient, aj keď by takéto mierne odchýlenie od normality významne neovplynilo výsledok pri použití Pearsonovho (výberového) korelačného koeficienta.

Proměnná	Testy normality N	max D	Lilliefors p	W	p
X	27	0,1830	p <,05	0,9093	0,0219
Y	27	0,1144	p >.20	0,9638	0,4490



Spočítame teda poradovú závislosť. Kedže máme k dispozícii realizácie náhodných veličín, určíme najprv ich poradové čísla (najnižšia hodnota má poradové číslo 1). Pre prehľadnosť a uľahčenie výpočtu Spearmanovho koeficienta r_s zostavíme najskôr tabuľku obsahujúcu realizácie náhodných veličín, ich poradia a druhú mocninu rozdielu týchto poradí.

x_i	y_i	R_i	Q_i	$(R_i - Q_i)^2$
8,6	7,5	12,5	20	56,3
5,2	6,9	6,5	17,5	121
8,6	5,3	12,5	10,5	4
8,7	3,7	14	2	144
14,6	8,4	20	25	25
2,1	4,7	2	7,5	30,3
7,3	4,5	10,5	6	20,3
10,9	8,3	16	23	49
31,7	8,3	27	23	16
14,4	8,3	19	23	16
13,2	6,1	17,5	13,5	16
13,2	3,9	17,5	3	210,3
4,2	6	4	12	64
3,5	4,3	3	4	1
6,8	4,7	9	7,5	2,3
7,3	7,4	10,5	19	72,3
5,2	6,4	6,5	15,5	81
18,1	3,2	23	1	484
8,9	4,4	15	5	100
28,2	9,6	26	26	0
22,4	8	25	21	16
1,6	6,4	1	15,5	210
18,5	4,8	24	9	225
5,1	11,1	5	27	484
15,9	6,9	21	17,5	12,3
17,5	6,1	22	13,5	72,3
5,8	5,3	8	10,5	6,3

$$r_s = 1 - \frac{6}{n(n^2-1)} \sum_{i=1}^{n} (R_i - Q_i)^2 = 1 - \frac{6}{27(27^2-1)} (56,3 + 121 + \cdots + 6,3) = 0,225.$$

Táto nízka hodnota koeficienta $0,225$ naznačuje pomerne slabú poradovú závislosť medzi veličinami. V nasledujúcom kroku otestujeme, či skutočne existuje.

Staviame nulovú hypotézu

$$H_0: R_s = 0 \quad \text{oproti} \quad H_1: R_s \neq 0.$$

Vzhľadom na veľkosť rozsahu výberu $n = 27 > 20$ použijeme testovaciu štatistiku T:

$$t = \frac{r_s\sqrt{n-2}}{\sqrt{1-r_s^2}} = \frac{0,225 \cdot 5}{\sqrt{1-0,0506}} = 1,155.$$

V tabuľkách nájdeme hodnotu kvantilu Studentovho rozdelenia $t_{0,975}(25) = 2,0595$ a zostrojíme kritický obor pre daný test $W = (-\infty; -2,0595\rangle \cup \langle 2,0595; \infty)$. Hodnota testovacej štatistiky $1,155$ neleží v kritickom obore W, a preto na hladine významnosti $5\,\%$ nezamietame nulovú hypotézu. Nepreukázali sme, že medzi veličinami existuje závislosť.

Riešenie (*Výpočet za pomoci systému STATISTICA.*)**.**
Postup riešenia je takmer identický ako v minulom príklade. Vytvoríme dátový súbor s 2 premennými a 27 prípadmi určujúcimi realizácie veličín X, Y. Nie je potrebné zadávať poradové čísla, tie si STATISTICA pri výpočte určí automaticky (Z dôvodu rozsiahleho počtu pozorovaní tabuľku neuvádzame.).

Postup: *Statistiky – Neparametrická statistika – Korelace (Spearman, . . .);OK –Vytvoriť; Detailní report – Proměnné; X, Y; OK – Spearman.R.*

	Spearmanovy korelace ChD vynechány párově Označ. korelace jsou významné na hl. p <,05000			
Dvojice proměnných	Počet plat.	Spearman R	t(N-2)	p-hodn.
X & Y	27	0,223582	1,146943	0,262264

Z výstupnej tabuľky nás predovšetkým zaujíma p-hodnota testu rovná $0,2623$, čo je hodnota väčšia ako $\alpha = 0,05$, a preto nulovú hypotézu nezamietame na hladine významnosti $5\,\%$.

Záver

Pri tvorbe tejto bakalárskej práce sme sa snažili oboznámiť čitateľa so základnou teóriou skúmania závislosti intervalových, pomerových či ordinálnych veličín v rámci jednoduchej korelačnej analýzy. Tieto teoretické poznatky sme ukážkovo aplikovali na 9 príkladoch pokrývajúcich testy, resp. úlohy, ktoré jednoduchá korelačná analýza rieši. V spolupráci s Centrom výpočetnej techniky na Fakulte informatiky sme vytvorili interaktívny študijný materiál s cieľom lepšie sprístupniť a uľahčiť štúdium zahŕňajúce danú tému. V rámci tohto projektu bude mať študent možnosť využívať videotutoriály, ktoré poskytujú postupy riešení príkladov v systéme STATISTICA alebo možnosť vyskúšať si tieto postupy samostatne za pomoci interaktívnych tutoriálov, kde bude po celú dobu vedený ako postupovať.

Prílohy

Test hypotézy o danej hodnote korelačného koeficienta – zdrojový kód

```
Option Base 1

Sub Main

Dim alpha As Double   'hladina vyznamnosti
Dim p As Double       'p hodnota
Dim r As Double       'vyberovy korelacni koeficient
Dim n As Double       'rozsah
Dim c As Double       'konstanta
Dim Z As Double       'Fisherova transformace
Dim U As Double       'testovaci statistika

Coef=InputBox("Zadejte hodnotu vyberoveho korelacniho koeficientu:",
"Koeficient r")
r=CDbl(Coef)

Range=InputBox("Zadejte rozsah vyberu:","Rozsah vyberu")
n=CDbl(Range)

LevOfSign=InputBox("Zadejte hladinu vyznamnosti, na ktere bude
hypoteza testovana:","Hladina vyznamnosti")
alpha=CDbl(LevOfSign)

'Vyber alternativni hypotezy
alt = DisplayListBox("Alternativni hypoteza","Oboustranna
alternativa|Levostranna alternativa|Pravostranna alternativa",1)

'Fisherova transformace
Z = 0.5*Log((1+r)/(1-r))

'testovaci statistika
U = (Z-0.5*Log((1+c)/(1-c))-c/(2*(n-1)))*Sqrt(n-3)
```

```
'kvantily
If alt = 0 Then End
If alt = 1 Then
   u_1 = -VNormal(1-alpha/2,0,1)
   u_2 = VNormal(1-alpha/2,0,1)

   p1 = INormal(U,0,1)    'vypocet p-hodnoty
   p2 = 1-INormal(U,0,1)
   If (pmen <= pvac) Then
     p = 2 * p1
   Else
     p = 2 * p2
   End If

'Vysledni tabulka oboustranne alternativy
Set Summary = Spreadsheets.New
Summary.SetSize(1,4)
Summary.Header="Test hypotezy o dane hodnote koeficientu R"
Summary.CaseName(1) = "Zhrnuti"
Summary.AutoFitCase
Summary.VariableName(1) = "n"
Summary.VariableName(2) = "U"
Summary.VariableName(3) = "u_1-alpha/2"
Summary.VariableName(4) = "p-value"

Summary.VariableFormatString(2) = "0.0000"
Summary.VariableFormatString(3) = "0.0000"
Summary.VariableFormatString(4) = "0.0000"

Summary.Variable(1).ColumnWidth=0.4
Summary.Value(1,1) = n
Summary.Variable(2).ColumnWidth=0.7
Summary.Value(1,2) =
U~Summary.Variable(3).ColumnWidth=0.9
Summary.Value(1,3) = u_2
Summary.Variable(4).ColumnWidth=0.7
Summary.Value(1,4) = p

'Kdyz je p-hodnota mensi nebo rovna nez alpha, zobrazi se cervene.
   If (p <= alpha) Then
      Summary.Cells(1,4).Font.Color = RGB(255,0,0)
      End If
      Summary.Visible=True
End If
```

```
If alt = 2 Then
   u_norm = VNormal(1-alpha,0,1)
   p = INormal(U,0,1) 'p-hodnota

'Vysledni tabulka levostranne alternativy
Set Summary = Spreadsheets.New
Summary.SetSize(1,4)
Summary.Header="Test hypotezy o dane hodnote koeficientu R"
Summary.CaseName(1) = "Zhrnuti"
Summary.AutoFitCase
Summary.VariableName(1) = "n"
Summary.VariableName(2) = "U"
Summary.VariableName(3) = "u_1-alpha"
Summary.VariableName(4) = "p-value"

Summary.VariableFormatString(2) = "0.0000"
Summary.VariableFormatString(3) = "0.0000"
Summary.VariableFormatString(4) = "0.0000"

Summary.Variable(1).ColumnWidth=0.4
Summary.Value(1,1) = n
Summary.Variable(2).ColumnWidth=0.7
Summary.Value(1,2) =
U~Summary.Variable(3).ColumnWidth=0.7
Summary.Value(1,3) = u_norm
Summary.Variable(4).ColumnWidth=0.7
Summary.Value(1,4) = p

'Kdyz je p-hodnota mensi nebo rovna nez alpha, zobrazi se cervene.
   If (p <= alpha) Then
      Summary.Cells(1,4).Font.Color = RGB(255,0,0)
   End If
   Summary.Visible=True
End If

If alt = 3 Then
   u_norm = -VNormal(1-alpha,0,1)
   p = 1 - INormal(U,0,1) 'p-hodnota

'Vysledni tabulka pravostranne alternativy
Set Summary = Spreadsheets.New
Summary.SetSize(1,4)
Summary.Header="Test hypotezy o dane hodnote koeficientu R"
Summary.CaseName(1) = "Zhrnuti"
```

```
Summary.AutoFitCase
Summary.VariableName(1) = "n"
Summary.VariableName(2) = "U"
Summary.VariableName(3) = "u_1-alpha"
Summary.VariableName(4) = "p-value"

Summary.VariableFormatString(2) = "0.0000"
Summary.VariableFormatString(3) = "0.0000"
Summary.VariableFormatString(4) = "0.0000"

Summary.Variable(1).ColumnWidth=0.4
Summary.Value(1,1) = n
Summary.Variable(2).ColumnWidth=0.7
Summary.Value(1,2) =
U~Summary.Variable(3).ColumnWidth=0.7
Summary.Value(1,3) = u_norm
Summary.Variable(4).ColumnWidth=0.7
Summary.Value(1,4) = p

'Kdyz je p-hodnota mensi nebo rovna nez alpha, zobrazi se cervene.
   If (p <= alpha) Then
      Summary.Cells(1,4).Font.Color = RGB(255,0,0)
   End If
   Summary.Visible=True
End If

End Sub
```

Test zhody *k* korelačných koeficientov – zdrojový kód

```
Option Base 1

Sub Main

Dim alpha As Double        'hladina vyznamnosti
Dim p As Double            'p hodnota
Dim b As Double            'koeficient b
Dim Chi As Double          'testovaci statistika
Dim chiKvantil As Double   'kvantil
Dim VarList () As Long     'seznam vybranych promennych
Dim Matrix() As Double     'datova matice
Dim suma As Double         'soucet rozsahu
Dim Z() As Double          'Fisherova transformace
```

```
numvar=ActiveSpreadsheet.NumberOfVariables
numcas=ActiveSpreadsheet.NumberOfCases

'Volba promennych
ReDim VarList(1 To numvar)

If 0=SelectVariables2(ActiveDataSet,"Vyber promennych",1,1,Varlist(1)
    Count,"Koeficienty",1,1,Varlist(2),Count,"Rozsahy") Then
    End
End If

s1 = Varlist(1)
s2 = Varlist(2)

If numcas < 3 Then
    MsgBox("Prilis malo koeficientu.","Chyba")
    End
End If

'soucet rozsahu
ReDim Preserve Matrix(numcas,numvar) As Double
Matrix = ActiveSpreadsheet.Data

For i = 1 To numcas
    suma = suma + Matrix(i,s2)
Next i

'Fisherova transformace
ReDim Z(1 To numcas)

For i = 1 To numcas
    Z(i) = 0.5*Log((1+Matrix(i,s1))/(1-Matrix(i,s1)))
Next i

'b
For i = 1 To numcas
    b = b + (Matrix(i,s2)-3)*Z(i)
Next i
b = b/(suma-3*numcas)

'statistika
For i = 1 To numcas
    Chi = Chi + (Matrix(i,s2)-3)*(Z(i)-b)^2
Next i
```

```
'p hodnota
p =1- IChi2(Chi, numcas-1)

'alpha
LevOfSign=InputBox("Zadejte hladinu vvznamnosti, na ktere
bude hypoteza testovana:","Hladina vyznamnosti")
alpha=CDbl(LevOfSign)

'kvantil
chiKvantil = VChi2(alpha, numcas-1)

'vysledni tabulka
Set Summary = Spreadsheets.New
Summary.SetSize(1,4)
Summary.Header="Test zhody k korelacnich koeficientu"
Summary.CaseName(1) = "Zhrnuti"
Summary.AutoFitCase
Summary.VariableName(1) = "k"
Summary.VariableName(2) = "ChiStat"
Summary.VariableName(3) = "Chi_1-alpha (k-1)"
Summary.VariableName(4) = "p-value"

Summary.VariableFormatString(2) = "0.0000"
Summary.VariableFormatString(3) = "0.0000"
Summary.VariableFormatString(4) = "0.0000"

Summary.Variable(1).ColumnWidth=0.4
Summary.Value(1,1) = numcas
Summary.Variable(2).ColumnWidth=0.7
Summary.Value(1,2) = Chi
Summary.Variable(3).ColumnWidth=1.2
Summary.Value(1,3) = chiKvantil
Summary.Variable(4).ColumnWidth=0.7
Summary.Value(1,4) = p

'Kdyz je p-hodnota mensi nebo rovna nez alpha, zobrazi se cervene.
If (p <= alpha) Then
   Summary.Cells(1,4).Font.Color = RGB(255,0,0)
End If
Summary.Visible=True

End Sub
```

Tukeyov test – zdrojový kód

```
Option Base 1

Sub Main

Dim VarList() As Long      'seznam vybranych promennych
Dim Z() As Double          'Fisherova transformace
Dim DZ() As Double         'rozdil Fisher transformaci
Dim C() As Double          'kriticka hodnota
Dim Matrix() As Double     'datova matice
Dim MatrixDif() As Double  'matice rozdilu Fisher transformaci
Dim MatrixCrit() As Double 'matice kritickych hodnot

numvar=ActiveSpreadsheet.NumberOfVariables
numcas=ActiveSpreadsheet.NumberOfCases

'Volba promennych
ReDim VarList(1 To numvar)

If 0=SelectVariables2(ActiveDataSet,"Vyber promennych"1,1,VarList(1),
   Count,"Koeficienty",1,1,VarList(2),Count,"Rozsahy") Then
   End
End If

s1 = VarList(1)
s2 = VarList(2)

'Fisherova transformace
ReDim Preserve Matrix(numcas,numvar) As Double
Matrix = ActiveSpreadsheet.Data
ReDim Z(1 To numcas)

For i = 1 To numcas
    Z(i) = 0.5*Log((1+Matrix(i,s1))/(1-Matrix(i,s1)))
Next i

'matice rozdilu Fisher transformaci
ReDim MatrixDif(numcas,numcas)
ReDim DZ(1 To numcas)

For i= 1 To numcas
    For j = 1 To numcas
        DZ(i)= Abs(Z(i)-Z(j))
        MatrixDif(i,j)=DZ(i)
```

```
    Next j
Next i

'matice kritickych hodnot
tabValue=InputBox("Zadajte tabelovanu hodnotu q_(k,inf)(alpha):",
"Tabelovana hodnota")
tv=CDbl(tabValue)

ReDim MatrixCrit(numcas,numcas)
ReDim C(1 To numcas)

For i= 1 To numcas
    For j= 1 To numcas
        C(i) = tv*Sqrt(0.5*(1/(Matrix(i,s2)-3)+ 1/(Matrix(j,s2)-3)))
        MatrixCrit(i,j)=C(i)
    Next j
Next i

'Vysledni tabulka kritickych hodnot
Dim tabC As New Spreadsheet
tabC.SetSize(numcas,numcas)
tabC.Header="Tukeyuv test: tabulka Crit"
tabC.Infobox.VerticalAlignment = 1
tabC.Infobox.HorizontalAlignment = 1
For i= 1 To numcas
    tabC.VariableName(i)=""
    tabC.VariableFormatString(i)="0.0000"
Next i

For i= 1 To numcas
    For j= 1 To numcas
        tabC.Cells(i,j)=MatrixCrit(i,j)
        tabC.Cells(i,i)= ""
    Next j
Next i
'tabC.Visible=True
'zobrazeni tabulky v pripade, ze chceme videt vypoctene hodnoty

'Vysledni tabulka rozdilu Fisher transformaci
Dim tabD As New Spreadsheet
tabD.SetSize(numcas,numcas)
tabD.Header="Tukeyov test:  Z(i) - Z(j) >= Crit"
tabD.InfoBox="  i / j "
tabD.InfoBox.VerticalAlignment = 1
tabD.InfoBox.HorizontalAlignment = 1
```

```
For i= 1 To numcas
    tabD.VariableName(i) = ""
    tabD.VariableFormatString(i) = "0.0000"
Next i

For i= 1 To numcas
    For j= 1 To numcas
        tabD.Cells(i,j)=MatrixDif(i,j)
    Next j
Next i

'Rozdily splnujuci nerovnost testu oznaci cervene
For i= 1 To numcas
    For j= 1 To numcas
        If tabD.Cells(i,j) >= MatrixCrit(i,j) Then
           tabD.Cells(i,j).Font.Color = RGB(255,0,0)
        End If
   Next j
Next i
tabD.Visible=True

End Sub
```

Zoznam použitej literatúry

[1] ANDĚL, Jiří. 2007. *Statistické metody*. 4., upr. vyd. Praha: Matfyzpress. ISBN 8073780038.

[2] BÍLKOVÁ, Diana, Petr BUDINSKÝ a Václav VOHÁNKA. 2009. *Pravděpodobnost a statistika*. Plzeň: Vydavatelství a nakladatelství Aleš Čeněk. ISBN 9788073802240.

[3] BUDÍKOVÁ, Marie, Maria KRÁLOVÁ a Bohumil MAROŠ. 2010. *Průvodce základními statistickými metodami*. Praha: Grada. Expert (Grada). ISBN 9788024732435.

[4] BUDÍKOVÁ, Marie, Tomáš LERCH a Štěpán MIKOLÁŠ. 2005. *Základní statistické metody*. Brno: Masarykova univerzita v Brně. ISBN 8021038861.

[5] HENDL, Jan. 2004. *Přehled statistických metod zpracování dat: analýza a metaanalýza dat*. Praha: Portál. ISBN 8071788201.

[6] HINDLS, Richard, Stanislava HRONOVÁ a Jan SEGER. c2004. *Statistika pro ekonomy*. 5. vyd. Praha: Professional Publishing. ISBN 8086419592.

[7] JEŘÁBEK, Tomáš a Jana ŠTOFILOVÁ. 2012. *Statistika v hotelnictví a cestovním ruchu I*. Brno: Vysoká škola obchodní a hotelová. ISBN 9788087300343.

[8] KLÍMEK, Petr. 2003. *Aplikovaná statistika pro ekonomy*. Zlín: Univerzita Tomáše Bati ve Zlíně. ISBN 8073181487.

[9] KOŽÍŠEK, Jan. 2002. *Statistická analýza*. Vyd. 4. přeprac. Praha: Vydavatelství ČVUT. ISBN 8001024962.

[10] MAREK, Luboš. 2015. *Statistika v příkladech*. Druhé vydání. Praha: Kamil Mařík - Professional Publishing. ISBN 9788074311536.

[11] MARKECHOVÁ, Diana, Beata STEHLÍKOVÁ a Anna TIRPÁKOVÁ. 2011. *Štatistické metódy a ich aplikácie*. Nitra: UKF. ISBN 9788080948078. Dostupné také z: http://www.km.fpv.ukf.sk/upload_publikacie/20120125_143707__1.pdf

[12] MELOUN, Milan a Jiří MILITKÝ. 2012. *Kompendium statistického zpracování dat*. Praha: Karolinum. ISBN 9788024621968.

[13] MELOUN, Milan, 1994. *Statistické zpracování experimentálních dat v chemometrii, biometrii, ekonometrii a v dalších oborech přírodních, technických a společenských věd*. Praha: Plus. Plus (Plus). ISBN 8085297566.

[14] StatSoft, Inc. (2013). STATISTICA (data analysis software system), version 12. www.statistica.io