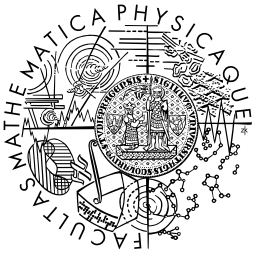


MATEMATICKO-FYZIKÁLNÍ
FAKULTA
Univerzita Karlova

BAKALÁRSKA PRÁCA

Roman Lacuš

Korelačné koeficienty a ich použitie

Katedra pravděpodobnosti a matematické statistiky

Vedúci bakalárskej práce:	doc. Ing. Marek Omelka, Ph.D.
Štúdijný program:	Matematika
Štúdijný odbor:	Finančná matematika

Praha 2022

Prehlasujem,že som túto bakalársku prácu vypracoval(a) samostatne a výhradne s použitím citovaných prameňov, literatúry a ďalších odborných zdrojov.

Beriem na vedomie, že sa na moju prácu vzťahujú práva a povinnosti vyplývajúce zo zákona č 121/2000Sb., autorského zákona v platnom znení, najmä skutočnosť, že Univerzita Karlova má právo na uzavretie licenčnej zmluvy o použití tejto práce ako školského diela podľa §60 odst. 1 autorského zákona.

V ..... dňa ..... Podpis autora

Chcel by som sa poďakovať môjmu vedúcemu bakalárskej práce doc. Ing. Marek Omelkovi, Ph.D. za vedenie práce a poskytnutie vecných rád, ktoré som aplikoval v tejto práci.

Názov práce: Korelačné koeficienty a ich použitie

Autor: Roman Lacuš

Katedra: Katedra pravděpodobnosti a matematické statistiky

Vedúci bakalárskej práce: doc. Ing. Marek Omelka, Ph.D., Katedra pravděpodobnosti a matematické statistiky

Abstrakt: Cieľom bakalárskej práce je teoretické oboznámenie sa s korelačnými koeficientmi a to konkrétne Pearsonovým, Spearmanovým a Kendallovým, a s ich štatistickými odhadmi a potom ich následné použitie v príkladoch a diskusia nad vhodnosťou a nevhodnosťou použitia jednotlivých korelačných koeficientov a simulačné štúdie za účelom porovnania sily testu hypotézy nezávislosti.

Kľúčové slová: korelačné koeficienty, pravdepodobnosť, štatistika

Title: Correlation coefficients and their use

Author: Roman Lacuš

Department: Department of Probability and Mathematical Statistics

Supervisor: doc. Ing. Marek Omelka, Ph.D., Department of Probability and Mathematical Statistics

Abstract: The aim of this bachelor thesis is to introduce the correlation coefficients especially Pearson's , Spearman's and Kendall's correlation coefficient and their statistic estimations following with application on data and discussion on validity of their usage and simulation studies with intention to compare the power of tests of independence.

Keywords: correlation coefficients, probability, statistics

Obsah

Úvod

Bežne sa stretávame v rôznych oblastiach praxe s javmi, ktoré sa dajú charakterizovat sadou 2 dát t.j. $(X_1, X_2, ..., X_n)$ a $(Y_1, Y_2, ..., Y_n)$, kde n je počet pozorovaní. Ak tieto dáta považujeme za realizácie náhodnéhňeho výberu z neznámych rozdelení X a Y, tak má zmysel skúmať ich korelovanost.
Potom môžeme interpretovať data aj ako výber z neznámeho rozdelenia náhodného vektoru $V = (X,Y)$, teda napozorované dáta sú náhodný výber z dvojrozmerého rozdelenia vektoru (X,Y) a chceli by sme vedieť, či sú náhodné rozdelenia X a Y medzi sebou korelované.

V prvej časti sa zoznámime s teóriou a konštrukciou štatistických odhadov korelačných koeficientov, ktoré použijeme na odhad korelácie z náhodného výberu. Táto časť slúži hlavne na vybudovanie matematického aparátu s ktorým budeme pracovať a potom ho aplikovať na data.

V druhej časti sa pozrieme na ilustraciu použitia na dátach a porovnáme empirické hodnoty korelačných koeficientov s reálnymi.

V tretej časti sa zameriame na simulačnú štúdiu a budeme testovať hypotétu nezávisloti na simulovaných výberoch a vyhodnotíme, ktorý test má najvačšiu silu pri danej hladine štatistického testu.

1. Matematický aparát

Teoretická časť

Zavedieme základné definície, tvrdenia a vety potrebné k zadefinovaniu korelačných koeficientov a ich následne štatistické odhady, použitie na príkladoch a testovanie hypotéz o nezávislosti.

1.1 Základné pojmy a definície

Definícia 1.1 (Zvára a Štepán (1997), str. 104 : **Kovariancia náhodných veličín**). *Nech pre náhodné veličiny X,Y existujú rozptyly. Výraz*

$$cov(Y,X) = E(X - EX)(Y - EY) \tag{1.1}$$

sa nazýva kovariancia náhodných veličín X,Y.

Tvrdenie 1.1 (*Vlastnosti kovariancie*)
Pre každé $a,b,c,d \in \mathbb{R}$ a náhodné veličiny X,Y platí
(i) $cov(X,Y) = E(XY) - (EX)(EY)$
(ii) $cov(aX,bY) = ab\ cov(X,Y)$
(iii) $cov(X,Y + Z) = cov(X,Y) + cov(X,Z)$
(iv) $cov(X,X) \geq 0$
(v) $cov(aX + b,cY + d) = ac\ cov(X,Y)$

1.2 Pearsonov korelačný koeficient

Definícia 1.2 (Hazewinkel (1989), str. 301 : **Pearsonov korelačný koeficient**). *Nech X a Y sú náhodné veličiny s konečnými druhými momentmi a s kladnými rozptylmi. Definujeme Pearsonov korelačný koeficient*

$$\rho_p = \frac{cov(X,Y)}{\sqrt{(varX)(varY)}}. \tag{1.2}$$

Pre korelačný koeficient platí

$$-1 \leq \rho_p \leq 1. \tag{1.3}$$

Rovnosť $\rho_p = 1$ platí práve vtedy, keď $Y = a + bX$ s pravdepodobnosťou 1, pričom $b > 0$, a $\in \mathbb{R}$. Analogicky rovnosť platí práve vtedy, keď $Y = a + bX$ s pravdepodobnosťou 1 pre $b < 0$, a $\in \mathbb{R}$.(Hazewinkel (1989), str. 301)

Dôkaz. ($-1 \leq \rho_p \leq 1$)
Zadefinujme si skalárny súčin a normu indukovanú skalárnym súčinom ako

$$< X,Y >= cov(X,Y), \quad ||X|| = \sqrt{< X,X >}, \tag{1.4}$$

kovariácia spĺňa definíciu skalárneho súčinu (viz. **Tvrdenie 1.1**) a teda platí Schwarzova nerovnosť. Zo Schwarzovej nerovnosti dostávame

$$|< X,Y >| \leq ||X||\,||Y|| \tag{1.5}$$

a po dosadení

$$< X,Y >= cov(X,Y), \quad ||X|| = \sqrt{< X,X >} \tag{1.6}$$

dostávame

$$|cov(X,Y)| \leq \sqrt{cov(X,X)cov(Y,Y)} \tag{1.7}$$

a teda po úprave

$$|cov(X,Y)| \leq \sqrt{var(X)var(Y)}. \tag{1.8}$$

Nakoniec vynásobíme obe strany $1/\sqrt{var(X)var(Y)}$ a dostávame

$$\frac{|cov(X,Y)|}{\sqrt{var(X)var(Y)}} \leq 1. \tag{1.9}$$

Nakoľko menovateľ je vždy > 0 tak rozšírenie absolútnej hodnoty na cely zlomok nám nezmení hodnotu zlomku, píšeme

$$\left| \frac{cov(X,Y)}{\sqrt{var(X)var(Y)}} \right| \leq 1 \tag{1.10}$$

a teda dostávame

$$|\rho_p| \leq 1. \qquad \square$$

Definícia 1.3 (Anděl (2007), str. 67 : ***Výberový priemer a rozptyl*)**. *Postupnosť nezávislých rovnako rozdelených veličín X_1,...,X_n sa nazýva náhodný výber. Číslo n je rozsah výberu, $n \geq 2$. Zavedieme veličiny*

$$\bar{X} = \sum_{i=1}^{n} \frac{X_i}{n}, \quad S^2 = \sum_{i=1}^{n} \frac{(X_i - \bar{X})^2}{n-1} \tag{1.11}$$

Veličina $\bar{X}$je výberový priemer. Veličina S^2 je výberový rozptyl.

Definícia 1.4 (Anděl (2007), str 229 : **Výberový Pearsonov korelačný koeficient)**. *Majme náhodný výber (X_1,Y_1),...,(X_n,Y_n) z nejakého dvojrozmerného rozdelenia. Označme $\bar{X}$ a S_X^2 ako výberový priemer a rozptyl výberu X_1,...,X_n a podobne $\bar{Y}$ a S_Y^2 pre Y_1,..., Y_n. Ďalej definujeme*

$$S_{XY} = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \bar{X})(Y_i - \bar{Y}). \tag{1.12}$$

Ak $S_X^2 > 0$ a $S_Y^2 > 0$, definujeme výberový Pearsonov korelačný koeficient vzorcom

$$r_p = \frac{S_{X,Y}}{\sqrt{S_X^2 S_Y^2}}. \tag{1.13}$$

Niekedy sa miesto r_p píše $r_{p_{X,Y}}$.

Pearsonov výberový koeficient nadobúda hodnoty $|r_p| \leq 1$. Hodnotu $r_p = 1$ nadobúda práve vtedy, keď náhodný výber $(X_1,Y_1),...(X_n,Y_n)$ leží na priamke.

Geomterická interpretácia. K tomu aby sme mohli geometricky interpretovať pearsonov korelačný koeficient musíme zadefinovať regresnú priamku a potom ukážeme, že Pearsonov korelačný koeficient je naviazaný na veľkosť uhla medzi 2 regresnými priamkami

Definícia 1.5 (Anděl (2007), str 111 : **Lineárny model)**. *Majme náhodný vektor* $\mathbf{Y}= (Y_1,...,Y_n)'$ *a maticu daných čísel* $X_{n\times k}$. Predpokladajme, že sa Y riadmi tkz. lineárnym modelom

$$\mathbf{Y} = \mathbf{X}\beta + \epsilon, \tag{1.14}$$

kde $\beta = (\beta_1,...,\beta_k)$ *je vektor neznámych parametrov a* $\epsilon = (\epsilon_1,...,\epsilon_n)'$ *je vektor náhodných veličín splňujúcich podmienky*

$$\mathbf{E}\,\epsilon = \mathbf{0}, \quad var\,\epsilon = \sigma^2\,\mathbf{I} \tag{1.15}$$

Definícia 1.6 (Zvára a Štepán (1997), str 185 : **Regresná priamka)**. *Majme náhodný vektor* $\mathbf{Y}= (Y_1,...,Y_n)'$ *a maticu daných čísel* $X_{n\times 2}$ a predpokladajme, že $\epsilon = (\epsilon_1,...,\epsilon_n)'$ sú nezávislé náhodné veličiny s rozdělením $N(0,\, \sigma^2)$. Potom model

$$Y_i = \beta_0 + x_i\beta_1 + \epsilon_i, \quad i = 1,...,n \tag{1.16}$$

je špeciálnym prípadom lineárneho regresného modelu a hovoríme o jednoduchej lineárnej regresii,kde $\beta = (\beta_0, \beta_1)'$ a

$$X = \begin{pmatrix} 1 & x_1 \\ ..... & \\ 1 & x_n \end{pmatrix}, \tag{1.17}$$

Veta 1.1 (Stuart a Ord (2010), str 287 : **Odhady parametrov** β_0 **a** β_1**)**. Ak máme linearny model definovaný tak ako v 1.16, tak odhady koeficientov β_0 a β_1 metodou najmenších súčtov štvorcov sú

$$\beta_1 = \frac{n\sum\limits_{i=1}^{n} x_i y_i - \sum\limits_{i=1}^{n} x_i \sum\limits_{i=1}^{n} y_i}{n\sum\limits_{i=1}^{n} x_i^2 - (\sum\limits_{i=1}^{n} x_i)^2} = \frac{\mu_{11}}{\sigma_1^2}, \tag{1.18}$$

$$\beta_0 = \frac{\sum\limits_{i=1}^{n} x_i^2 \sum\limits_{i=1}^{n} y_i - \sum\limits_{i=1}^{n} x_i \sum\limits_{i=1}^{n} x_i y_i i}{n\sum\limits_{i=1}^{n} x_i^2 - (\sum\limits_{i=1}^{n} x_i)^2} = \mu_y - \beta_1\mu_x, \tag{1.19}$$

kde σ_1 je smerodatna odchylka X (nezavislej premennej), σ_2 je smerodatna odchylka Y (zavislej premennej) a μ_{11} je kovariancia medzi X a Y. Pak po dosadení 1.18 a 1.19 a aplikovaní strednej hodnoty dostavame

$$y = \mu_y - \beta_1\mu_x + \beta_1 x \tag{1.20}$$

Upravou

$$(y - \mu_y) = \beta_1(x - \mu_x) \tag{1.21}$$

Potom je definovám Pearsonov korelačný koeficient v novom značení

$$p = \frac{\mu_{11}}{\sigma_1 \sigma_2} \tag{1.22}$$

Veta 1.2 (Stuart a Ord (2010), str 287 : **Veľkosť uhla medzi 2 regresnými priamkami)**. Veľkosť uhla Θ medzi dvoma regresnými priamkami je potom rovná

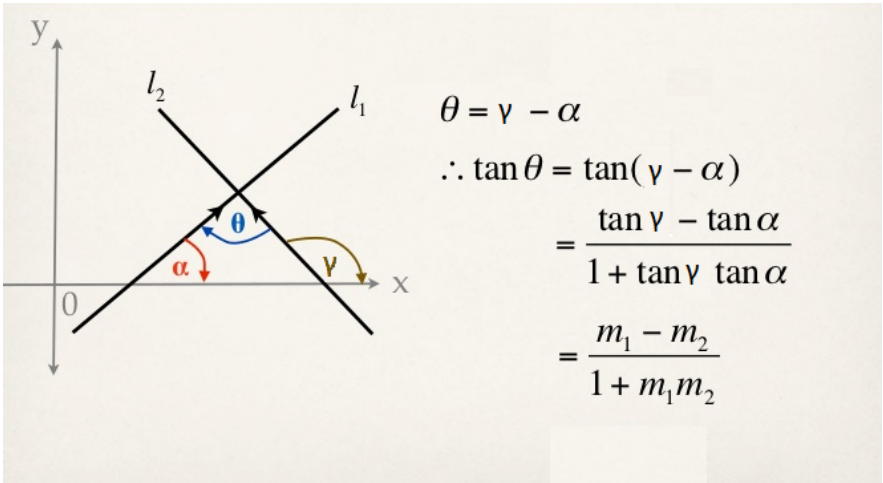
$$tan\ \Theta = \frac{\sigma_1 \sigma_2}{\sigma_1^2 + \sigma_2^2} \left(\frac{1}{p} - p \right) \tag{1.23}$$

Dôkaz. (Veľkost uhla medzi 2 regresnými priamkami)
Chceme ukázať, že 1.23 platí a k tomu si musíme pripomenúť jednu vlastnost tangensu uhla, ktorý zvierajú dve priamky.

Lemma. (Uhol medzi 2 priamkami). Uhol Θ, ktorý zvierajú dve priamky medzi sebou je rovný rozdielu uhlov γ a α, ktoré zvierajú priamky s osou x. Potom platí

$$tan\ \Theta = \frac{tan\ \gamma - tan\ \alpha}{1 + tan\ \gamma\ tan\ \alpha} \tag{1.24}$$

Jednotlivé kroky dôkazu su jasne vidieť z obrázku 1.1



Obr. 1.1: Uhol medzi 2 priamkami
Zdroj: https://www.slideshare.net/SimonBorgert/angle-between-2-lines

Uvažujme dve regresívne rovnice priamok. Prvá je taká, že nech Y je zavislá premenná a X je nezávislá premenná, potom rovnica regresnej priamky po aplikovaní strednej hodnoty je rovná

$$y - \mu_y = \beta_2(x - \mu_x) \tag{1.25}$$

za β dosadíme 1.19 a upravíme tak, aby sme dostali Pearsonov korelačný koeficient v rovnici, zo vzťahu 1.22. Teda

$$y - \mu_y = p\frac{\sigma_y}{\sigma_x}(x - \mu_x) \tag{1.26}$$

Druhá regresná rovnica, kde teraz X je závislá premenná a Y je nezávislá premenná má tvar

$$x - \mu_x = \beta_2(y - \mu_y) \tag{1.27}$$

Po rovnakej úprave ako v úprave 1.26

$$x - \mu_x = p\frac{\sigma_x}{\sigma_y}(y - \mu_y) \tag{1.28}$$

Keď si z prvej rovnice vyjadrím

$$y - \mu_y = \frac{\sigma_y}{p\,\sigma_x}(x - \mu_x) \tag{1.29}$$

Z 1.29 a 1.26 dostávame

$$m_1 = p\frac{\sigma_y}{\sigma_x}, \quad m_2 = \frac{\sigma_y}{p\,\sigma_x} \tag{1.30}$$

Po dosadení potom do 1.24 dostávame

$$tan\ \Theta = \frac{\frac{\sigma_y}{p\,\sigma_x} - \frac{p\,\sigma_y}{\sigma_x}}{1 + \frac{p\,\sigma_y\sigma_y}{p\,\sigma_x\sigma_x}} \tag{1.31}$$

Zjednodušíme a vyjmeme pred zátvorku

$$tan\Theta = \frac{\frac{\sigma_y}{\sigma_x}\left[\frac{1}{p} - p\right]}{1 + \frac{\sigma_y^2}{\sigma_x^2}} \tag{1.32}$$

Rozšírime vrchny zlomok s σ_x/σ_x a potom skrátime σ_x^2 so zlomkom v menovateli a dostávame

$$tan\ \Theta = \frac{\sigma_y\sigma_x}{\sigma_x^2 + \sigma_y^2}\left[\frac{1 - p^2}{p}\right] \qquad \square$$

Pozorovanie. Ak p = 1, tak tangens uhla Θ, ktorý zovierajú regresné priamky je 0, čo odpovedá uhlu 0°, teda priamky zvierajú uhol rovný 0°. Ak p konverguje do 0 z prava, tak tangens ∞ je rovný 90° a uhol, ktorý zvierajú regresné priamky je rovný 90°, teda sú na seba kolmé. Teda, keby zoberieme ako náhodnú veličinu iba realizácie stredných hodnôt na regresnej priamke, tak by sa dalo hovoriť o tom, že Pearsonov korelačný koeficient hovorí o uhle, ktorý zvierajú takto špecialne vybrané náhodné veličiny.

Veta 1.3 (Anděl (2007), str.231 : **Rozdelenie výberového Pearsonovho kor. koeficientu)**. Nech $(X_1,Y_1),...,(X_n,Y_n)$ je náhodný výber z dvojrozmerného normálneho rozdelenia, ktoré ma kladné rozptyly a nech ešte $\rho_p = 0$ a $n \geq 3$. Potom náhodná veličina

$$T_p = \frac{r_p}{\sqrt{1-r_p^2}}\sqrt{n-2} \sim t_{n-2}. \tag{1.33}$$

Potom rozdelenie štatistiky (1.33) nám umožňuje testovať hypotézu $H_0 : \rho_p = 0$ oproti alternatíve $H_1 : \rho_p \neq 0$. Ak bude testová štatistika $|T| \geq t_{(n-2)}(1-\frac{\alpha}{2})$ tak hovoríme, že zamietame nulovú hypotézu na hladine α, kde $t_{(n-2)}$ je kvantilová funkcia Študentovho t-rozdelenia o $n-2$ stupňoch voľnosti.

Poznámka.(skr, str. 176). Nech $(X_1,Y_1),...,(X_n,Y_n)$ je náhodný výber z dvojrozmerného rozdelenia s konečnou nesingulárnou variačnou maticou, tak pri predpoklade, že X_i a Y_i sú nezávislé má testová štatistika rozdelenie

$$T_p = \frac{r_p}{\sqrt{1-r_p^2}}\sqrt{n-2} \overset{as.}{\sim} N(0,1), \tag{1.34}$$

kde $\overset{as.}{\sim}$ znamená, konvergenciu v distribucii pre n $\to \infty$.

Poznámka. (Athreya a Lahiri (2006), str. 288). Nech náhodný výber X_n, $n > 0$ a označme F_{X_n} ako distribučnú funkciu, z ktorej nahodný výber X_n, $n > 0$ pochádza. Potom postupnosť $\{X_n\}_{n=1}^{\infty}$ konverguje v distribúcii k X, ak

$$\lim_{n\to\infty} F_{X_n}(x) = F_X(x) \tag{1.35}$$

v každom bode x, kde $F_X(x)$ je spojitá.

Teda za platnosti hypotézy, že Xi a Yi jsou nezávislé a počet pozorovaní n bude rásť nad všetky meze, tak testová štatistika T_p bude konvergovať v distribucií k rozdeleniu N(0,1).

Veta 1.4 (Anděl (2007), str 232 : **Veta o Fisherovej z-transformácii**). Keby chceme testovať hypotézu $H_0 : \rho = \rho_0$, kde $\rho_0 \neq 0$, proti alternatíve $H_1 : \rho \neq \rho_0$ tak použijeme rozptyl stabilizujúcu Fisherovu z-transformáciu

$$Z = \frac{1}{2}\ln\frac{1+r_p}{1-r_p}. \tag{1.36}$$

Teda

$$\zeta_0 = \frac{1}{2}\ln\frac{1+\rho_0}{1-\rho_0}. \tag{1.37}$$

Za platnosti $H_0 : \rho = \rho_0$ a ak dáta pochádzajú z normálneho dvojrozmerného rozdelenia, tak má náhodná veličina

$$U = \sqrt{n-3}(Z-\zeta_0) \overset{as.}{\sim} N(0,1) \tag{1.38}$$

Potom H_0 zamietneme, keď $|U| \geq u_{(1-\frac{\alpha}{2})}$,kde u je kvantilová funkcia rozdelenia N(0,1). Pomocou tejto transformácie vieme skonštruovať konfidenčný interval pre ρ . Ak ρ je skutočná hodnota korelačného koeficientu, tak

$$P\left[\sqrt{n-3}\left|Z - \frac{1}{2}\ln\frac{1+\rho}{1-\rho}\right| < u_{(1-\frac{\alpha}{2})}\right] \doteq 1-\alpha, \tag{1.39}$$

kde $\doteq$ znamená aproximáciu danému výrazu.

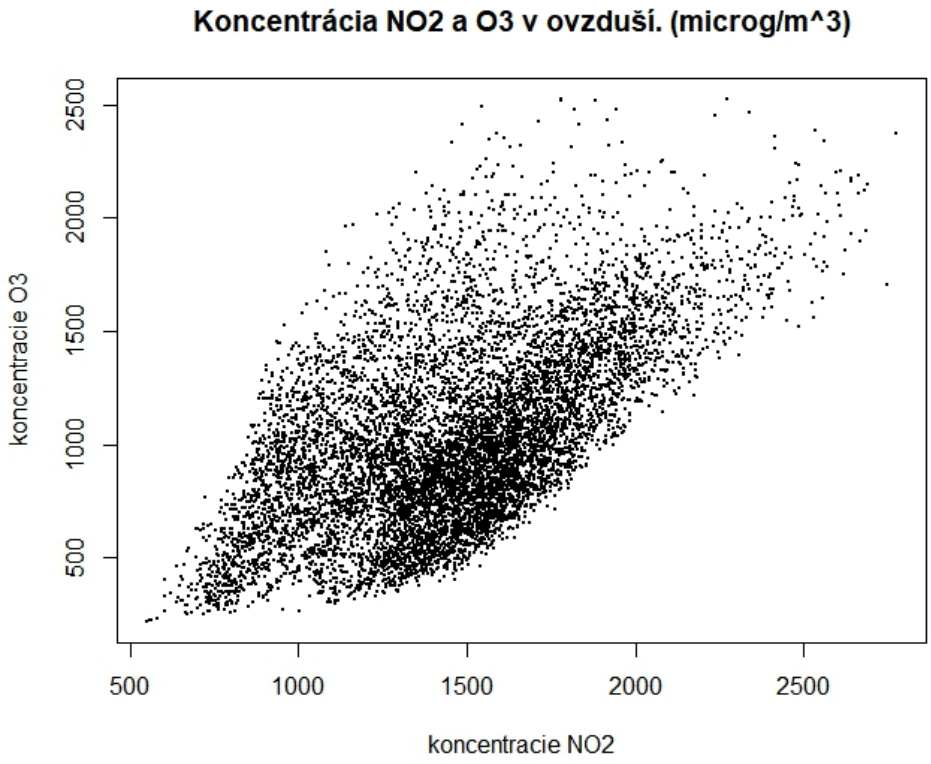
Potom po úprave dostaneme, že interval spoľahlivosti pre ρ je

$$\left(\frac{D-1}{D+1}, \frac{H-1}{H+1}\right),$$

kde

$$D = exp\left\{2Z - \frac{2u_{(1-\frac{\alpha}{2})}}{\sqrt{n-3}}\right\}, H = exp\left\{2Z + \frac{2u_{(1-\frac{\alpha}{2})}}{\sqrt{n-3}}\right\} \tag{1.40}$$

Príklad 1.1
Mame 9358 meraní , ktoré merali kvalitu ovzdušia od marca 2014 do februára 2015 s hodinovým intervalom , sledovali sme ukazatele oxidu dusičného(NO_2) a ozónu(O_3) v ovzduší. Data som zobral zo zdroja (dts).
Pozrime sa na bodový diagram závislosti NO_2 a O_3



Obr. 1.2: Koncentrácia NO2 a O3

Poznámka. Dáta pochádzajú z časovej rady a nespĺňajú definíciu náhodného výberu, čo nám za účelom ilustrácie nevadí, ale keby chceme na dáta testovať napríklad hypotézu nezávislosti, museli by okrem iného spĺňať definíciu náhodného výberu. Ale to nie je účelom tejto ukážky.

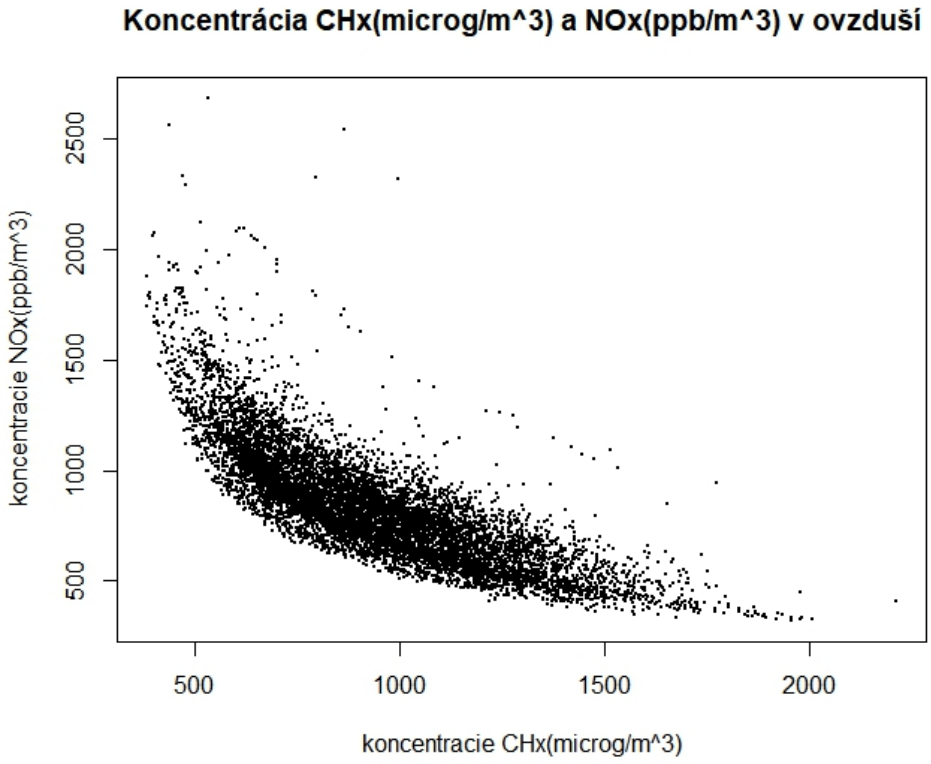
Z bodového diagramu je vidieť, že by mohla existovať závislosť medzi veličinami koncentrácie NO_2 a O_3 v ovzduší. Z dát je zrejme, že by medzi nimi mohla byť lineárna závislosť, preto k určeniu vzťahu medzi veličinami použije Pearsonov korelačný koeficient

$$r_p = 0{,}59.$$

Interpretácia je nasledovná, keď nameriame vyššiu koncentraciu $NO_2(O_3)$ tak môžeme očakávať aj zvýšenú koncentráciu $O_3(NO_2)$. Nedá sa nič presnejšie o koncentráciách NO_2 a O_3 povedať, aj keby sme mali hodnotu pearsonovho korelačného koeficientu rovnú 1(-1) tak maximálne môžeme iba povedať, že pozorovanie O_3 je bližšie neurčený kladný(záporný) násobok čísla b koncentrácie NO_2 v ovzduší zvýšený (znížený) o neznámu konštantu a.

Teda v tomto prípade nenulovosť Personovho korelačného koeficientu $r_p = 0{,}59$ iba poukazuje na to, že existuje stredne-silná lineárna závislosť medzi koncentráciami NO_2 a O_3 v ovzduší.

Príklad 1.2
Použijeme tie iste dáta , ale teraz sa pozrieme na koncentráciu nekovových zlúčenín uhľovodíka (ozn. CHx) a celkovú koncentráciu zlúčenín oxidu dusnatého (ozn. NOx). Celkový počet pozorovaní je 8991. Pozrime sa na bodový diagram koncentrácii NOx na CHx v ovzduší.



Obr. 1.3: Koncentrácia CHx a NOx

Poznámka. ppb znamená Parts per billion("počet častíc na jednu miliardu")

Z bodového diagramu vidíme, že dáta by sme mali použiť niečo robustnejšie, nie sa len zamerať na jej lineárnu časť (dáta neležia na priamke, skôr na parabolickom oblúku), preto by sme nemali použiť Personov korelačný koeficient, ktorý meria iba mieru lineárnej závislosti, ale niečo viac obecnejšie, čo nám zachytí koreláciu medzi dátami, ktorá je spôsobená nie len linearnou zložkou, a tým je Spearmanov alebo Kendallov korelačný koeficient.

1.3 Spearmanov korelačný koeficient

Pearsonov korelačný koeficient meria mieru lineárnej závislosti medzi veličinami. Keď potrebujeme zistiť koreláciu v dátach, ktoré nie su len lineárne závislé, tak sa môžeme zamerať na celkovú monotóniu v dátach. To robí Spearmanov a Kendallov korelačný koeficient.

Ďalšia výhoda Spearmanovho korelačného koeficientu je, že niekedy v náhodnom výbere $(X_1,Y_1),...,(X_n,Y_n)$ sa nedajú hodnoty uvedených náhodných veličín presne stanoviť a je k dispozícii iba ich poradie. Ak sú poradia náhodného výberu $(X_1,..., X_n)$ a $(Y_1,..., Y_n)$ dosť podobné, nepochybne to svedčí o istej závislosti medzi nimi.

Definícia 1.7 (Hazewinkel (1989), str. 376 : **Výberový Spearmanov korelačný koeficient)**. *Nech $(X_1,Y_1)',...,(X_n,Y_n)'$ je výber zo spojitého dvojrozmerného rozdelenia. Nech $R_1,...,R_n$ je usporiadané poradie veličín $X_1,...,X_n$ a nech $Q_1,...,Q_n$ je usporiadané poradie veličín $Y_1,...,Y_n$ potom definujeme ako Spearmanov korelačný koeficient*

$$r_s = \frac{S_{RQ}}{\sqrt{S_R^2 S_Q^2}}, \tag{1.41}$$

kde

$$S_{RQ} = \frac{1}{n-1}\sum_{i=1}^{n}(R_i - \bar{R})(Q_i - \bar{Q}) \tag{1.42}$$

a S_R^2,S_Q^2 sú výberové rozptyly počítané z poradia.

Príklad.1.3
Ak sa opýtame dvoch športových komentátorov nech ohodnotia hráčov pridelením poradia od najlepšieho po najhoršieho. Ak obaja komentátori ohodnotia hráčov dosť podobne, teda ich rebríček hráčov od najhoršieho po najlepšieho bude podobný alebo rovnaký, môžeme hovoriť o kladnej korelácii. Existuje istá závislosť hodnotenia jedného komentátora od druhého, ak prvý komentátor umiestni hráča vysoko v rebríčku, môžeme očakávať, že druhý komentátor ho tiež umiestni vysoko v rebríčku.

Populačná verzia Spearmanovho korelačného koeficientu je definovaná nasledovne:

Definícia 1.8 (Nelsen (2006), str. 170 : **Spearmanov korelačný koeficient)**. *Nech X a Y sú náhodné veličiny, ktoré majú distribučné funkcie F a G (t.j. F(x)= $P[X \leq x]$). Potom náhodné veličiny U,V zadefinujeme ako U=F(X) a V=G(Y). Potom pre tieto nové náhodné veličiny U a V definujeme Spearmanov korelačný koeficient*

$$\rho_s = \frac{E(UV) - E(U)E(V)}{\sqrt{(varU)(varV)}} = \frac{cov(U,V)}{\sqrt{(varU)(varV)}}. \tag{1.43}$$

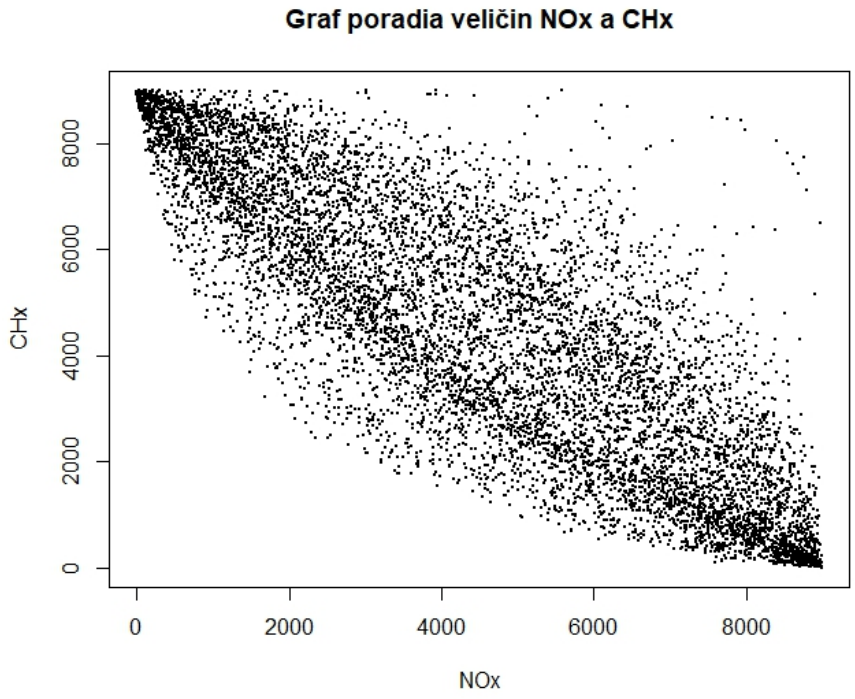
Spearmanov korelačný koeficient je vlastne Pearsonov korelačný koeficient aplikovaný na náhodné veličiny U a V.

Veta 1.5 (skr, str 167 : **Testovanie hypotéz pre Spearmanov korelačný koeficient)**. Nech $(X_1,Y_1),...,(X_n,Y_n)$ je náhodný výber z dvojrozmerného rozdelenia s konečnou nesingulárnou variačnou maticou, tak pri predpoklade, že X_i a Y_i sú nezávislé má testová štatistika rozdelenie

$$T_s = \frac{r_s}{\sqrt{1-r_s^2}}\sqrt{n-2} \overset{as.}{\approx} N(0,1). \tag{1.44}$$

Hypotézu o nezávislosti zamietame ak $|T_s| > u_{1-\frac{\alpha}{2}}$.

Príklad 1.2(Pokračovanie)
Teraz máme matematicky aparát aby sme preskúmali závislosť CHx a NOx, viz.bodový diagram 1.2.
Spearmanov korelačný koeficient sa počíta rovnako ako Pearsonov korelačný koeficient, ale dátové sady sú odlišné, Pearsonov sa počíta zo samotných dát, zatiaľ čo Spearmanov sa počíta z poradia dát.



Obr. 1.4: Poradie dát CHx a NOx

Na takýto súbor poradí použijeme Personov korelačný koeficient. Je vidieť, že poradia majú lineárny charakter, tak je jeho použitie vhodné.

Vypočítajme teda Spearmanov korelačný koeficient

$$r_s = -0{,}85$$

Spearmanov korelačný koeficient poukazuje na silnú negatívnu koreláciu v dátach a zameriava na celkovú monotónnosť v dátach, teda je robustnejší jak Pearsonov, ktorý sa zameriava iba na lineárnu zložku závislosti v dátach.
Pre porovnanie vypočítame Pearsonov korelačný koeficient nad rovnakými dátami

$$r_p = -0{,}80.$$

Pearsonov korelačný koeficient tiež dobre zachytáva závislosť v dátach. V dátach nie sú žiadne odľahlé pozorovania čo zapríčiňuje, že Pearsonov korelačný koeficient sa dá dobre použiť na odhad korelácie v dátach. Keby sme však v dátach mali nejaké odľahlé pozorovania tak by Pearsonov korelačný koeficient bol značne ovplyvnený týmito pozorovaniami.

1.4 Kendallov korelačný koeficient

Definícia 1.9 (Anděl (2007), str 240 : **Výberový Kendallov korelačný koeficient)**. *Nech* $(X_1,Y_1),...,(X_n,Y_n)$ *je výber zo spojitého dvojrozmerného rozdelenia. Nech* $R_1,...,R_n$ *je poradie veličín* $X_1,...,X_n$ *a nech* $Q_1,...,Q_n$ *je poradie veličín* $Y_1,...,Y_n$ *potom Kendallov koeficient* τ *je definovaný*

$$\tau_n = \frac{1}{n(n-1)} \sum \sum_{i \neq j} sign(R_i - R_j) sign(Q_i - Q_j), \tag{1.45}$$

kde pre $\forall z \in R$ je fukncia $sign(z)$ definovana nasledovne:

$$\begin{aligned} sign(z) &= -1, \quad z < 0, \\ sign(z) &= \quad 0, \quad z = 0, \\ sign(z) &= +1, \quad z > 0. \end{aligned} \tag{1.46}$$

Poznámka. Kendallov korelačný koeficient sa dá počítať i rovno z náhodného výberu. Uvedomíme si, že platí

$$sign(R_i - R_j) = sign(x_i - x_j).$$

Potom

$$\tau_n = \frac{1}{n(n-1)} \sum \sum_{i \neq j} sign(x_i - x_j) sign(y_i - y_j). \tag{1.47}$$

Kendallov korelačný koeficient počíta počet súhlasných a nesúhlasných párov. To, že páry sú súhlasné znamená, že $sign(x_i - x_j) = sign(y_i - y_j)$, inak povedané, ak sa v danej dvojici indexov (i,j), i<j, pozorovanie zníži (zvýši) v x-ovej súradnici, tak sa aj zníži (zvýši) jeho hodnota v y-ovej súradnici. Nesúhlasné páry su také, pre ktoré $sing(x_i - x_j) = -sign(y_i - y_j)$, teda ak sa v danej dvojici indexov (i,j), i<j, pozorovanie zníži (zvýši) v x-ovej súradnici, tak sa ďalšie pozorovanie v y-ovej súradnici zvýši (zníži).

Korelácia v náhodnom výbere bude vysoká, keď budú mať páry (x_i,y_i) malý rozdiel medzi ich poradiami R_i a Q_i. Teda ak je korelácia kladná, tak očakávame, že keď bude poradie pozorovania x_i veľké (relatívne k ostatným pozorovaniam x_j,

$i \neq j$), tak aj poradie pozorovania y_i bude veľké (relatívne k ostatným pozorovaniam y_j, $i \neq$j).

Populačná verzia je definovaná nasledovne.

Definícia 1.10 (Nelsen (2006), str. 158 : **Populačný Kendallov korelačný koeficient)**. *Nech* (X_1,Y_1) *a* (X_2,Y_2) *sú 2 nahodné vektory zo spojitého rozdelenia H, tak hodnota Kendallovho korelačného koeficientu* τ *je definovaná*

$$\tau = P[(X_1 - X_2)(Y_1 - Y_2) > 0] - P[(X_1 - X_2)(Y_1 - Y_2) < 0]. \tag{1.48}$$

Kendallov korelačný koeficient nadobúda hodnoty od 1 po -1. Hodnotu 1 nadobúda práve vtedy, keď sú všetky páry v náhodnom výbere súhlasné. Ak sú všetky páry z náhodného výberu $(x_1,y_1)'$,...$(x_n,y_n)'$ súhlasné, to znamená, že pre ich poradie platí $R_i = Q_i$ pre $\forall i$, $i \in n$, a $x_{R(1)} < x_{R(2)} < x_{R(3)} ... < x_{R(n)}$ a $y_{Q(1)} < y_{Q(2)} < y_{Q(3)} < ... < y_{Q(n)}$, kde n je veľkosť náhodného výberu a $X_{R_{(i)}} = X_i$ a $Y_{S_{(i)}} = Y_i$. Hodnotu -1 nadobúda ak zameníme znamienka nerovností.

Poznámka. (Anděl (2007), str. 241)
Máme náhodný výber (X_1,Y_1),...,(X_n,Y_n). Nech T je počet dvojic indexov (i,j), i<j, ktoré sú súhlasné. Potom

$$\tau_n = \frac{4T}{n(n-1)} - 1. \tag{1.49}$$

Ak (X_1,Y_1),...,(X_n,Y_n) je náhodný výber zo spojitého dvojrozmerného rozdelenia. Potom za nezávislosti X_i a Y_i platí

$$E\tau_n = 0 \quad , var\ \tau_n = \frac{2(2n+5)}{9n(n-1)} \tag{1.50}$$

a štatistika

$$T_k = \frac{\tau_n}{\sqrt{var\ \tau_n}} \overset{as.}{\approx} N(0,1) \tag{1.51}$$

Hypotézu o nezávislosti zamietame ak $|T_k| > u_{1-\frac{\alpha}{2}}$.

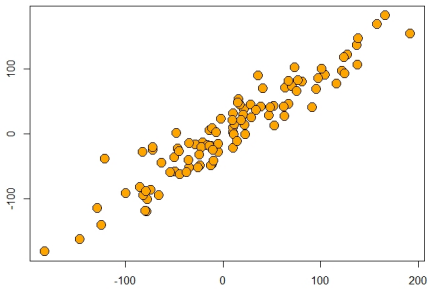
Poznámka. Kendallov korelačný koeficient nám tiež odhaduje mieru monotónnosť medzi náhodnými veličinami X a Y, ale iným spôsobom ako Pearsonov alebo Spearmanov korelačný koeficient.

2. Ilustrácia na dátach

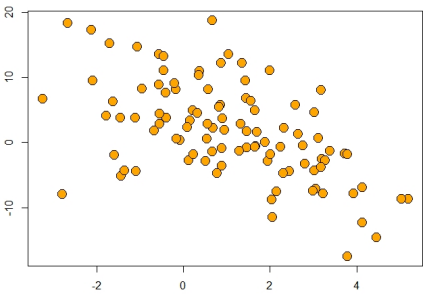
V tejto časti sa zameriame na ilustráciu použitia korelačných koeficientov, ktoré budú pozostávať z aplikovania odhadov korelačných koeficientov na rovnaký typ dát. Zameriame sa na porovnanie presnosti odhadov a vyhodnoteníme silné a slabé stránky korelačných koeficientov nad danými dátami.

2.1 Ilustračné dáta

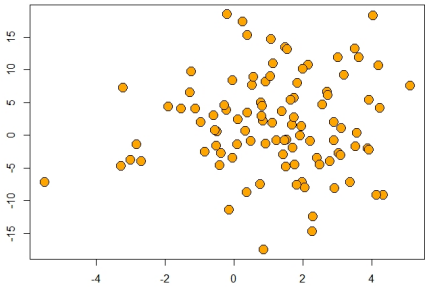
Dáta si vygenerujeme z dvojrozmerného normálneho rozdelenia a aplikujeme na nich výberové verzie korelačných koeficientov.



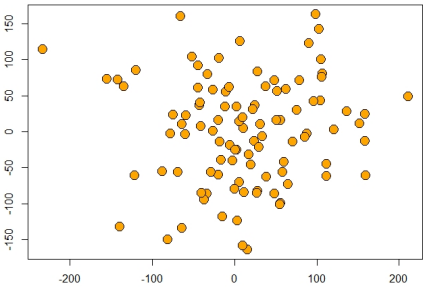
Obr. 2.1: σ_x=80, σ_y=80, ρ=0,95



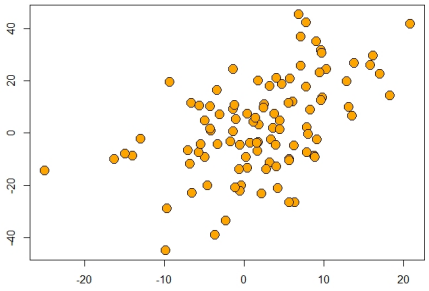
Obr. 2.2: σ_x =2, σ_y =8, ρ=-0,6



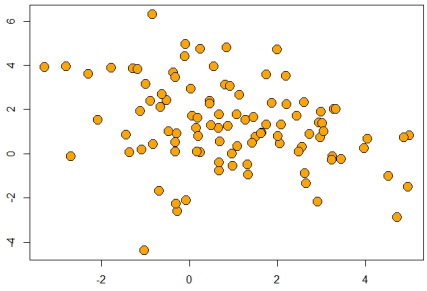
Obr. 2.3: σ_x =2, σ_y =8, ρ=0



Obr. 2.4: σ_x=80, σ_y=80, ρ=0,1



Obr. 2.5: σ_x=8, σ_y=20, ρ=0,5



Obr. 2.6: σ_x=2, σ_y=2, ρ=-0,35

Mame 6 sad dát, 100 pozorovaní z dvojrozmerného normálneho rozdelenia, ktoré sme vygenerovali na základe parametrov, ktoré sme si sami určili, teda stredné hodnoty, rozptyly a kovarianciu medzi veličinami X a Y.
Náhodné výbery majú parametre normálneho rozdelenia, z ktorého sme ich generovali, pod grafmi, okrem stredných hodnôt, ktoré sú pre každú sadu dát rovnané a to $EX = 1$ a $EY = 1$.

Poznámka. Na stredných hodnotách EX a EY nezáleží, pretože z vlastností kovariancie **(Tvrdenie 1.1, vlastnosť (V))** vyplýva, že stredné hodnoty nemajú vplyv na korelačný koeficient.

Na dáta aplikujeme Pearsonov, Spearmanov a Kendalov korelačný koeficient. Zobrazíme v prehľadnej tabuľke.

Tabuľka 2.1: Tabuľka korelačných koeficientov

	1	2	3	4	5	6
Pearson	0,944	-0,550	0,050	0,043	0,503	-0,299
Spearman	0,941	-0,531	-0,004	0,039	0,497	-0,291
Kendal	0,796	-0,387	-0,006	0,029	0,340	-0,202
Skutočný	0,950	-0,600	0,000	0,100	0,500	-0,350

Ak chceme porovnať, ako sú odhadnuté korelačné koeficienty rozdielne od skutočných korelačných koeficientov, musíme najskôr zaviesť transformačné vzťahy, pretože Spearmanov a Kendallov korelačný koeficient nekoresponduje s ρ ako parametrom korelácie z ktorého daný náhodný výber pochádza.

Veta 1.6 (Anděl (2007), strana 239 : **Transformácia Spearmanovho korelačného koeficientu)**. Keď náhodný výber pochádza z dvojrozmerného normálneho rozdelenia, tak pre Spearmanov korelačný koeficient je daná aproximácia

$$r_p \doteq 2\sin(\frac{\pi}{6}r_s), \tag{2.1}$$

kde $\doteq$ znamená aproximáciu danému výrazu.

Veta 1.7 (Kendall (1970), strana 126 : **Transformácia Kendallovho korelačného koeficientu)**. Keď náhodný výber pochádza z dvojrozmerného normálneho rozdelenia, tak pre Kendallov korelačný koeficient je daná aproximácia

$$r_p \doteq \sin(\frac{\pi}{2}\tau_n), \tag{2.2}$$

kde $\doteq$ znamená aproximáciu danému výrazu.

Teda tabuľka po transformácii má tvar:

Tabuľka 2.2: Tabuľka transformovaných korelačných koeficientov

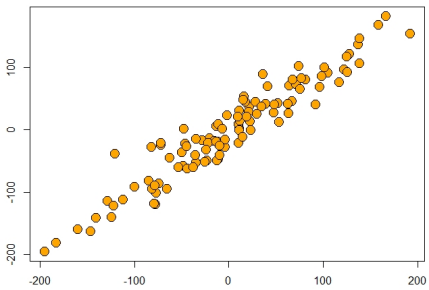
	1	2	3	4	5	6
Pearson	0,944	-0,550	0,050	0,043	0,503	-0,299
Spearman_T	0,946	-0,549	-0,004	0,041	0,515	-0,304
Kendal_T	0,949	-0,571	-0,009	0,046	0,509	-0,312
Skutočny	0,950	-0,600	0,000	0,100	0,500	-0,350

Po transformácii vidíme, že všetky korelačné koeficienty vcelku dobre vystihuju koreláciu v našom výbere.

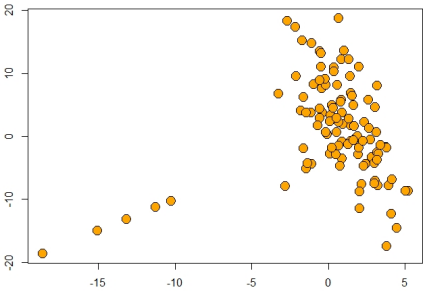
Vyskúšajme sa teraz pozrieť na vplyv odľahlých pozorovaní na tuto sadu dát. Do dát pridáme 5 hodnôt, ktoré budú 90,92,95,97 a 99 % kvantilom daného dvojrozmerného normálneho rozdelenia.

Nakoľko neexistuje analytická kvantilova funkcia pre dvojrozmerné normálne rozdelenie, tak sa spoľahneme na numericky nájdené korene pre kvantilovú funkciu. Budeme brať hodnoty, ktoré sa budú nachádzať na spodnej hranici kvantilu. Numerická metóda používa v svojom hľadaní také body (x_α,y_α), $x_\alpha = y_\alpha$ tak, aby $P[X < x_\alpha, Y < y_\alpha] \doteq 1 - \alpha$, kde $\doteq$ znamená aproximáciu danému výrazu.

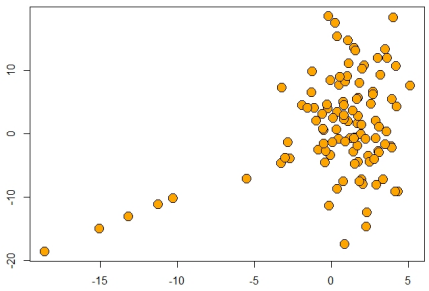
Po pridaní odľahlých pozorovaní naše dáta majú tvar .



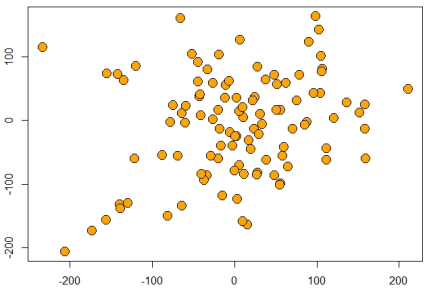
Obr. 2.7: σ_x=80, σ_y=80, ρ=0,95



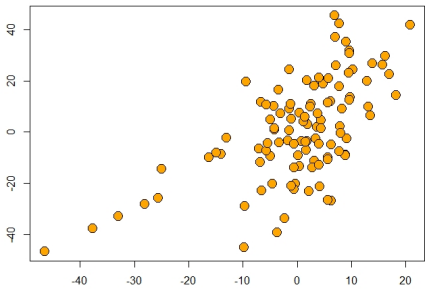
Obr. 2.8: σ_x =2, σ_y =8, ρ=-0,6



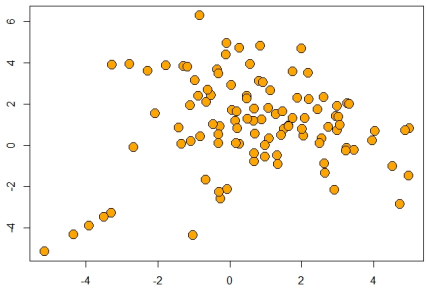
Obr. 2.9: σ_x =2, σ_y =8, ρ=0



Obr. 2.10: σ_x=80, σ_y=80, ρ=0,1



Obr. 2.11: σ_x=8, σ_y=20, ρ=0,5



Obr. 2.12: σ_x=2, σ_y=2, ρ=-0,35

Potom výsledná tabuľka ma tvar.

Tabuľka 2.3: Tabuľka s odľahlými hodnotami

	1	2	3	4	5	6
Pearson	0,954	0,129	0,391	0,231	0,615	0,056
Spearman	0,949	-0,333	0,127	0,157	0,562	-0,115
Kendal	0,814	-0,261	0,084	0,112	0,396	-0,092
Skutočny	0,950	-0,600	0,000	0,100	0,500	-0,350

Po transformácii

Tabuľka 2.4: Tabuľka s odľahlými hodnotami po transformácii

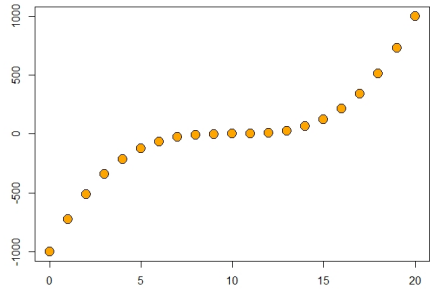
	1	2	3	4	5	6
Pearson	0,954	0,129	0,391	0,231	0,615	0,056
Spearman_t	0,953	-0,347	0,133	0,164	0,580	-0,120
Kendal_t	0,958	-0,399	0,132	0,175	0,583	-0,144
Skutočny	0,950	-0,600	0,000	0,100	0,500	-0,350

Vidíme, že po pridaní odľahlých koeficientov sa hodnota Pearsonovho koeficientu viac odchyľuje od skutočnej hodnoty na rozdiel od Spearmanovho alebo Kendallovho korelačného koeficientu, ktoré sú robustnejšie nad výbermi, ktoré majú v sebe odľahlé pozorovania.
Je to spôsobené tým, že Pearsonov korelačný koeficient berie nominálne hodnoty výberu a to má za následok jeho vychýlenie. Čím viac vzdialené bude toho pozorovanie, tým väčší dopad bude mať na výsledok odhadu.
Na rozdiel od Spearmanovho korelačného koeficientu, ktorý berie v potaz iba poradie, teda zanedbáva veľkosť odľahlého koeficientu. Kendallov korelačný koeficient je tiež menej citlivý na odľahlé pozorovania.

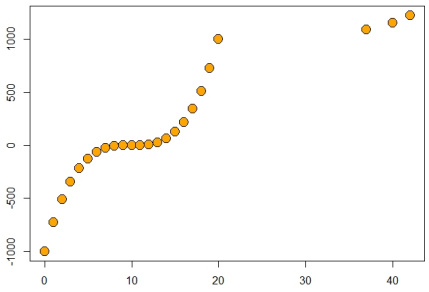
Preto sú odhady pomocou Spearmanovho a Kendallovho korelačného koeficientu robustnejšie vo výberoch, kde sa nachádzajú odľahlé pozorovania.
Spermanov a Kendallov korelačný koeficient sa tiež nazýva poradový korelačný koeficient. Pretože je citlivý na zmenu poradia v dátach.

Ďalšia časť. Preskúmajme ako dobre budú koeficienty odhadovať koreláciu na dátach, ktoré nepochádzajú z normálneho rozdelenia. Tieto dáta budu vygenerovane zámerne, aby sme sa pozreli ako dobre odhadujú korelačné koeficienty závislosť.

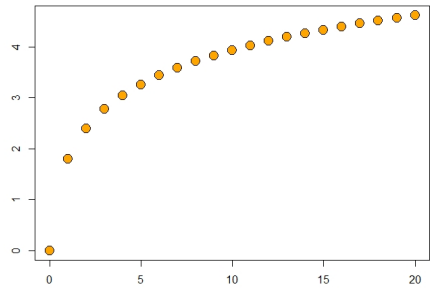
Naše data maju tvar



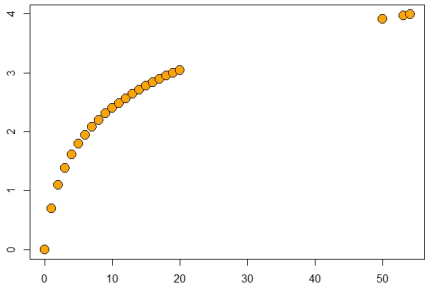
Obr. 2.13: Dáta so kubickou závisloťou



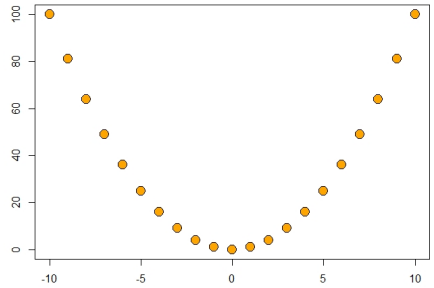
Obr. 2.14: Dáta so kubickou závisloťou s odlľhalými pozorovaniami



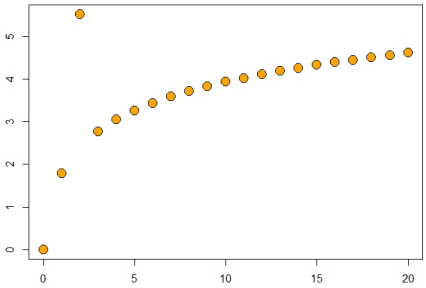
Obr. 2.15: Dáta s logaritmickou závislosťou



Obr. 2.16: Dáta s logaritmickou závislosťou s odlahlými pozorovaniami



Obr. 2.17: Špecialny prípad kvadratickej zavislosti



Obr. 2.18: Dáta s logaritmickou závislosťou s chybou vo výbere

Dáta sú vygenerované tak, aby sme poukázali na rozdiely medzi jednotlivými korelačnými koeficientmi.

Výstupom je tabuľka.

Tabuľka 2.5: Tabuľka korelačných koeficientov na dátach nepochádzajúcich z normálneho rozdelenia

	1	2	3	4	5	6
Pearson	0,918	0,928	0,862	0,850	0,000	0,657
Spearman	1,000	1,000	1,000	1,000	0,000	0,778
Kendal	1,000	1,000	1,000	1,000	0,000	0,829

Pearsonov korelačný koeficient zachytáva iba mieru lineárnej závislosti, preto keď ho používame nad dátami, ktoré nepochádzajú z normálneho rozdelenia, tak nezohľadňuje nelineárnu zložku závislosti v dátach. Preto je lepšie použiť Spearmanov alebo Kendallov korelačný koeficient na odhad korelácie v takejto sade dát.

Keď sa pozrieme na bodový diagram 2.17, je to špecialny prípad, pri ktorom všetky korelačné koeficienty sú rovné 0, pričom je jasné, že dáta majú kvadratickú závislosť.

3. Simulačné štúdie

V tejto kapitole sa budeme venovať testovaniu hypotézy nezávislosti

$$H_0 : data\ sú\ nezávislé$$

$$H_1 : data\ nie\ sú\ nezávislé$$

pre všetky 3 koeficienty nad rovnakými dátami a porovnáme silu testov na rovnakej hladine pre rôzne rozsahy výberov. Na testovanie použijeme matematický aparát, ktoré sme si vybudovali v 1 Sekcii. V prvej sade dát použijeme testy 1.33, 1.44 a 1.51. V druhej sade dát, ktoré nie su z normálneho rozdelenia, použijeme iba asymptotické testy 1.34, 1.44 a 1.51.

3.1 Simulačné dáta

Použijeme 2 sady dát, v prvej budú dáta z normálneho dvojrozmerného rozdelenia s parametrami $EX = 1, EY = 1$,$\sigma_x = 1$, $\sigma_y = 1$, $\rho \in \{0; 0{,}05; 0{,}15; 0{,}4; 0{,}7\}$. Veľkosť výberu bude $n \in \{20{,}70{,}200\}$

Pre 1000 simulacii o výbere veľkosti n=20, $\rho = 0$ dostávame

Tabuľka 3.1: Empirická hladina testu n=20

$\rho = 0$	Empirická hladina testu
Pearson	0,048
Spearman	0,039
Kendal	0,044

Pre 1000 simulácii o výbere veľkosti n=200, $\rho = 0$ dostávame

Tabuľka 3.2: Empirická hladina testu n=200

$\rho = 0$	Empirická hladina testu
Pearson	0,041
Spearman	0,051
Kendal	0,049

3.2 Sila testu hypotézy nezávislosti

Výsledkom je tabuľka, ktorá závisí na počte pozorovaní a koeficiente r a určuju odhadovanú hladinu testu.

Tabuľka 3.3: Sila testu, výber z normálneho rozdelenia

	Metóda	r=0,05	r=0,15	r=0,40	r=0,70
n=20	Pearson	0,050	0,097	0,425	0,953
	Spearman	0,049	0,089	0,374	0,924
	Kendal	0,038	0,084	0,358	0,921
n=70	Pearson	0,083	0,264	0,939	1
	Spearman	0,074	0,245	0,921	1
	Kendal	0,077	0,248	0,923	1
n=200	Pearson	0,109	0,575	1	1
	Spearman	0,096	0,534	1	1
	Kendal	0,094	0,531	1	1

Veľkosť sily testu hypotézy o nezávislosti závisí od počtu pozorovaní a koeficientu r, ktorý predpokladáme, že ním budú dáta korelované.

Nedá sa povedať, že pre dáta, ktoré pochádzajú z normálneho rozdelenia definovaného na začiatku, že by bol niektorý test o hypotéze nezávislosti výrazne lepší ako druhý. Ale pozorujeme, že sila testu hypotézy o nezávislosti pri použití Pearsonovho korelačného koeficientu je o trochu väčšia jak u ostatných testov

Ďalšia časť. Máme výber, ktorý nepochádzajú z normálneho rozdelenia. Dáta vznikli transformáciou z normálneho dvojrozmerného rozdelenia umocnením jednej zložky na tretiu.

Výsledkom je opäť tabuľka, ktorá závisí na počte pozorovaní a koeficiente r a určujú odhadovanú silu testu.

Tabuľka 3.4: Sila testu, výber nie je z normálneho rozdelenia

	Metóda	r=0,05	r=0,15	r=0,40	r=0,70
n=20	Pearson	0,055	0,077	0,296	0,842
	Spearman	0,049	0,089	0,374	0,924
	Kendal	0,038	0,084	0,358	0,921
n=70	Pearson	0,06	0,157	0,782	1
	Spearman	0,074	0,245	0,921	1
	Kendal	0,077	0,248	0,923	1
n=200	Pearson	0,078	0,386	0,998	1
	Spearman	0,096	0,534	1	1
	Kendal	0,094	0,531	1	1

Z tabuľky 3.4 pozorujeme, že sila testu hypotézy nezávislosti pri použití Pearsonovho korelačného koeficientu je menšia ako u Spearmanovho alebo Kendallovho korelačného koeficientu. Preto ak budeme testovať hypotézu o nezávislosti nad dátami, ktoré budeme vedieť, že nepochádzajú z dvojrozmerného normálneho rozdelenia, tak by sme mali použiť test s vyššou silou pri danej hladine α.

Záver

Práci boli predstavené 3 korelačné koeficienty. Zoznámili sme sa s ich populačnými a výberovými verziami a ich použitím na testovanie hypotézy nezavislosti.

Následne bolo ukázané praktické použitie korelačných koeficientov na odhadnutie sily korelácie na dátach, boli ukázané rozdiely v korelačných koeficientoch v závislosti na aký druh dát ich aplikujeme.

V tretej časti sme sa pozreli na simulačné štúdie testu hypotézy nezávislosti a počítali sme silu testu, ktorá keď výber pochádzal z normálneho rozdelenia, tak najvyššiu silu dosahoval Pearsonov korelačný koeficient, ale keď dáta neboli z normálneho rozdelenia, tak vačšiu silu testu hypotézy nezavislosti mali Spearmanov a Kendallov korelačný koeficient.

Zoznam použitej literatúry

Air quality data set. http://archive.ics.uci.edu/ml/datasets/Air+Quality. Accessed: 2019-05-17.

Matematická statistika 1,poznámky k přednášce. https://www.karlin.mff.cuni.cz/~omelka/Soubory/nmsa331/ms1.pdf. Accessed: 2020-06-27.

Anděl, J. (2007). *Statistické metody*. Čtvrté přepracované vydání. Matfyzpress, Praha. ISBN 80-7378-003-8.

Athreya, K. B. a Lahiri, S. N. (2006). *Measure theory and probability theory*. Springer Science & Business Media. ISBN 0387354344.

Hazewinkel, M. (1989). *Encyclopaedia of mathematics. Volume 3. D–Feynman measure*. Kluwer. ISBN 978-1-55608-002-9.

Kendall, M. G. (1970). *Rank correlation methods*. Griffin. ISBN 0852641990.

Nelsen, R. B. (2006). *An introduction to copulas*. Springer Science & Business Media. ISBN 0387286594.

Stuart, A. a Ord, K. (2010). *Kendall's Advanced Theory of Statistics, Distribution Theory*. Wiley. ISBN 9780470665305.

Zvára, K. a Štepán, J. (1997). Pravdepodobnost a matematická statistika (1. vydání). *Matfyzpress, MFF UK, Praha*.

Zoznam obrázkov

Zoznam tabuliek