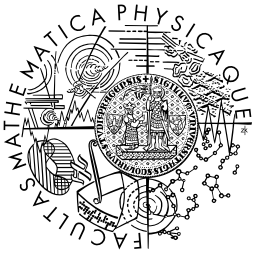


MATEMATICKO-FYZIKÁLNÍ FAKULTA
Univerzita Karlova

BAKALÁŘSKÁ PRÁCE

Vendula Rusá

Intervalový odhad korelačního koeficientu

Katedra pravděpodobnosti a matematické statistiky

Vedoucí bakalářské práce: doc. RNDr. Arnošt Komárek, Ph.D.
Studijní program: Obecná matematika

Praha 2024

Prohlašuji, že jsem tuto bakalářskou práci vypracoval(a) samostatně a výhradně s použitím citovaných pramenů, literatury a dalších odborných zdrojů. Tato práce nebyla využita k získání jiného nebo stejného titulu.

Beru na vědomí, že se na moji práci vztahují práva a povinnosti vyplývající ze zákona č. 121/2000 Sb., autorského zákona v platném znění, zejména skutečnost, že Univerzita Karlova má právo na uzavření licenční smlouvy o užití této práce jako školního díla podle §60 odst. 1 autorského zákona.

V ..... dne ....

Podpis autora

Chtěla bych tímto poděkovat doc. RNDr. Arnoštu Komárkovi, Ph.D. za jeho vedení mé bakalářské práce. Především bych chtěla vyzdvihnout empatické chování, trpělivost, rychlé reakce, možnost upravení obsahu práce dle mých představ a připomínky jak k obsahu tak i formální úpravě práce.

Dále mé velké díky patří mé rodině a Martinovi Hruškovi za obrovskou psychickou podporu a Aničce Švarcové za možnost sdílení našich strastí na cestě k napsání bakalářské práce.

Název práce: Intervalový odhad korelačního koeficientu

Autor: Vendula Rusá

Katedra: Katedra pravděpodobnosti a matematické statistiky

Vedoucí bakalářské práce: doc. RNDr. Arnošt Komárek, Ph.D., Katedra pravděpodobnosti a matematické statistiky

Abstrakt: Korelační koeficienty jsou standardní mírou vztahu mezi dvěma náhodnými veličinami. V práci si představíme různé metody pro konstrukci intervalového odhadu o spolehlivosti $(1 - \alpha)$ pro Pearsonův a Kendallův korelační koeficient. Zaměříme se na Fisherovu metodu z-transformace a dvě metody založené na empirické věrohodnosti pro Pearsonův korelační koeficient. Pro Kendallův korelační koeficient uvedeme dvě metody vycházející z vlastností funkce vlivu pro Kendallův korelační koeficient, z nichž jedna je rovněž založená na empirické věrohodnosti. Přidanou hodnotou metod založených na empirické věrohodnosti je jejich vhodnost i pro neznámé dvojrozměrné rozdělení. Nakonec provedeme simulační studii, kde porovnáme rozebrané metody z pohledu pravděpodobnosti pokrytí a průměrné délky intervalů spolehlivosti pro konečné rozsahy.

Klíčová slova: korelační koeficient, intervalový odhad, Kendallův korelační koeficient, Pearsonův korelační koeficient, empirická věrohodnost

Title: Interval estimation of the correlation coefficient

Author: Vendula Rusá

Department: Department of Probability and Mathematical Statistics

Supervisor: doc. RNDr. Arnošt Komárek, Ph.D., Department of Probability and Mathematical Statistics

Abstract: Correlation coefficients are a standard measure of the relationship between two random variables. In this paper, we will present various methods for constructing a $(1 - \alpha)$ level confidence interval for Pearson and Kendall correlation coefficients. We focus on Fisher's z-transformation method and two methods based on empirical likelihood for the Pearson correlation coefficient. For the Kendall correlation coefficient, we will present two methods based on the properties of the influence function for the Kendall correlation coefficient, one of which is also based on empirical likelihood. The added value of the methods based on empirical likelihood is their suitability even for the unknown bivariate distributions. Finally, we conduct a simulation study where we compare the discussed methods in terms of coverage probabilities and average length of confidence intervals for finite ranges.

Keywords: correlation coefficient, confidence interval, Kendall correlation coefficient, Pearson correlation coefficient, empirical likelihood

Obsah

Úvod

Korelační koeficienty jsou nástrojem k měření vztahu mezi dvěma náhodnými veličinami. Nejčastěji se používá Pearsonův korelační koeficient, který velmi dobře odhaluje především lineární závislost, či Kendallův korelační koeficient. V této práci se budeme zabývat metodami intervalových odhadů právě pro tyto dva korelační koeficienty.

První kapitola nám představí definice těchto korelačních koeficientů, jejich výběrové odhady a ilustrace několika příkladů. Na závěr této kapitoly si uvedeme odvození intervalového odhadu pro korelační koeficient pomocí Fisherovy metody z-transformace.

Ve druhé kapitole rozebereme neparametrickou věrohodnost a její základní vlastnost. Ukážeme si spojitost mezi neparametrickou věrohodností a intervalovými odhady Pearsonova korelačního koeficientu založenými na metodě věrohodnosti, které byly poprvé uvedeny v článku (Hu, Jung a Qin, 2020).

Článek (Hu a kol., 2020) si vyžádal zkoumání metody věrohodnosti založené na funkci vlivu pro Kendallův korelační koeficient, kterého se zhostili (Huang a Qin, 2022) a navíc ještě odvodili normální aproximační metodu pro Kendallův korelační koeficient. Ve třetí kapitole můžeme tedy nalézt definici a rozbor vlastností funkce vlivu Kendallova korelačního koeficientu a úpravu metody věrohodnosti založené na funkci vlivu pro Kendallův korelační koeficient. Obě části nás dovedou ke složitějším tvrzením, jejichž důkazy si podrobně odvodíme. Z obou tvrzení následně vyplývá tvar intervalového odhadu Kendallova korelačního koeficientu.

Čtvrtá kapitola se zabývá simulační studií, která zkoumá pravděpodobnost pokrytí a průměrnou délku intervalových odhadů jednotlivých metod pro konečné rozsahy výběru, různá rozdělení a odlišné hodnoty korelačního koeficientu.

1. Základní pojmy a ilustrace

V této kapitole si představíme základní pojmy, zavedeme značení, ukážeme si několik grafických znázornění a nakonec se podíváme na hojně užívanou metodu intervalového odhadu korelačního koeficientu: Fisherovu z-transformaci.

1.1 Základní pojmy

Definice 1 (korelační koeficient)**.** *Nechť X a Y jsou náhodné veličiny s konečnými druhými momenty,* $\operatorname{var} X > 0$ *a* $\operatorname{var} Y > 0$. *Pak definujeme* korelační koeficient X a Y *jako*

$$\rho = \frac{\operatorname{cov}(X,Y)}{\sqrt{\operatorname{var} X \operatorname{var} Y}}.$$

Korelační koeficient nabývá hodnot $[-1,1]$. Přičemž $\rho = 1$, pokud $Y = a + bX$ s pravděpodobností 1, kde $a \in \mathbb{R}$ a $b > 0$. Obdobně $\rho = -1$, pokud $Y = a + bX$ s pravděpodobností 1, kde $a \in \mathbb{R}$ a $b < 0$. Jsou-li navíc náhodné veličiny X a Y nezávislé, pak zřejmě $\rho = 0$.

Dále pro $\begin{pmatrix} X_1 \\ Y_1 \end{pmatrix}, \begin{pmatrix} X_2 \\ Y_2 \end{pmatrix}, ..., \begin{pmatrix} X_n \\ Y_n \end{pmatrix}$ náhodný výběr z dvojrozměrného rozdělení $\begin{pmatrix} X \\ Y \end{pmatrix}$ značíme

$$\overline{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i, \qquad \overline{Y}_n = \frac{1}{n}\sum_{i=1}^{n} Y_i, \qquad S_X^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X}_n)^2,$$

$$S_Y^2 = \frac{1}{n-1}\sum_{i=1}^{n}(Y_i - \overline{Y}_n)^2 \qquad a \qquad S_{XY} = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X}_n)(Y_i - \overline{Y}_n).$$

Definice 2 (Pearsonův výběrový korelační koeficient)**.** Výběrový korelační koeficient *definujeme jako*

$$r_n = \frac{S_{XY}}{\sqrt{S_X^2 S_Y^2}}.$$

Ze Schwarzovy nerovnosti plyne $-1 \leq r_n \leq 1$. Chceme-li zkoumat lineární závislost mezi náhodnými veličinami z dvojrozměrného rozdělení, obvykle použijeme Pearsonův výběrový korelační koeficient. V případech, kdy potřebujeme zkoumat složitější vztah, náš výběr obsahuje odlehlá pozorování nebo známe pouze pořadí a ne přesné hodnoty náhodných veličin, můžeme využít Kendallův výběrový korelační koeficient.

Nejprve si definujeme teoretický Kendallův korelační koeficient a poté si ukážeme jeho konzistentní odhad vytvořený pomocí náhodného výběru.

Definice 3 (Kendallův korelační koeficient)**.** *Nechť* $\begin{pmatrix} X \\ Y \end{pmatrix}$ *je dvojrozměrný náhodný vektor s distribuční funkcí $F_0(x,y)$, pak číslo*

$$\rho_K(F_0) = \mathsf{E}_{F_0}[sign\,((X_1 - X_2)(Y_1 - Y_2))] = 2P_{F_0}[(X_1 - X_2)(Y_1 - Y_2) > 0] - 1,$$

kde $\begin{pmatrix} X_1 \\ Y_1 \end{pmatrix}$ *a* $\begin{pmatrix} X_2 \\ Y_2 \end{pmatrix}$ *jsou dvě nezávislé kopie* $\begin{pmatrix} X \\ Y \end{pmatrix}$*, nazýváme* Kendallův korelační koeficient.

Definice 4 (Kendallův výběrový korelační koeficient)**.** Kendallův výběrový korelační koeficient *je definován vzorcem*

$$\tau_n = \frac{2}{n(n-1)} \sum_{i<j} \sum sign\left((X_i - X_j)(Y_i - Y_j)\right).$$

Nechť $R_1, R_2, ..., R_n \in \{1, ..., n\}$ jsou pořadí veličin $X_1, X_2, ..., X_n$ a $Q_1, Q_2, ..., Q_n \in \{1, ..., n\}$ jsou pořadí veličin $Y_1, Y_2, ..., Y_n$. Neznáme-li přesné hodnoty náhodných veličin, můžeme zřejmě Kendallův výběrový korelační koeficient ekvivalentně zapsat jako

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

Stejně jako Pearsonův korelační koeficient $\tau_n \in [-1, 1]$. Pokud pořadí jsou totožná, tak $\tau_n = 1$. Naopak $\tau_n = -1$, pokud pořadí jednoho je přesným opakem druhého.

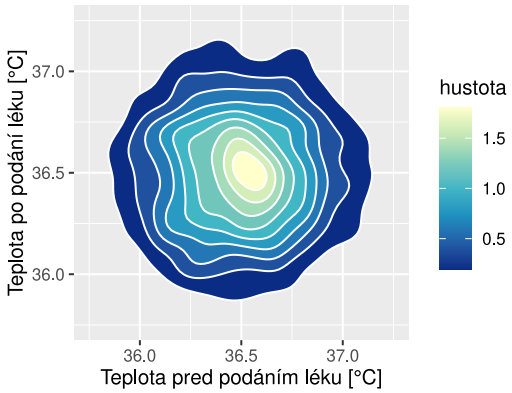
1.2 Ilustrace

Korelační koeficient měří vztah dvou náhodných veličin. Tento vztah můžeme někdy také pozorovat pouhým okem například na grafu jejich sdružené hustoty. Nyní si ukážeme zobrazení odhadu hustoty dvojrozměrného rozdělení náhodného vektoru získané pomocí náhodného výběru o rozsahu 2000 (resp. 365 pro 1.1b) pro pět příkladů, jimiž jsou:

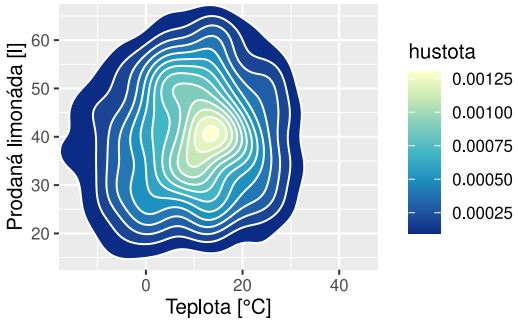
- tělesná teplota člověka před podáním léku a tělesná teplota člověka po podání léku 1.1a,
- teplota vzduchu a objem prodeje limonády v malém podniku 1.1b,
- výška a váha ženy 1.2a,
- váha běžce a jeho rychlost při běhu na 100 metrů 1.2b,
- rychlost auta a jeho brzdná dráha 1.3.

Na obrázku 1.1a můžeme pozorovat, že ačkoli libovolně zvolíme teplotu před podáním léku od $36\,°C$ do $37\,°C$, tak můžeme očekávat stejný rozsah hodnot teploty po podání léku. Zároveň z obrázku 1.1b můžeme předpokládat, že počet prodaných litrů limonády v malém podniku dosáhne čísla mezi dvaceti a padesáti pěti, přestože se bude teplota ovzduší pohybovat od $-5\,°C$ do $25\,°C$. Z toho, je zřejmé, že lineární závislost je zde velmi slabá pro oba případy. Pearsonův výběrový korelační koeficient je roven 0.01, resp. 0.05. Kendallův korelační koeficient nabývá hodnot přibližně 0.00, resp. 0.03.

Na grafu 1.2a sledujeme, že se zvyšující se výškou ženy můžeme očekávat vyšší váhu. Naopak z 1.2b vyšší váha muže znamená, že jeho dosažená rychlost v běhu na 100 metrů bude nejspíše nižší. Hodnoty Pearsonova výběrového koeficientu, resp. Kendallova korelačního koeficientu, v těchto případech jsou 0.76, resp. 0.55, a -0.89, resp. -0.70.

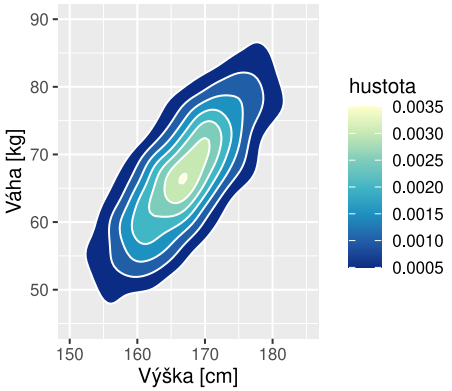


(a) Teploty před a po podání léku.

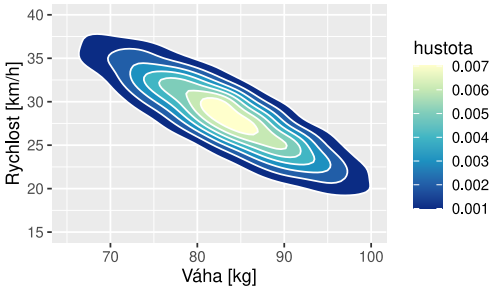


(b) Teplota a prodej limonády v malém podniku.

Obrázek 1.1: Odhady hustoty dvojrozměrného rozdělení vektoru dvou nezávislých náhodných veličin.

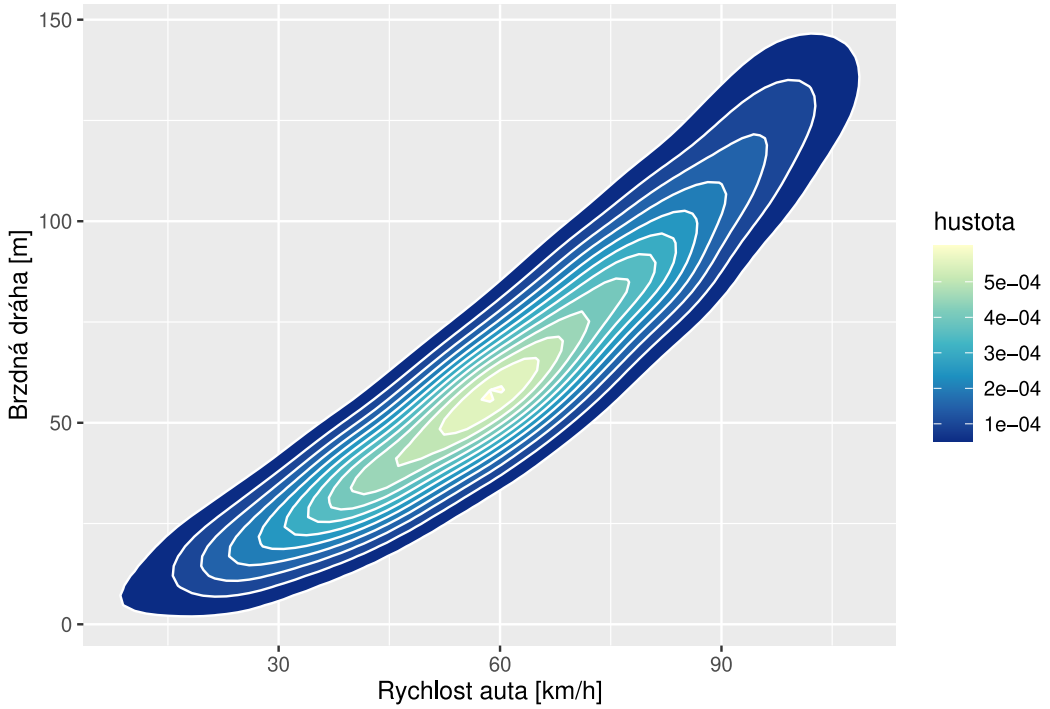


(a) Výška a váha žen.



(b) Váha a rychlost běžců při běhu na 100 metrů.

Obrázek 1.2: Odhady hustoty dvojrozměrného náhodného vektoru.



Obrázek 1.3: Rychlost a brzdná dráha auta.

Nakonec na obrázku 1.3 máme znázorněnou brzdnou dráhu auta v závislosti na jeho rychlosti před začátkem brzdění. Odhady korelačního koeficientu zde nabývají hodnot 0.97 pro Pearsonův korelační koeficient a 0.92 pro Kendallův korelační koeficient. Zároveň se dle obrázku můžeme domnívat, že závislost nebude lineární, ale bude odpovídat nějaké složitější funkci. V tomto případě bychom pro sestavení intervalového odhadu korelačního koeficientu využili Kendallův korelační koeficient.

1.3 Fisherova z-transformace

Fisher navrhl metodu z-transformace

$$Z(r_n) = \operatorname{arctgh}(r_n) = \frac{1}{2} \log \frac{1 + r_n}{1 - r_n}$$

pro odvození intervalového odhadu korelačního koeficientu, které si ukážeme v této kapitole.

Pro začátek budeme vycházet z následujícího tvrzení:

Tvrzení 1 (Anděl, 2011, strana 95–96)**.** *Pro výběrový korelační koeficient r_n náhodného výběru z dvojrozměrného normálního rozdělení s korelačním koeficientem ρ o rozsahu n platí*

$$\sqrt{n} \cdot r_n \xrightarrow[n \longrightarrow \infty]{d} \mathcal{N}\left(\rho, (1 - \rho^2)^2\right).$$

Z něj můžeme nyní odvodit tvar intervalu o spolehlivosti $1 - \alpha$ pro ρ. Nejprve jednoduchou úpravou vzorce z tvrzení 1, odečtením ρ od r_n, dostaneme, že

$$\sqrt{n}(r_n - \rho) \xrightarrow[n \longrightarrow \infty]{d} \mathcal{N}\left(0, (1 - \rho^2)^2\right).$$

K odvození z-transformace využijeme transformaci stabilizující rozptyl na předchozí výraz. Spočítáme tedy funkci g, která nám bude stabilizovat rozptyl, následujícím předpisem:

$$g(\rho) = c \int \frac{1}{\sqrt{(1-\rho^2)^2}}\, d\rho = c \int \frac{1}{1-\rho^2}\, d\rho = \frac{1}{2} c \ln \frac{1+\rho}{1-\rho},$$

kde c je libovolná konstanta. Transformace stabilizující rozptyl nám říká, že asymptoticky $\operatorname{var} g(\rho) = c^2$, a tak položíme $c = 1$, abychom získali konvergenci k normovanému normálnímu rozdělení. A tedy $g(\rho) = \frac{1}{2} \ln \frac{1+\rho}{1-\rho} = \operatorname{arctgh} \rho$. Poté pro $Z(r_n) = \frac{1}{2} \ln \frac{1+r_n}{1-r_n}$ za použití Δ-metody platí

$$\sqrt{n} \left(Z(r_n) - \frac{1}{2} \ln \frac{1+\rho}{1-\rho} \right) \xrightarrow[n \longrightarrow \infty]{d} \mathcal{N}(0,\, 1). \tag{1.1}$$

Odtud bychom již mohli vyjádřit intervalový odhad pro $g(\rho)$ a následovně pro ρ, ale podařilo se ukázat, že malou úpravou získáme lepší asymptotické vlastnosti pro intervalový odhad.

Přenásobíme konvergenci (1.1) výrazem $1 = \frac{\sqrt{n-3}}{\sqrt{n-3}}$ a použitím Cramérovy-Sluckého věty dostaneme, že

$$\sqrt{n-3} \left(Z(r_n) - \frac{1}{2} \ln \frac{1+\rho}{1-\rho} \right) \xrightarrow[n \longrightarrow \infty]{d} \mathcal{N}\left(0, 1\right), \tag{1.2}$$

jelikož $\frac{\sqrt{n}}{\sqrt{n-3}} \xrightarrow[n \longrightarrow \infty]{} 1$.

Odtud již získáme asymptotický interval o spolehlivosti $1 - \alpha$ pro $g(\rho)$:

$$\left(\frac{1}{2} \ln \frac{1+r_n}{1-r_n} - u_{1-\alpha/2} \frac{1}{\sqrt{n-3}}, \quad \frac{1}{2} \ln \frac{1+r_n}{1-r_n} + u_{1-\alpha/2} \frac{1}{\sqrt{n-3}} \right),$$

kde $u_{1-\alpha/2}$ je $(1-\alpha/2)$-kvantil normovaného normálního rozdělení. Nakonec tedy asymptotický interval o spolehlivosti $1-\alpha$ pro ρ vytvořený pomocí inverzní funkce

$$g^{-1}(z) = \operatorname{tgh}(\mathrm{z}) = \frac{e^{2z} - 1}{e^{2z} + 1}$$

vypadá následovně:

$$\left(\frac{\frac{1+r_n}{1-r_n} \exp\left(-\frac{2}{\sqrt{n-3}} u_{1-\alpha/2} \right) - 1}{\frac{1+r_n}{1-r_n} \exp\left(-\frac{2}{\sqrt{n-3}} u_{1-\alpha/2} \right) + 1}, \quad \frac{\frac{1+r_n}{1-r_n} \exp\left(\frac{2}{\sqrt{n-3}} u_{1-\alpha/2} \right) - 1}{\frac{1+r_n}{1-r_n} \exp\left(\frac{2}{\sqrt{n-3}} u_{1-\alpha/2} \right) + 1} \right). \tag{1.3}$$

2. Metody intervalového odhadu pro Pearsonův korelační koeficient

Empirická věrohodnost je silná neparametrická metoda vhodná ke konstrukci intervalových/oblastních odhadů neznámých parametrů. Intervalové odhady založené na empirické věrohodnosti byly vytvořeny též pro korelační koeficienty a v této kapitole se s nimi seznámíme.

2.1 Základní definice

Empirickou věrohodností se jako první zabýval Art B. Owen. Nejdříve se podíváme na její zavedení z knihy (Owen, 2001) a doplníme ho do souvislosti s maximálně věrohodným odhadem.

Symbolem $\mathcal{F}$ budeme značit prostor všech neklesajících zprava spojitých funkcí nabývajících hodnot od 0 do 1, neboli prostor všech distribučních funkcí. Mějme $X_1, X_2,..., X_n$ nezávislé reálné náhodné veličiny s distribuční funkcí $F_0 \in \mathcal{F}$.

Definice 5 (empirická distribuční funkce)**.** Empirickou distribuční funkcí *nazýváme*

$$F_n(x) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{[X_i \leq x]} \qquad pro -\infty < x < \infty.$$

Dále podle (Owen, 2001, str. 7-12) zavedeme neparametrickou věrohodnost.

Definice 6 (neparametrická věrohodnost)**.** Neparametrickou věrohodností distribuční funkce $F \in \mathcal{F}$ *nazveme*

$$\mathcal{L}(F) = \prod_{i=1}^{n} \left(F(X_i) - F(X_i-)\right).$$

Víme, že $\mathcal{L}(F) = 0$, když F odpovídá spojitému rozdělení. Navíc ukážeme, že empirická distribuční funkce maximalizuje neparametrickou věrohodnost.

Tvrzení 2. *Nechť F_n je empirická distribuční funkce $X_1, X_2,..., X_n$ a $F \in \mathcal{F}$ je libovolná distribuční funkce. Pokud $F \neq F_n$, potom $\mathcal{L}(F) < \mathcal{L}(F_n)$.*

Důkaz. Označme si $Z_1, Z_2,..., Z_m$ navzájem různé hodnoty $X_1, X_2,..., X_n$; n_j počet X_i rovných Z_j, neboli $n_j = \sum_{i=1}^{n} \mathbb{1}_{[X_i = Z_j]}$, a $p_j = F(Z_j) - F(Z_j-)$ pro $j \in \{1, 2,..., m\}$. Díky tomu si můžeme neparametrické věrohodnosti funkcí F a F_n zapsat následujícím způsobem:

$$\mathcal{L}(F) = \prod_{j=1}^{m} {p_j}^{n_j},$$

$$\mathcal{L}(F_n) = \prod_{j=1}^{n} \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{[X_i = X_j]} = \prod_{j=1}^{m} \left(\frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{[X_i = Z_j]}\right)^{n_j} = \prod_{j=1}^{m} \left(\frac{n_j}{n}\right)^{n_j}.$$

Pokud se nějaké $p_j = 0$, pak $\mathcal{L}(F) = 0 < \mathcal{L}(F_n)$. Předpokládejme tedy dále, že $p_j > 0$. Navíc z předpokladu $F \neq F_n$ máme, že alespoň jedno $p_j \neq \frac{n_j}{n}$. Stačí nám ověřit, že $\log \frac{\mathcal{L}(F)}{\mathcal{L}(F_n)} < 0$, místo $\mathcal{L}(F) < \mathcal{L}(F_n)$, jelikož tyto nerovnosti jsou ekvivalentní. Tedy

$$\log \frac{\mathcal{L}(F)}{\mathcal{L}(F_n)} = \log \frac{\prod_{j=1}^{m} {p_j}^{n_j}}{\prod_{j=1}^{m} \left(\frac{n_j}{n}\right)^{n_j}} = \sum_{j=1}^{m} \log \left(\frac{p_j}{\frac{n_j}{n}}\right)^{n_j} =$$

$$= \sum_{j=1}^{m} n_j \cdot \log \frac{p_j}{\frac{n_j}{n}} \overset{(1)}{<} \sum_{j=1}^{m} n_j \cdot \left(\frac{p_j}{\frac{n_j}{n}} - 1\right) = n \cdot \sum_{j=1}^{m} (p_j - 1) \overset{(2)}{\leq} 0.$$

Přičemž v (1) jsme využili toho, že $\log x \leq x - 1$ a rovnost zde nastane, když $x = 1$. Jelikož $\frac{p_j}{\frac{n_j}{n}} \neq 1$ alespoň pro jedno j, tak získáme ostrou nerovnost. (2) vychází z toho, že $p_j \in (0,1]$ a součet členů menších nebo rovných nule přenásobený kladným číslem je rovněž menší nebo rovný nule.

□

Jak jsme již na začátku kapitoly zmínili, tak neparametrickou věrohodnost můžeme využít na testování hypotéz a konstrukci intervalových odhadů. K tomu budeme potřebovat navíc poměr věrohodnosti.

Definice 7 (poměr věrohodnosti a log-věrohodnostní poměr)**.** *Pro rozdělení $F \in \mathcal{F}$ definujeme pomocí neparametrické věrohodnosti* poměr věrohodnosti *jako*

$$R(F) = \frac{\mathcal{L}(F)}{\mathcal{L}(F_n)}$$

a log-věrohodnostní poměr *následujícím vzorcem:*

$$l(F) = -2logR(F).$$

2.2 Plug-in metoda věrohodnosti

Zde si představíme plug-in intervalový odhad pro korelační koeficient z článku (Hu, Jung a Qin, 2020), který lze jednoduše použít v praxi.

Pomocí korelačního koeficientu ρ zjišťujeme korelaci dvou náhodných veličin. Mějme tedy dvojrozměrný náhodný výběr $\begin{pmatrix} X_1 \\ Y_1 \end{pmatrix}, \begin{pmatrix} X_2 \\ Y_2 \end{pmatrix}, ..., \begin{pmatrix} X_n \\ Y_n \end{pmatrix}$ z rozdělení $\begin{pmatrix} X \\ Y \end{pmatrix}$ s distribuční funkcí F_0, se skutečnými středními hodnotami $\begin{pmatrix} \mu_X \\ \mu_Y \end{pmatrix} = \begin{pmatrix} \mathsf{E}\ X \\ \mathsf{E}\ Y \end{pmatrix}$ a kladnými konečnými rozptyly $\begin{pmatrix} \sigma_X^2 \\ \sigma_Y^2 \end{pmatrix} = \begin{pmatrix} \operatorname{var} X \\ \operatorname{var} Y \end{pmatrix}$. Navíc využijeme toho, že $\begin{pmatrix} \sigma_X \\ \sigma_Y \end{pmatrix} = \begin{pmatrix} \sqrt{\sigma_X^2} \\ \sqrt{\sigma_Y^2} \end{pmatrix}$ značí směrodatné odchylky rozdělení X a Y, a uvažujeme zde analogie definic z kapitoly 2.1 pro dvojrozměrný náhodný výběr. Dále budeme využívat značení

$$F_0(\boldsymbol{w}-) = P_0(X < x, Y < y), \quad \text{kde } \boldsymbol{w} = (x, y)^T$$

a P_0 je pravděpodobnostní míra jednoznačně určená danou distribuční funkcí F_0.

Jelikož nás zajímá pouze korelační koeficient ρ a nepotřebujeme znát přesné rozdělení F_0, tak upravíme neparametrickou věrohodnost, aby nezávisela na $F \in \mathcal{F}$. Nechť $\boldsymbol{W_i} = \begin{pmatrix} X_i \\ Y_i \end{pmatrix}$ pro $i = 1,...,n$. Označme si $\boldsymbol{p_0} = (p_{0,1},...,p_{0,n})$, kde

$$p_{0,i} = F_0(\boldsymbol{W_i}) - F_0(\boldsymbol{W_i}-) \quad \text{pro } i = 1,...,n.$$

Každé $p_{0,i}$ nám znázorňuje pravděpodobnost výběru pozorování $\boldsymbol{W_i}$, a proto $\sum_{i=1}^n p_{0,i} = 1$. Z toho získáváme, že budeme hledat supremum $\prod_{i=1}^n p_i$ přes pravděpodobnostní vektory $\boldsymbol{p} = (p_1,...,p_n)$. Protože pro korelační koeficient ρ platí následující rovnost:

$$\begin{aligned} \mathsf{E}\left(\frac{X-\mu_X}{\sigma_X}\cdot\frac{Y-\mu_Y}{\sigma_Y}\right) - \rho &= \mathsf{E}\left(\frac{XY-\mu_X Y - X\mu_Y + \mu_X\mu_Y}{\sigma_X\cdot\sigma_Y}\right) - \rho \\ &= \frac{\mathsf{E}\,(XY) - \mu_X\mu_Y}{\sigma_X\cdot\sigma_Y} - \mathsf{E}\;\rho = \mathsf{E}\left(\frac{\operatorname{cov}(X,Y)}{\sqrt{\operatorname{var} X \operatorname{var} Y}} - \rho\right) = 0, \end{aligned} \quad (2.1)$$

tak $\boldsymbol{p}$ musí splňovat $\sum_{i=1}^n p_i(V(\boldsymbol{W_i}) - \rho) = 0$, kde $V(\boldsymbol{W_i}) = \frac{X_i-\mu_X}{\sigma_X}\cdot\frac{Y_i-\mu_Y}{\sigma_Y}$ pro $i = 1,...,n$.

Empirickou věrohodnost ρ můžeme definovat následujícím způsobem:

$$L_0(\rho) = \sup_{\boldsymbol{p}}\left\{\prod_{i=1}^n p_i : p_i \geq 0, \sum_{i=1}^n p_i = 1, \sum_{i=1}^n p_i(V(\boldsymbol{W_i}) - \rho) = 0\right\}, \quad (2.2)$$

kde $\mathbf{p} = (p_1,...,p_n)$ je pravděpodobnostní vektor.

Skutečné střední hodnoty $(\mu_X, \mu_Y)^T$ a skutečné směrodatné odchylky $(\sigma_X, \sigma_Y)^T$ jsou v praxi neznámé, ale $(\mu_X, \mu_Y)^T$ mohou být odhadnuty příslušnými výběrovými průměry $(\overline{X}_n, \overline{Y}_n)^T$ a $(\sigma_X, \sigma_Y)^T$ lze odhadnout odmocninami výběrových rozptylů $(S_X, S_Y)^T$. Po vložení těchto odhadů do předchozího vzorce dostáváme následující plug-in empirickou věrohodnost pro korelační koeficient.

Definice 8 (plug-in empirická věrohodnost pro korelační koeficient)**.** Plug-in empirickou věrohodností pro korelační koeficient ρ *budeme nazývat*

$$\widehat{L}(\rho) = \sup_{\boldsymbol{p}}\left\{\prod_{i=1}^n p_i : p_i \geq 0, \sum_{i=1}^n p_i = 1, \sum_{i=1}^n p_i(\widehat{V}(\boldsymbol{W_i}) - \rho) = 0\right\},$$

kde $\widehat{V}(\boldsymbol{W_i}) = \frac{X_i-\overline{X}_n}{S_X}\cdot\frac{Y_i-\overline{Y}_n}{S_Y}, \quad i = 1,...,n.$

Vyjádření pro p_i získáme použitím metody Lagrangeových multiplikátorů na logaritmickou plug-in věrohodnostní funkci, kde zároveň využijeme toho, že pravá strana poslední rovnosti z podmínky je po přenásobení hodnotou $(-n)$ stále rovna nule. Jinak řečeno nahradíme ji podmínkou $-n\sum_{i=1}^n p_i(\widehat{V}(\boldsymbol{W_i}) - \rho) = 0$. Hledáme tedy maximum funkce $\sum_{i=1}^n \log p_i$ na množině, kde $p_i \geq 0$ a platí rovnosti $\sum_{i=1}^n p_i = 1$, $-n\sum_{i=1}^n p_i(\widehat{V}(\boldsymbol{W_i}) - \rho) = 0$. Lagrangeova funkce značená $\boldsymbol{\mathcal{L}}$ v tomto případě vypadá následujícím způsobem:

$$\boldsymbol{\mathcal{L}}(p_1, p_2,..., p_n, \lambda_1, \lambda_2) = \sum_{i=1}^n \log p_i + \lambda_1\left(\sum_{i=1}^n p_i - 1\right) - \lambda_2 n\left(\sum_{i=1}^n p_i\left(\widehat{V}(\boldsymbol{W_i}) - \rho\right)\right). \quad (2.3)$$

Dále využijeme toho, že v bodě extrému jsou všechny parciální derivace funkce nulové.

$$\begin{aligned}
\frac{\partial \boldsymbol{\mathcal{L}}}{\partial p_1} &= \frac{1}{p_1} + \lambda_1 - \lambda_2 n \left(\widehat{V}(\boldsymbol{W_1}) - \rho\right) = 0 \\
\frac{\partial \boldsymbol{\mathcal{L}}}{\partial p_2} &= \frac{1}{p_2} + \lambda_1 - \lambda_2 n \left(\widehat{V}(\boldsymbol{W_2}) - \rho\right) = 0 \\
&\vdots \\
\frac{\partial \boldsymbol{\mathcal{L}}}{\partial p_n} &= \frac{1}{p_n} + \lambda_1 - \lambda_2 n \left(\widehat{V}(\boldsymbol{W_n}) - \rho\right) = 0
\end{aligned}$$

Rovnice přenásobíme příslušným p_i pro všechna $i = 1,...,n$ a sečteme je.

$$\left.\begin{aligned}
1 + p_1\lambda_1 - p_1\lambda_2 n \left(\widehat{V}(\boldsymbol{W_1}) - \rho\right) = 0 \\
1 + p_2\lambda_1 - p_2\lambda_2 n \left(\widehat{V}(\boldsymbol{W_2}) - \rho\right) = 0 \\
\vdots \\
1 + p_n\lambda_1 - p_n\lambda_2 n \left(\widehat{V}(\boldsymbol{W_n}) - \rho\right) = 0
\end{aligned}\right\} \quad + \tag{2.4}$$

Získáme díky tomu rovnici:

$$n + \lambda_1 \sum_{i=1}^{n} p_i - \lambda_2 n \left(\sum_{i=1}^{n} p_i \left(\widehat{V}(\boldsymbol{W_i}) - \rho\right)\right) = 0,$$

kterou můžeme upravit do tvaru $n + 1 \cdot \lambda_1 + 0 \cdot \lambda_2 = 0$, jelikož hledáme řešení, které splňuje $\sum_{i=1}^{n} p_i = 1,\ -n\sum_{i=1}^{n} p_i(\widehat{V}(\boldsymbol{W_i}) - \rho) = 0$. Máme tedy, že $\lambda_1 = -n$.

Dosadíme řešení pro λ_1 do 2.4 a získáváme pro všechna $i = 1,...,n$ rovnost

$$1 - np_i - np_i\lambda_2 \left(\widehat{V}(\boldsymbol{W_i}) - \rho\right) = 0,$$

ze které si jednoduše vyjádříme, že

$$p_i = \frac{1}{n}\{1 + \lambda_2(\widehat{V}(\boldsymbol{W_i}) - \rho)\}^{-1}, \qquad i = 1,..,n.$$

A nakonec λ_2 získáme z podmínky $p_i = \frac{1}{n}\{1 + \lambda_2(\widehat{V}(\boldsymbol{W_i}) - \rho)\}^{-1} \geq 0$ a z řešení následující rovnice:

$$\frac{1}{n}\sum_{i=1}^{n} \frac{\widehat{V}(\boldsymbol{W_i}) - \rho}{1 + \lambda_2(\widehat{V}(\boldsymbol{W_i}) - \rho)} = 0. \tag{2.5}$$

Vzhledem k tomu, že $\sum_{i=1}^{n} p_i = 1$, tak $\prod_{i=1}^{n} p_i$ nabývá svého maxima n^{-n}, když $p_i = n^{-1}$ pro $i = 1,..,n$. Tedy plug-in věrohodnostní poměr pro ρ je dán vzorcem:

$$R(\rho) = \prod_{i=1}^{n} \frac{p_i}{n^{-1}} = \prod_{i=1}^{n}(np_i) = \prod_{i=1}^{n} \{1 + \lambda_2(\widehat{V}(\boldsymbol{W_i}) - \rho)\}^{-1}. \tag{2.6}$$

Odpovídající plug-in empirický log-věrohodnostní poměr pro ρ je tvaru

$$l(\rho) = -2\log R(\rho) = 2\sum_{i=1}^{n} \log\{1 + \lambda_2(\widehat{V}(\boldsymbol{W_i}) - \rho)\}. \tag{2.7}$$

Věta 3 (Hu a kol., 2020)**.** *Když ρ je skutečná hodnota korelačního koeficientu, potom asymptotické rozdělení $l(\rho)$ je škálované chí-kvadrát rozdělení o jednom stupni volnosti, tedy*

$$A \cdot l(\rho) \xrightarrow[n\to\infty]{d} \chi_1^2,$$

kde pomocná konstanta $A = \sigma_0^2/\sigma_V^2$ s

$$\sigma_V^2 = \operatorname{var}\left[\frac{X-\mu_X}{\sigma_X}\cdot\frac{Y-\mu_Y}{\sigma_Y} - \frac{1}{2}\rho\left(\left(\frac{X-\mu_X}{\sigma_X}\right)^2 + \left(\frac{Y-\mu_Y}{\sigma_Y}\right)^2\right)\right],$$

$$\sigma_0^2 = \operatorname{var}\left[\frac{X-\mu_X}{\sigma_X}\cdot\frac{Y-\mu_Y}{\sigma_Y}\right].$$

Důkaz. Podrobně rozepsaný důkaz lze najít v bakalářské práci (Farda, 2021). □

K intervalovému odhadu korelačního koeficientu ρ potřebujeme odhadnout pomocnou konstantu A. Nechť

$$A_{1i} = \frac{X_i - \overline{X}_n}{S_X}\cdot\frac{Y_i-\overline{Y}_n}{S_Y} - \frac{1}{2}r_n\left(\left(\frac{X_i-\overline{X}_n}{S_X}\right)^2 + \left(\frac{Y_i-\overline{Y}_n}{S_Y}\right)^2\right),$$

$$A_{2i} = \frac{X_i - \overline{X}_n}{S_X}\cdot\frac{Y_i-\overline{Y}_n}{S_Y},$$

$$\widehat{\sigma}_V^2 = \frac{1}{n}\sum_{i=1}^{n}\left(A_{1i} - \frac{1}{n}\sum_{i=1}^{n}A_{1i}\right)^2,$$

$$\widehat{\sigma}_0^2 = \frac{1}{n}\sum_{i=1}^{n}\left(A_{2i} - \frac{1}{n}\sum_{i=1}^{n}A_{2i}\right)^2.$$

Potom $\widehat{A} = \widehat{\sigma}_0^2/\widehat{\sigma}_V^2$ je konzistentní odhad pro pomocnou konstantu A a interval o spolehlivosti $(1-\alpha)$ založený na plug-in metodě věrohodnosti pro ρ můžeme zkonstruovat následujícím způsobem:

$$\{\rho : \widehat{A}\cdot l(\rho) \le \chi_1^2(1-\alpha)\},$$

kde $\chi_1^2(1-\alpha)$ je $(1-\alpha)$-kvantil χ_1^2.

Poznámka. Najít řešení rovnice 2.5 je výpočetně náročné. Alternativně můžeme dle (Lazar, 2021) minimalizovat přes $\mathbb{R}^2$ funkci

$$-\sum_{i=1}^{n}\log_*\{1+\lambda_2(\widehat{V}(\boldsymbol{W_i})-\rho)\}, \tag{2.8}$$

kde $\log_* z$ je pseudologaritmická funkce definovaná jako

$$\log_* z = \begin{cases}\log z, & \text{je-li } z \ge 1/n,\\ \log\frac{1}{n} - 1.5 - 2nz - \frac{(nz)^2}{2}, & \text{je-li } z \le 1/n.\end{cases} \tag{2.9}$$

2.3 Metoda věrohodnosti založená na funkci vlivu

Intervalový odhad založený na plug-in metodě věrohodnosti se používá jednoduše, ale je třeba odhadnout pomocnou konstantu A. K odstranění pomocné konstanty definujeme empirickou věrohodnost pro ρ založenou na funkci vlivu.

Nechť $\begin{pmatrix} X \\ Y \end{pmatrix}$ je dvojrozměrný vektor s distribuční funkcí F_0. Z kapitoly 2.1 použijeme značení pro náhodný výběr $\begin{pmatrix} X_1 \\ Y_1 \end{pmatrix}, \begin{pmatrix} X_2 \\ Y_2 \end{pmatrix}, ..., \begin{pmatrix} X_n \\ Y_n \end{pmatrix}$ z rozdělení $\begin{pmatrix} X \\ Y \end{pmatrix}$. Opět nepotřebujeme znát přesné rozdělení F_0, tudíž budeme usilovat o to, aby empirická věrohodnost nezávisela na tomto parametru. V článku (Hu a kol., 2020) pracují s funkcí vlivu dle následující definice:

Definice 9. Funkcí vlivu[1] pro ρ *nazýváme*

$$V_I(\boldsymbol{W_i}, \rho) = \left(\frac{X_i - \mu_X}{\sigma_X} \cdot \frac{Y_i - \mu_Y}{\sigma_Y} - \rho\right) - \frac{1}{2}\rho\left[\left(\left(\frac{X_i - \mu_X}{\sigma_X}\right)^2 - 1\right) + \left(\left(\frac{Y_i - \mu_Y}{\sigma_Y}\right)^2 - 1\right)\right]. \quad (2.10)$$

Obecně nám funkce vlivu znázorňuje, jaký bude rozdíl mezi funkcionálem se skutečným rozdělením a funkcionálem s rozdělením, jehož vychýlení od skutečného rozdělení jde limitně k nule. V našem případě uvažujeme ρ jako funkcionál se skutečným rozdělením a $\widehat{V}(\boldsymbol{W_i}) = \frac{X_i - \overline{X}_n}{S_X} \cdot \frac{Y_i - \overline{Y}_n}{S_Y}$ jako funkcionál s limitně nulovou změnou rozdělení pro $n \to \infty$. Podrobné odvození tvaru funkce vlivu pro ρ můžeme nalézt v důkazu Lemma 1 (i) z článku (Hu a kol., 2020, strana 35).

Nyní si definujeme empirickou věrohodnost pro ρ, ve které využijeme toho, že platí rovnost

$$\mathsf{E}\ V_I\left((X, Y)^T, \rho\right) = 0.$$

Definice 10. Empirickou věrohodnost založenou na funkci vlivu *definujeme následujícím způsobem:*

$$\mathcal{L}_I(\rho) = \sup_{\boldsymbol{p}} \left\{ \prod_{i=1}^{n} p_i : p_i \geq 0, \sum_{i=1}^{n} p_i = 1, \sum_{i=1}^{n} p_i \widehat{V}_I(\boldsymbol{W_i}, \rho) = 0 \right\},$$

kde $\boldsymbol{p} = (p_1, ..., p_n)$ *je pravděpodobnostní vektor a pro* $i = 1, ..., n$
$\widehat{V}_I(\boldsymbol{W_i}, \rho) = \left(\frac{X_i - \overline{X}_n}{S_X} \cdot \frac{Y_i - \overline{Y}_n}{S_Y} - \rho\right) - \frac{1}{2}\rho\left[\left(\left(\frac{X_i - \overline{X}_n}{S_X}\right)^2 - 1\right) + \left(\left(\frac{Y_i - \overline{Y}_n}{S_Y}\right)^2 - 1\right)\right].$

Vyjádření pro p_i získáme analogicky k odvození z kapitoly 2.2 pomocí metody Lagrangeových multiplikátorů. Získáme tedy, že

$$p_i = \frac{1}{n}\{1 + \lambda_I \widehat{V}_I(\boldsymbol{W_i}, \rho)\}^{-1}, \qquad i = 1, .., n,$$

kde λ_I řeší následující rovnici:

$$\frac{1}{n}\sum_{i=1}^{n} \frac{\widehat{V}_I(\boldsymbol{W_i}, \rho)}{1 + \lambda_I \widehat{V}_I(\boldsymbol{W_i}, \rho)} = 0. \quad (2.11)$$

[1] Anglicky influence function

Ze získaného řešení p_i a λ_I můžeme dále sestavit log-věrohodnostní poměr empirické věrohodnosti založené na funkci vlivu pro ρ jako

$$l_I(\rho) = 2\sum_{i=1}^{n} \log\{1 + \lambda_I \widehat{V}_I(\boldsymbol{W_i}, \rho)\}. \tag{2.12}$$

Navíc díky následující větě známe asymptotické rozdělení log-věrohodnostního poměru založeného na funkci vlivu, které se obejde bez škálovací konstanty, a můžeme tak i jednoduše sestavit intervalový odhad.

Věta 4 (Hu a kol., 2020)**.** *Když ρ je skutečná hodnota korelačního koeficientu, potom asymptotické rozdělení $l_I(\rho)$ je chí-kvadrát rozdělení o jednom stupni volnosti, tedy*

$$l_I(\rho) \xrightarrow[n\to\infty]{d} \chi^2_1.$$

Důkaz. Důkaz je analogický důkazu věty 3. □

Interval o spolehlivosti $(1-\alpha)$ získaný pomocí empirické věrohodnosti založené na funkci vlivu pro ρ tedy je

$$\{\rho : l_I(\rho) \leq \chi^2_1(1-\alpha)\}.$$

3. Metody odhadu pro Kendallův korelační koeficient

V návaznosti na (Hu, Jung a Qin, 2020) byl vydán článek (Huang a Qin, 2022) pojednávající o intervalových odhadech pro Kendallův korelační koeficient založených na funkci vlivu. V této kapitole si podrobněji rozebereme jednotlivé metody uvedené v článku a rozvedeme důkaz věty o konvergenci empirického log-věrohodnostního poměru založeného na funkci vlivu pro Kendallův korelační koeficient.

3.1 Normální aproximační metoda

Normální aproximační metoda využívá konvergence funkce vlivu pro Kendallův korelační koeficient, jejíž odvození si zde předvedeme. Rovněž si ukážeme tvar intervalového odhadu o spolehlivosti $1-\alpha$ této metody.

Nechť $\begin{pmatrix} X \\ Y \end{pmatrix}$ je dvojrozměrný vektor s distribuční funkcí $F_0(x,y)$ a konečnými středními hodnotami $\begin{pmatrix} \mu_X \\ \mu_Y \end{pmatrix}$, $\rho_K(F_0)$ Kendallův korelační koeficient.

Definice 11. Funkci vlivu pro Kendallův korelační koeficient $\rho_K(F_0)$ *definujeme jako*

$$F_{vl}\left((x,y),\rho_K,F_0\right) = \lim_{\varepsilon\to 0}\frac{\rho_K((1-\varepsilon)F_0+\varepsilon\Delta_{(x,y)})-\rho_K(F_0)}{\varepsilon},$$

kde $\Delta_{(x,y)}$ *je Diracova míra v bodě* $(x,y)\in\mathbb{R}^2$.

Funkce vlivu pro $\rho_K(F_0)$ nám tedy ukazuje, jaký efekt bude mít minimální změna v bodě (x,y) na Kendallův korelační koeficient pro náhodné vektory z rozdělení F_0. Tuto definici bychom mohli zobecnit na definici funkce vlivu pro jakýkoliv funkcionál R náhodného vektoru z rozdělení H používaný ve statistice pouze nahrazením R za ρ_K, H za F_0 a bodu $(x,y)\in\mathbb{R}^2$ bodem odpovídajícího rozměru.

Nechť dále $\begin{pmatrix} X_1 \\ Y_1 \end{pmatrix}, \begin{pmatrix} X_2 \\ Y_2 \end{pmatrix}, ..., \begin{pmatrix} X_n \\ Y_n \end{pmatrix}$ je náhodný výběr z rozdělení $\begin{pmatrix} X \\ Y \end{pmatrix}$.

Tvrzení 5. *Pro funkci vlivu v bodě* $(x,y)\in\mathbb{R}^2$ *pro Kendallův korelační koeficient* $\rho_K(F_0)$ *platí:*

(i) $F_{vl}\left((x,y),\rho_K,F_0\right) = -2\rho_K(F_0) + 2\,P_{F_0}\left((X-x)(Y-y)>0\right) - 2\,P_{F_0}\left((X-x)(Y-y)<0\right).$

(ii) Je-li rozdělení F_0 *navíc spojité, tak*

$$F_{vl}\left((x,y),\rho_K,F_0\right) = 2\left\{-\rho_K(F_0) + 2\,P_{F_0}\left([(X-x)(Y-y)]>0\right) - 1\right\}.$$

(iii) $\sqrt{n}\left(\frac{2}{n}\sum_{i=1}^{n}\mathbb{1}_{[(X_i-\mu_X)(Y_i-\mu_Y)>0]} - 1 - \rho_K(F_0)\right) \xrightarrow[n\to\infty]{d} \mathcal{N}(0,\sigma^2)$, *kde*

$$\sigma^2 = 4\,P((X-\mu_X)(Y-\mu_Y)>0)\cdot\{1-P((X-\mu_X)(Y-\mu_Y)>0)\}.$$

Důkaz.

(i) Tvrzení získáme následujícími úpravami:

$$
\begin{aligned}
&F_{vl}\left((x,y),\rho_K,F_0\right) = \lim_{\varepsilon\to 0}\frac{\rho_K((1-\varepsilon)F_0+\varepsilon\Delta_{(x,y)})-\rho_K(F_0)}{\varepsilon}\\
&= \lim_{\varepsilon\to 0}\left\{\frac{(1-\varepsilon)^2\,\mathsf{E}_{F_0}[sign(X-\tilde{X})(Y-\tilde{Y})]}{\varepsilon}\right.\\
&\quad + \frac{2\varepsilon(1-\varepsilon)\,\mathsf{E}_{F_0}[sign(X-x)(Y-y)]}{\varepsilon}\\
&\quad \left. + \frac{\varepsilon^2 sign[(x-x)(y-y)]-\mathsf{E}_{F_0}[sign(X-\tilde{X})(Y-\tilde{Y})]}{\varepsilon}\right\}\\
&= -2\,\mathsf{E}_{F_0}[sign(X-\tilde{X})(Y-\tilde{Y})]+2\,\mathsf{E}_{F_0}[sign(X-x)(Y-y)]\\
&= -2\rho_K(F_0)+2\,\mathsf{P}_{F_0}\left((X-x)(Y-y)>0\right)-2\,\mathsf{P}_{F_0}\left((X-x)(Y-y)<0\right),
\end{aligned}
$$

kde $(\tilde{X},\,\tilde{Y})^T$ je nezávislá kopie $(X,\,Y)^T$.

(ii) Je-li rozdělení F_0 navíc spojité, tak můžeme pokračovat s úpravami funkce dále.

$$
\begin{aligned}
&F_{vl}\left((x,y),\rho_K,F_0\right) = -2\rho_K(F_0)+2\,\mathsf{P}_{F_0}\left([(X-x)(Y-y)]>0\right)\\
&-2\left(1-\mathsf{P}_{F_0}\left([(X-x)(Y-y)]>0\right)\right)\\
&= 2\left\{-\rho_K(F_0)+2\,\mathsf{P}_{F_0}\left([(X-x)(Y-y)]>0\right)-1\right\}
\end{aligned}
$$

(iii) Plyne z přímo z centrální limitní věty, protože $2\cdot\mathbb{1}_{[(X_i-\mu_x)(Y_i-\mu_y)>0]}$ jsou nezávislé stejně rozdělené náhodné veličiny,

$$
\begin{aligned}
\mathsf{E}\;2\cdot\mathbb{1}_{[(X-\mu_X)(Y-\mu_Y)>0]} &= 2\,\mathsf{P}\left((X-\mu_X)(Y-\mu_Y)>0\right)\\
&= \frac{1}{2}F_{vl}\left((\mu_X,\mu_Y),\rho_K,F_0\right)+\rho_K(F_0)+1 = \rho_K(F_0)+1,
\end{aligned}
$$

jelikož zřejmě $F_{vl}\left((\mu_X,\mu_Y),\rho_K,F_0\right)=0$, a nakonec

$$
\begin{aligned}
\sigma^2 &= \operatorname{var}\left[2\cdot\mathbb{1}_{[(X-\mu_X)(Y-\mu_Y)>0]}\right]\\
&= 4\left(\mathsf{E}\left[\mathbb{1}_{[(X-\mu_X)(Y-\mu_Y)>0]}\right]^2-\left[\mathsf{E}\;\mathbb{1}_{[(X-\mu_X)(Y-\mu_Y)>0]}\right]^2\right)\\
&= 4\left(\mathsf{P}\left((X-\mu_X)(Y-\mu_Y)>0\right)-\left[\mathsf{P}\left((X-\mu_X)(Y-\mu_Y)>0\right)\right]^2\right)\\
&= 4\,\mathsf{P}\left((X-\mu_X)(Y-\mu_Y)>0\right)\cdot\left\{1-\mathsf{P}\left((X-\mu_X)(Y-\mu_Y)>0\right)\right\}.
\end{aligned}
$$

Důkaz je takto kompletní.

□

Za pomoci tvrzení 5 (iii) a nestranných konzistentních odhadů $(\overline{X}_n,\,\overline{Y}_n)^T$, $(S_X,\,S_Y)^T$ středních hodnot a rozptylů sestavíme intervalový odhad pro $\rho_K(F_0)$. Označme si

$$
\hat{\sigma}^2 = \frac{4}{n}\sum_{i=1}^{n}\mathbb{1}_{[(X_i-\overline{X}_n)(Y_i-\overline{Y}_n)>0]}\left\{1-\frac{1}{n}\sum_{i=1}^{n}\mathbb{1}_{[(X_i-\overline{X}_n)(Y_i-\overline{Y}_n)>0]}\right\}.
$$

Potom interval

$$\left(\frac{2}{n}\sum_{i=1}^{n}\mathbb{1}_{[(X_i-\overline{X}_n)(Y_i-\overline{Y}_n)>0]}-1-u_{1-\alpha/2}\sqrt{\frac{\hat{\sigma}^2}{n}},\right.$$
$$\left.\frac{2}{n}\sum_{i=1}^{n}\mathbb{1}_{[(X_i-\overline{X}_n)(Y_i-\overline{Y}_n)>S0]}-1+u_{1-\alpha/2}\sqrt{\frac{\hat{\sigma}^2}{n}}\right), \quad (3.1)$$

kde $u_{1-\alpha/2}$ značí $(1-\alpha/2)$-kvantil normovaného normálního rozdělení, je interval o spolehlivosti $(1-\alpha)$ pro $\rho_K(F_0)$.

3.2 Metoda věrohodnosti založená na funkci vlivu pro Kendallův korelační koeficient

S metodou věrohodnosti založené na funkci vlivu jsme se v této práci již setkali při odvozování intervalového odhadu pro Pearsonův korelační koeficient. Zde si ukážeme, že tuto metodu s několika úpravami můžeme využít i pro Kendallův korelační koeficient, a navíc si zde předvedeme důkaz věty, která je zásadní pro tvorbu intervalového odhadu pro $\rho_K(F_0)$.

V této části budeme postupovat jako v kapitole 2.3. Rovněž budeme používat stejné značení pro $\begin{pmatrix}X_1\\Y_1\end{pmatrix}, \begin{pmatrix}X_2\\Y_2\end{pmatrix},..., \begin{pmatrix}X_n\\Y_n\end{pmatrix}$ náhodný výběr z dvourozměrného rozdělení $\begin{pmatrix}X\\Y\end{pmatrix}$ s distribuční funkcí F_0 z kapitoly 2.1.

Z tvrzení 5 funkce vlivu pro $(x, y) = (\mu_X, \mu_Y)$ získáváme, že

$$\mathsf{E}\left(2\cdot\mathbb{1}_{[(X-\mu_X)(Y-\mu_Y)>0]}-1-\rho_K(F_0)\right)=0. \quad (3.2)$$

Buď $\boldsymbol{p} = (p_1,...,p_n)$ pravděpodobnostní vektor, tj. $p_i \geq 0$, $\sum_{i=1}^n p_i = 1$. Z (3.2) $\boldsymbol{p}$ musí navíc splňovat $\sum_{i=1}^n p_i(V_K(\boldsymbol{W_i}) - \rho_K(F_0)) = 0$, kde

$$V_K(\boldsymbol{W_i}) = 2\cdot\mathbb{1}_{[(X_i-\mu_X)(Y_i-\mu_Y)>0]}-1 \text{ pro } i = 1,...,n.$$

Empirickou věrohodnost $\rho_K(F_0)$ definujeme následovně:

$$L_{K,0}(\rho_K(F_0)) = \sup_{\boldsymbol{p}}\left\{\prod_{i=1}^{n}p_i : p_i\geq 0, \sum_{i=1}^{n}p_i = 1, \sum_{i=1}^{n}p_i(V_K(\boldsymbol{W_i})-\rho_K(F_0)) = 0\right\}. \quad (3.3)$$

Jelikož zpravidla neznáme $(\mu_X, \mu_Y)^T$, tak využijeme jejich odhadu $(\overline{X}_n, \overline{Y}_n)^T$ a sestavíme plug-in empirickou věrohodnost založenou na funkci vlivu jako

$$\widehat{L}_K(\rho_K(F_0)) = \sup_{\boldsymbol{p}}\left\{\prod_{i=1}^{n}p_i : p_i\geq 0, \sum_{i=1}^{n}p_i = 1, \sum_{i=1}^{n}p_i(\widehat{V}_K(\boldsymbol{W_i})-\rho_K(F_0)) = 0\right\},$$

kde $\widehat{V}_K(\boldsymbol{W_i}) = 2\cdot\mathbb{1}_{[(X_i-\overline{X}_n)(Y_i-\overline{Y}_n)>0]}-1$ pro $i = 1,...,n$. Rovněž za použití Lagrangeovy metody 2.3 získáme vyjádření pro p_i:

$$p_i = \frac{1}{n}\{1+\lambda_K(\widehat{V}_K(\boldsymbol{W_i})-\rho_K(F_0))\}^{-1}, \qquad i = 1,..,n,$$

a odpovídající empirický log-věrohodnostní poměr založený na funkci vlivu:

$$l_K(\rho_K(F_0)) = 2\sum_{i=1}^{n} \log\{1 + \lambda_K(\widehat{V}_K(\boldsymbol{W_i}) - \rho_K(F_0))\}.$$

V další části si dokážeme, že log-věrohodností poměr Kendallova korelačního koeficientu konverguje za jistých předpokladů v distribuci k chí-kvadrát rozdělení o jednom stupni volnosti. Než se podíváme přímo na důkaz této věty, tak si zavedeme značení, které budeme využívat a dokážeme si lemma, které využijeme v závěru důkazu věty na odvození konvergence v distribuci.

Symbolem $Z_n = o_p(a_n)$ pro náhodnou veličinu Z_n a odpovídající posloupnost konstant a_n rozumíme, že $\frac{Z_n}{a_n} \xrightarrow[n\to\infty]{\mathsf{P}} 0$. Dále se budeme zabývat avizovaným lemmatem, které nalezneme v práci (Huang a Qin, 2022). Především si zde rozvedeme důkaz jeho druhé části.

Lemma 6. *Pokud* $\mathsf{E}\,X < \infty$ *a* $\mathsf{E}\,Y < \infty$*, potom platí,že*

(i) $\frac{1}{\sqrt{n}}\sum_{i=1}^{n}(\widehat{V}_K(\boldsymbol{W_i}) - \rho_K(F_0)) \xrightarrow[n\to\infty]{d} \mathcal{N}(0, \sigma^2)$,

(ii) $\frac{1}{n}\sum_{i=1}^{n}(\widehat{V}_K(\boldsymbol{W_i}) - \rho_K(F_0))^2 \xrightarrow[n\to\infty]{P} \sigma^2$. *, kde*

$$\sigma^2 = 4\,\mathsf{P}((X-\mu_X)(Y-\mu_Y) > 0) \cdot \{1 - \mathsf{P}((X-\mu_X)(Y-\mu_Y) > 0)\}\,.$$

Důkaz.

(i) Důkaz nalezneme v příloze článku (Huang a Qin, 2022) jako důkaz Lemma 1.

(ii) Na levé straně přičteme a odečteme $V_K(\boldsymbol{W_i})$ uvnitř druhé mocniny a poté funkci roznásobíme.

$$\begin{aligned}
&\frac{1}{n}\sum_{i=1}^{n}(\widehat{V}_K(\boldsymbol{W_i}) - \rho_K(F_0))^2 \\
&\qquad = \frac{1}{n}\sum_{i=1}^{n}\Big((\widehat{V}_K(\boldsymbol{W_i}) - V_K(\boldsymbol{W_i})) + (V_K(\boldsymbol{W_i}) - \rho_K(F_0))\Big)^2 \\
&\qquad = \underbrace{\frac{1}{n}\sum_{i=1}^{n}(\widehat{V}_K(\boldsymbol{W_i}) - V_K(\boldsymbol{W_i}))^2}_{=:I_1} \\
&\qquad + \underbrace{\frac{2}{n}\sum_{i=1}^{n}(\widehat{V}_K(\boldsymbol{W_i}) - V_K(\boldsymbol{W_i}))(V_K(\boldsymbol{W_i}) - \rho_K(F_0))}_{=:I_2} \\
&\qquad + \underbrace{\frac{1}{n}\sum_{i=1}^{n}(V_K(\boldsymbol{W_i}) - \rho_K(F_0))^2}_{=:I_3}
\end{aligned} \tag{3.4}$$

Pro I_1 a I_2 využijeme toho, že

$$\begin{aligned}
\widehat{V}_K(\boldsymbol{W_i}) - V_K(\boldsymbol{W_i}) &= 2\left(\mathbb{1}_{[(X_i-\overline{X}_n)(Y_i-\overline{Y}_n)>0]} - 1 - \mathbb{1}_{[(X_i-\mu_X)(Y_i-\mu_Y)>0]} + 1\right) \\
&\xrightarrow[n\to\infty]{\mathsf{P}} 2\left(\mathbb{1}_{[(X_i-\mu_X)(Y_i-\mu_Y)>0]} - \mathbb{1}_{[(X_i-\mu_X)(Y_i-\mu_Y)>0]}\right) = 0,
\end{aligned} \tag{3.5}$$

ale pro I_2 potřebujeme ještě omezit druhý člen součinu.

$$\begin{aligned}
&|V_K(\boldsymbol{W_i}) - \rho_K(F_0)| \\
&= \left|2 \cdot \mathbb{1}_{[(X_i-\mu_X)(Y_i-\mu_Y)>0]} - 1 - (2\,\mathsf{P}_{F_0}\left([(X_1 - X_2)(Y_1 - Y_2)] > 0\right) - 1)\right| \\
&= 2 \cdot \left|\mathbb{1}_{[(X_i-\mu_X)(Y_i-\mu_Y)>0]} - \mathsf{P}_{F_0}\left([(X_1 - X_2)(Y_1 - Y_2)] > 0\right)\right| \leq 2.
\end{aligned} \tag{3.6}$$

Celkem tedy z věty o spojité transformaci a vlastnosti limit pro omezenou funkci

$$I_1 = \frac{1}{n}\sum_{i=1}^{n}(\widehat{V}_K(\boldsymbol{W_i}) - V_K(\boldsymbol{W_i}))^2 \xrightarrow[n\to\infty]{\mathsf{P}} 0 \tag{3.7}$$

$$I_2 = \frac{2}{n}\sum_{i=1}^{n}(\widehat{V}_K(\boldsymbol{W_i}) - V_K(\boldsymbol{W_i}))(V_K(\boldsymbol{W_i}) - \rho_K(F_0)) \xrightarrow[n\to\infty]{\mathsf{P}} 0. \tag{3.8}$$

Navíc I_3 je výběrový průměr náhodných veličin $(\widehat{V}_K(\boldsymbol{W_i}) - \rho_K(F_0))^2$ a platí pro něj, že

$$\begin{aligned}
I_3 &= \frac{1}{n}\sum_{i=1}^{n}(\widehat{V}_K(\boldsymbol{W_i}) - \rho_K(F_0))^2 \xrightarrow[n\to\infty]{\mathsf{P}} \mathsf{E}\left[\widehat{V}_K(\boldsymbol{W}) - \rho_K(F_0)\right]^2 \\
&= \mathsf{E}\left[\widehat{V}_K(\boldsymbol{W})\right]^2 - 2\rho_K(F_0)\,\mathsf{E}\left[\widehat{V}_K(\boldsymbol{W})\right] + [\rho_K(F_0)]^2 \\
&= \mathsf{E}\left[\widehat{V}_K(\boldsymbol{W})\right]^2 - [\rho_K(F_0)]^2 \\
&= 4 * \mathsf{E}\left[\mathbb{1}_{[(X_i-\mu_X)(Y_i-\mu_Y)>0]}\right]^2 - [\rho_K(F_0)]^2 = \sigma^2.
\end{aligned} \tag{3.9}$$

Nakonec tedy z (3.7), (3.8) a (3.9) pro (3.4) platí, že

$$\frac{1}{n}\sum_{i=1}^{n}(\widehat{V}_K(\boldsymbol{W_i}) - \rho_K(F_0))^2 = I_1 + I_2 + I_3 \xrightarrow[n\to\infty]{\mathsf{P}} \sigma^2. \tag{3.10}$$

Důkaz je tímto dokončen.

□

Nyní se již podíváme na větu o asymptotickém rozdělení $l_K(\rho_K(F_0))$, ze které pak jednoduše sestavíme intervalový odhad pro $\rho_K(F_0)$.

Věta 7 (Huang a Qin, 2022)**.** *Pokud $\mathsf{E}\,X < \infty$ a $\mathsf{E}\,Y < \infty$, potom asymptotické rozdělení $l_K(\rho_K(F_0))$ je chí-kvadrát rozdělení o jednom stupni volnosti, t.j.*

$$l_K(\rho_K(F_0)) \xrightarrow[n\to\infty]{d} \chi^2_1.$$

Důkaz. V důkazu budeme využívat, že $\lambda_K = o_p(n^{-1/2})$. Odvození této vlastnosti pro Personův korelační koeficient bychom mohli najít v práci (Owen, 1990), přičemž použitím stejných argumentů bychom ho získali i pro Kendallův korelační koeficient. Nejprve použijeme Taylorův rozvoj funkce

$$l_K(\rho_K(F_0)) = 2\sum_{i=1}^{n}\log\{1 + \lambda_K(\widehat{V}_K(\boldsymbol{W_i}) - \rho_K(F_0))\}.$$

Víme, že funkce $\log(1+x)$ v bodě 0 má Taylorův rozvoj $\log(1+x) = x - \frac{x^2}{2} + \frac{x^3}{3} - \frac{x^4}{4} + \cdots$ a dosazením $\lambda(\widehat{V}_K(\boldsymbol{W_i}) - \rho_K(F_0))$ za x získáme, že

$$\begin{aligned} l_K(\rho_K(F_0)) &= 2\sum_{i=1}^{n} \log\left\{1 + \lambda_K(\widehat{V}_K(\boldsymbol{W_i}) - \rho_K(F_0))\right\} = \\ &= 2\sum_{i=1}^{n}\left[\lambda_K(\widehat{V}_K(\boldsymbol{W_i}) - \rho_K(F_0)) - \frac{\left(\lambda_K(\widehat{V}_K(\boldsymbol{W_i}) - \rho_K(F_0))\right)^2}{2} + r_i\right], \end{aligned} \tag{3.11}$$

kde

$$r_i = \frac{\left(\lambda_K(\widehat{V}_K(\boldsymbol{W_i}) - \rho_K(F_0))\right)^3}{3} - \frac{\left(\lambda_K(\widehat{V}_K(\boldsymbol{W_i}) - \rho_K(F_0))\right)^4}{4} + \cdots$$

je zbytek Taylorova polynomu funkce $l_K(\rho_K(F_0))$ stupně 2. Z vlastností Taylorova rozvoje víme, že

$$|r_i| \leq \frac{C_i}{(2+1)!}\left|\lambda_K(\widehat{V}_K(\boldsymbol{W_i}) - \rho_K(F_0))\right|^3, \quad \text{kde } C_i \text{ je konstanta.}$$

Pro součet zbytků tímto získáváme následující omezení:

$$\begin{aligned} \left|2 \cdot \sum_{i=1}^{n} r_i\right| &\leq C \cdot \sum_{i=1}^{n} |\lambda_K|^3 \cdot \left|\widehat{V}_K(\boldsymbol{W_i}) - \rho_K(F_0)\right|^3 \\ &= C\,|\lambda_K|^3 \cdot \sum_{i=1}^{n} \underbrace{\left|\widehat{V}_K(\boldsymbol{W_i}) - \rho_K(F_0)\right|^3}_{\leq 2} \\ &\leq Cn\,|\lambda_K|^3 = Cn\,o_p(n^{-3/2}) = o_p(n^{-1/2}), \quad \text{kde } C \text{ je konstanta.} \end{aligned}$$

Dále z plug-in empirické věrohodnosti založené na funkci vlivu platí rovnost

$$\frac{1}{n}\sum_{i=1}^{n} \frac{\widehat{V}_K(\boldsymbol{W_i}) - \rho_K(F_0)}{1 + \lambda_K\left(\widehat{V}_K(\boldsymbol{W_i}) - \rho_K(F_0)\right)} = 0, \tag{3.12}$$

díky níž si můžeme vyjádřit λ_K a druhý člen Taylorova rozvoje z (3.11). Pro zjednodušení zápisu označme $Z_i = \widehat{V}_K(\boldsymbol{W_i}) - \rho_K(F_0)$. Pak z Taylorova rozvoje funkce $\dfrac{x}{1+\lambda_K x}$ v bodě 0 platí:

$$0 = \frac{1}{n}\sum_{i=1}^{n} \frac{Z_i}{1 + \lambda_K Z_i} = \frac{1}{n}\sum_{i=1}^{n}(Z_i - \lambda_K Z_i^2 + r_{2,i}).$$

Převedením $\frac{1}{n}\sum_{i=1}^{n} Z_i$ na levou stranu pak získáme, že

$$\frac{1}{n}\sum_{i=1}^{n} Z_i = \lambda_K \frac{1}{n}\sum_{i=1}^{n} Z_i^2 - \frac{1}{n}\sum_{i=1}^{n} r_{2,i}, \tag{3.13}$$

přičemž poslední sumu můžeme rovněž omezit z Taylorova rozvoje

$$\begin{aligned} \left|\frac{1}{n}\sum_{i=1}^{n} r_{2,i}\right| &\leq \frac{C}{n}\sum_{i=1}^{n} |\lambda_K|^2 \cdot |Z_i|^3 \leq \frac{C}{n}\sum_{i=1}^{n} |\lambda_K|^2 \cdot \left|\widehat{V}_K(\boldsymbol{W_i}) - \rho_K(F_0)\right|^3 \\ &= \frac{C}{n}\lambda_K^2 \cdot \sum_{i=1}^{n} \underbrace{\left|\widehat{V}_K(\boldsymbol{W_i}) - \rho_K(F_0)\right|^3}_{\leq 2} \leq C\lambda_K^2 = o_p(n^{-1}). \end{aligned} \tag{3.14}$$

Z (3.13) a (3.14) a dalšími úpravami získáme vyjádření, která nám umožní zapsat empirický log-věrohodnostní poměr založený na funkci vlivu tak, že za pomoci lemma 6 (i) a (ii) dokážeme určit jeho asymptotické vyjádření. Přímo z předchozího plyne, že

$$\frac{1}{n}\sum_{i=1}^{n} Z_i = \lambda_K \frac{1}{n}\sum_{i=1}^{n} Z_i^2 + o_p(n^{-1}). \tag{3.15}$$

Z toho dostáváme vyjádření pro λ_K a druhý člen Taylorova rozvoje z (3.11). Pro λ_K rovnost vydělíme pomocí $\frac{1}{n}\sum_{i=1}^{n} Z_i^2$ a pro druhý člen Taylorova rozvoje ji přenásobíme λ_K a vynásobíme n.

$$\begin{aligned} \lambda_K &= \frac{\sum_{i=1}^{n} Z_i}{\sum_{i=1}^{n} Z_i^2} + o_p(n^{-1}) \\ \sum_{i=1}^{n} \lambda_K Z_i &= \sum_{i=1}^{n} (\lambda_K Z_i)^2 + o_p(n^{-1/2}) \end{aligned} \tag{3.16}$$

Dosazením zpět za Z_i obdržíme rovnosti ve tvaru:

$$\lambda_K = \frac{\sum_{i=1}^{n}(\widehat{V}_K(\boldsymbol{W_i}) - \rho_K(F_0))}{\sum_{i=1}^{n}(\widehat{V}_K(\boldsymbol{W_i}) - \rho_K(F_0))^2} + o_p(n^{-1}) \tag{3.17}$$

$$\sum_{i=1}^{n} \lambda_K(\widehat{V}_K(\boldsymbol{W_i}) - \rho_K(F_0)) = \sum_{i=1}^{n} \Big(\lambda_K(\widehat{V}_K(\boldsymbol{W_i}) - \rho_K(F_0))\Big)^2 + o_p(n^{-1/2}). \tag{3.18}$$

Z lemma 6 (i) a (ii) tedy získáváme, že

$$\begin{aligned} & l_K(\rho_K(F_0)) \overset{(3.11)}{=} \\ & = 2\sum_{i=1}^{n}\left[\lambda_K(\widehat{V}_K(\boldsymbol{W_i}) - \rho_K(F_0)) - \frac{\Big(\lambda_K(\widehat{V}_K(\boldsymbol{W_i}) - \rho_K(F_0))\Big)^2}{2}\right] + o_p(n^{-1/2}) \\ & \overset{(3.18)}{=} 2\sum_{i=1}^{n}\lambda_K(\widehat{V}_K(\boldsymbol{W_i}) - \rho_K(F_0)) - \sum_{i=1}^{n}\lambda_K(\widehat{V}_K(\boldsymbol{W_i}) - \rho_K(F_0)) + o_p(n^{-1/2}) \\ & = \sum_{i=1}^{n}\lambda_K(\widehat{V}_K(\boldsymbol{W_i}) - \rho_K(F_0)) + o_p(n^{-1/2}) \\ & \overset{(3.17)}{=} \frac{\left[\overbrace{\frac{1}{\sqrt{n}}\sum_{i=1}^{n}(\widehat{V}_K(\boldsymbol{W_i}) - \rho_K(F_0))}^{\xrightarrow{d}\mathcal{N}(0,\,\sigma^2)}\right]^2}{\underbrace{\frac{1}{n}\sum_{i=1}^{n}(\widehat{V}_K(\boldsymbol{W_i}) - \rho_K(F_0))^2}_{\xrightarrow{\mathsf{P}}\sigma^2}} + o_p(n^{-1/2}) \\ & = \left[\frac{\overbrace{\sum_{i=1}^{n}(\widehat{V}_K(\boldsymbol{W_i}) - \rho_K(F_0))}^{\xrightarrow{d}\mathcal{N}(0,\,1)}}{\sqrt{\sum_{i=1}^{n}(\widehat{V}_K(\boldsymbol{W_i}) - \rho_K(F_0))^2}}\right]^2 + o_p(n^{-1/2}) \xrightarrow[n\to\infty]{d} \chi_1^2. \end{aligned} \tag{3.19}$$

Tím máme tvrzení věty dokázáno.

□

Na závěr interval o spolehlivosti $(1-\alpha)$ odvozený z věty 7 pro $\rho_K(F_0)$ je tvaru

$$\{\rho_K(F_0) : l_K(\rho_K(F_0)) \leq \chi^2_1(1-\alpha)\}.$$

4. Simulace

V této kapitole provedeme simulační studii pravděpodobnosti pokrytí a průměrné délky intervalových odhadů o spolehlivosti 95 % zkonstruovaných pomocí metod představených v předchozích kapitolách. Porovnáme si jejich výsledky pro konečné rozsahy výběrů $n = 50$, 100 a 250. Pro tři různé hodnoty korelačního koeficientu ρ : -0.5, 0.2, 0.85, neboli pro negativní střední, slabou a pozitivní silnou závislost X a Y, budeme uvažovat následující rozdělení.

(i) Dvojrozměrné normované normální rozdělení:

$$\begin{pmatrix} X \\ Y \end{pmatrix} \sim \mathcal{N}_2 \left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix} \right). \tag{4.1}$$

(ii) Dvojrozměrné nenormované normální rozdělení:

$$\begin{pmatrix} X \\ Y \end{pmatrix} \sim \mathcal{N}_2 \left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 5 & \sqrt{10}\rho \\ \sqrt{10}\rho & 2 \end{pmatrix} \right). \tag{4.2}$$

(iii) Dvousložková směs normálních rozdělení:

$$\begin{pmatrix} X \\ Y \end{pmatrix} \sim 0.9 \cdot \mathcal{N}_2 \left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix} \right) + 0.1 \cdot \mathcal{N}_2 \left(\begin{pmatrix} 0 \\ 2 \end{pmatrix}, \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix} \right). \tag{4.3}$$

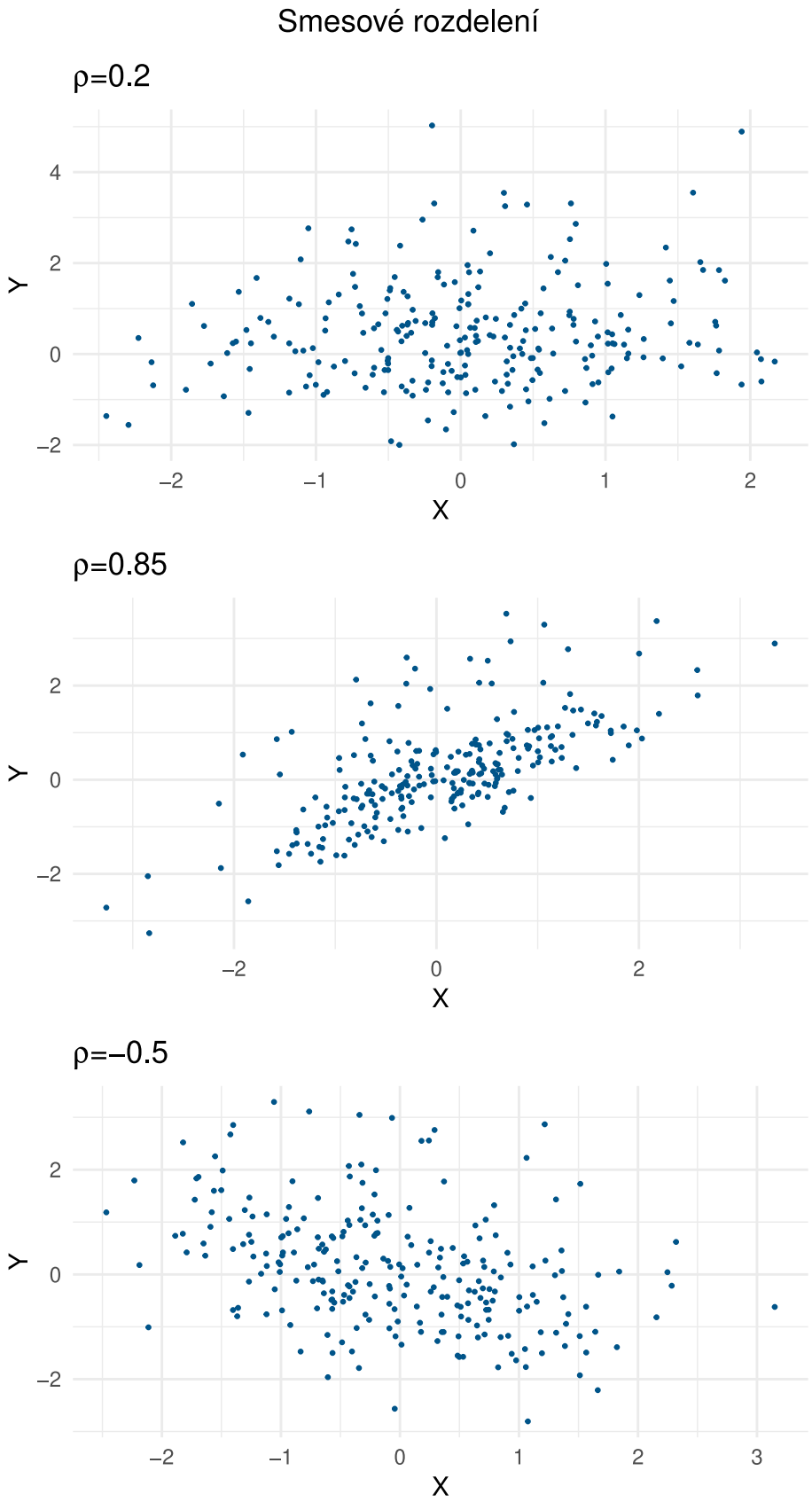
(iv) Kvadratické rozdělení:

$$\begin{aligned} &X \sim \mathcal{R}(0{,}1), \\ &Y \sim sign(\rho) \sqrt{\frac{36\rho^2}{15 - 16\rho^2}} X^2 + \mathcal{N}_1 \left(0, 0.2\right). \end{aligned} \tag{4.4}$$

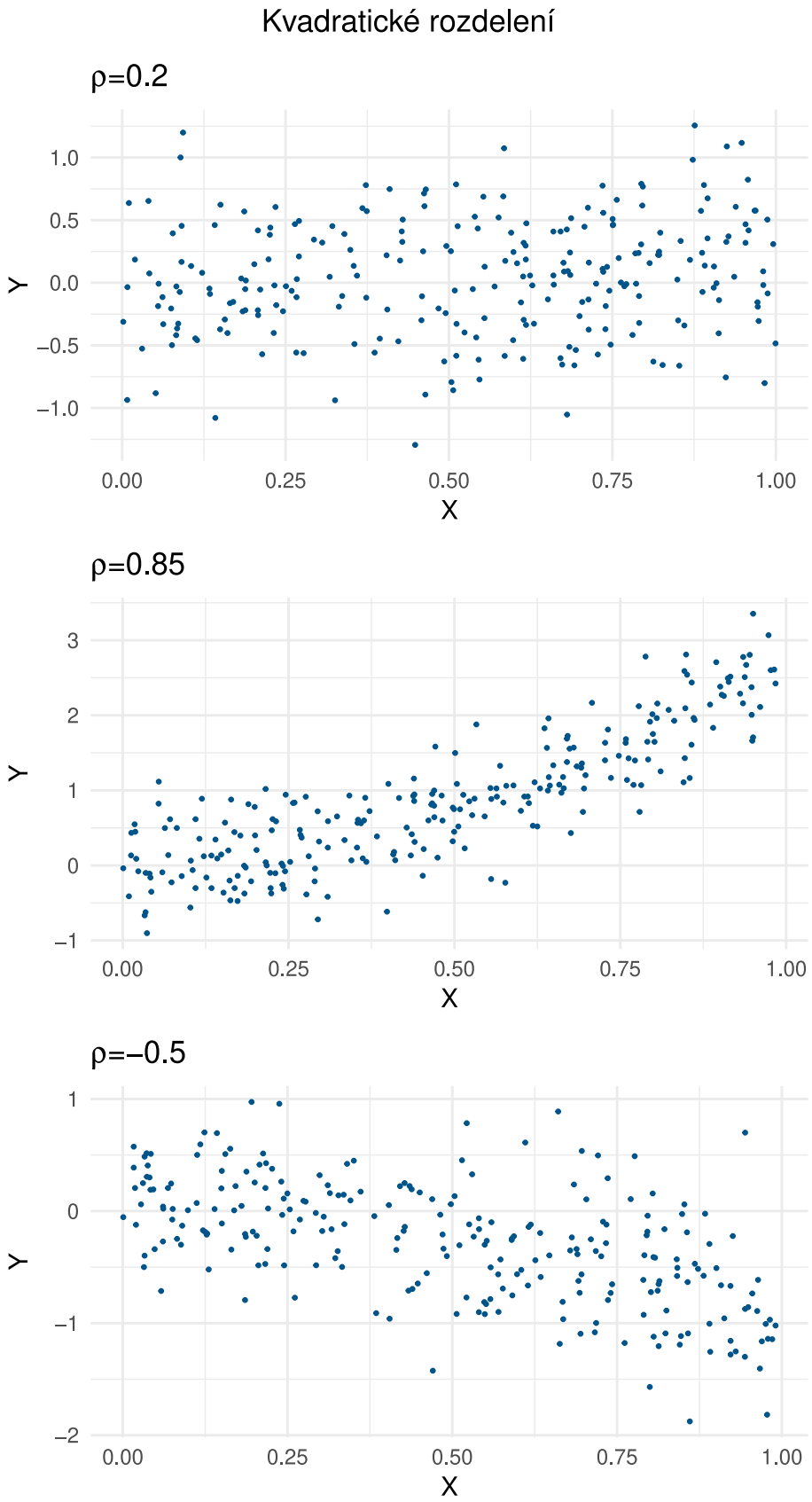
Ke každé z kombinací generujeme 5000 náhodných výběrů a počítáme pravděpodobnost pokrytí korelačního koeficientu a průměrnou délku intervalu spolehlivosti.

Zápisem dvousložkové směsi normálních rozdělení míníme, že 90 % pochází z prvního normálního rozdělení a 10 % z druhého normálního rozdělení. Toto rozdělení jsme zvolili, abychom měli možnost pozorovat chování jednotlivých metod v okamžiku, kdy naše pozorování budou obsahovat odlehlé hodnoty. Neboli našim cílem je zjistit pokrytí korelačního koeficientu ρ a ρ_K pouze prvního rozdělení (ne celé dvousložkové směsi), i když náhodný výběr obsahuje 10 % odlehlých hodnot. V reálném světě si pod tím můžeme představit, že jsme měřili hmotnost a rozdíl tepu před a po jednom dřepu běžkyň před závodem a zajímá nás korelační koeficient těchto veličin pro ženy trénující pětkrát týdně, přičemž se závodů účastní 10 % žen, které se běhání věnují třikrát týdně. Zde nás zajímá, zdali i při 10% přítomnosti odlehlých hodnot bude intervalový odhad stále pokrývat ρ či ρ_K.

Poznámka. Při pohledu na kvadratické rozdělení (4.4) není zřejmé, že korelačí koeficient je roven ρ, a tak si ukážeme jeho odvození.



Obrázek 4.1: Náhodný výběr ze směsového rozdělení (4.3) o rozsahu 500.



Obrázek 4.2: Náhodný výběr z kvadratického rozdělení (4.4) o rozsahu 500.

Označíme si $Z \sim \mathcal{N}_1\left(0, 0.2\right)$,

$$a = sign(\rho)\sqrt{\frac{36\rho^2}{15 - 16\rho^2}}$$

a ρ_T teoretickou hodnotu korelačního koeficientu. Potom platí:

$$\begin{aligned}
\rho_T &= \frac{\mathrm{cov}(X,Y)}{\sqrt{\mathrm{var}X\,\mathrm{var}Y}}, \qquad (4.5)\\
\mathrm{cov}(X,Y) &= \mathsf{E}\ XY - \mathsf{E}\ X\,\mathsf{E}\ Y\\
&= a\,\mathsf{E}\ X^3 + \mathsf{E}\ XZ - \mathsf{E}\ X\left(a\,\mathsf{E}\ X^2 + \underbrace{\mathsf{E}\ Z}_{=0}\right)\\
&= a\,\mathsf{E}\ X^3 + \mathsf{E}\ X\underbrace{\mathsf{E}\ Z}_{=0} - a\,\mathsf{E}\ X\,\mathsf{E}\ X^2\\
&= a\,\mathsf{E}\ X^3 - a\,\mathsf{E}\ X\,\mathsf{E}\ X^2,
\end{aligned}$$

jelikož střední hodnota součinu dvou nezávislých náhodných veličin X a Z je součin jejich středních hodnot.

Dále se zaměříme na výpočty, které plynou z toho, že X má rovnoměrné rozdělení na intervalu (0, 1).

$$\begin{aligned}
\mathsf{E}\ X &= \int_{-\infty}^{\infty} x\mathbb{1}_{(0,1)}\,dx = \left[\frac{x^2}{2}\right]_0^1 = \frac{1}{2}\\
\mathsf{E}\ X^2 &= \frac{1}{3}\\
\mathsf{E}\ X^3 &= \frac{1}{4}\\
\mathsf{E}\ X^4 &= \frac{1}{5}\\
\mathrm{var}X &= \mathsf{E}\ X^2 - (\mathsf{E}\ X)^2 = \frac{1}{3} - \frac{1}{2}^2 = \frac{1}{12}
\end{aligned}$$

Nyní dopočítáme rozptyl náhodné veličiny Y a budeme moct všechny částečné výpočty dosadit do vzorce 4.5.

$$\begin{aligned}
\mathrm{var}Y &= \mathrm{var}\,aX^2 + \mathrm{var}Z + \underbrace{\mathrm{cov}(aX^2,Z)}_{=0}\\
&= a^2\mathrm{var}X^2 + \frac{1}{5}\\
&= \frac{36\rho^2}{15 - 16\rho^2}\left(\frac{1}{5} - \frac{1}{3}^2\right) + \frac{1}{5}\\
&= \frac{36\rho^2}{15 - 16\rho^2}\cdot\frac{4}{45} + \frac{1}{5}
\end{aligned}$$

Celkem tedy získáváme ověření, že ρ_T odpovídá námi předpokládanému korelač-

ρ	Normované (4.1)	Nenormované (4.2)	Kvadratické (4.4)
-0.5	-0.296	-0.372	-0.368
0.2	0.111	0.143	0.139
0.85	0.577	0.720	0.725

Tabulka 4.1: Monte Carlo aproximace Kendallova korelačního koeficientu.

nímu koeficientu.

$$\begin{aligned}
\rho_T &= \frac{sign(\rho)\sqrt{\frac{36\rho^2}{15-16\rho^2}} \cdot \left(\frac{1}{4} - \frac{1}{2} \cdot \frac{1}{3}\right)}{\sqrt{\frac{1}{12} \cdot \left(\frac{36\rho^2}{15-16\rho^2} \cdot \frac{4}{45} + \frac{1}{5}\right)}} \\
&= \frac{sign(\rho)\sqrt{\frac{36\rho^2}{15-16\rho^2}} \cdot \frac{1}{12}}{\sqrt{\frac{1}{12} \cdot \frac{3}{(15-16\rho^2)}}} \\
&= sign(\rho)\frac{6\,|\rho|}{\sqrt{15-16\rho^2}} \cdot \frac{1}{12} \cdot \frac{2\sqrt{15-16\rho^2}}{1} \\
&= sign(\rho) \cdot |\rho| = \rho
\end{aligned}$$

Před zkoumáním daných výsledků, které je možno vidět v tabulkách 4.2, 4.3, 4.4 a 4.5, si připomeneme jednotlivé metody. Pro korelační koeficient můžeme využít Fisherovu z-transformaci (FZ), plug-in metodu věrohodnosti (PEL) a metodu věrohodnosti založené na funkci vlivu (IFEL). Kendallův korelační koeficient můžeme určit přesně pouze pro normální rozdělení s distribuční funkcí F_0 na základě vztahu (Huang a Qin, 2022, viz)

$$\rho_K(F_0) = \frac{2}{\pi} arcsin(\rho).$$

Za účelem jednotnosti jsme zvolili Monte Carlo aproximaci skutečné hodnoty $\rho_K(F)$ pro distribuční funkce F všech uvažovaných rozdělení (viz tabulka 4.1). Můžeme pozorovat, že výsledky této aproximace zhruba odpovídají výše uvedenému vzorci i pro ostatní rozdělení a pro simulaci je nahrazujeme za teoretické hodnoty v normální metodě (KN) a metodě věrohodnosti založená na funkci vlivu pro Kendallův korelační koeficient (KIFEL). Může být zarážející, proč v tabulce 4.1 chybí hodnoty pro směsové rozdělení (4.3). V době, kdy chceme zkoumat vliv odlehlých hodnot na výpočet Kendallova korelačního koeficientu, tak jako teoretickou hodnotu uvažujeme Kendallův korelační koeficient pouze rozdělení, jehož podíl je 90 %, t.j. normovaného normálního rozdělení (4.1).

V tabulkách můžeme pozorovat, že pro dvojrozměrné normální rozdělení normované 4.2 i nenormované 4.3, mají Fisherovy intervaly spolehlivosti pro ρ velmi dobré pokrytí již v případě, že rozsah výběru $n = 50$. Metody PEL a IFEL mají problémy s nízkou pravděpodobností pokrytí pro malé rozsahy výběru. Při jejich porovnání zjistíme, že PEL si vede hůře pro $\rho = -0.5$ a 0.2 naopak IFEL má nižší hodnoty pravděpodobnosti pokrytí pro $\rho = 0.85$. Průměrná délka intervalů

		Pravděpodobnost pokrytí					Průměrná délka intervalu				
n	ρ	FZ	PEL	IFEL	KN	KIFEL	FZ	PEL	IFEL	KN	KIFEL
50	−0.50	0.945	0.918	0.924	0.925	0.951	0.422	0.403	0.401	0.510	0.501
50	0.20	0.950	0.909	0.931	0.934	0.954	0.527	0.526	0.502	0.543	0.533
50	0.85	0.949	0.949	0.930	0.923	0.959	0.165	0.148	0.156	0.374	0.372
100	−0.50	0.949	0.934	0.937	0.943	0.943	0.296	0.291	0.291	0.362	0.359
100	0.20	0.950	0.931	0.941	0.945	0.945	0.374	0.377	0.366	0.386	0.383
100	0.85	0.956	0.954	0.946	0.938	0.957	0.113	0.107	0.110	0.268	0.267
250	−0.50	0.955	0.947	0.952	0.941	0.946	0.186	0.185	0.185	0.230	0.229
250	0.20	0.956	0.948	0.951	0.952	0.952	0.238	0.239	0.236	0.245	0.244
250	0.85	0.948	0.949	0.946	0.937	0.944	0.070	0.068	0.069	0.170	0.170

Tabulka 4.2: Intervalové odhady ρ a ρ_K pro normované normální rozdělení (4.1).

		Pravděpodobnost pokrytí					Průměrná délka intervalu				
n	ρ	FZ	PEL	IFEL	KN	KIFEL	FZ	PEL	IFEL	KN	KIFEL
50	−0.50	0.955	0.927	0.936	0.947	0.947	0.420	0.401	0.399	0.517	0.508
50	0.20	0.949	0.911	0.929	0.936	0.936	0.527	0.525	0.501	0.544	0.534
50	0.85	0.954	0.948	0.931	0.944	0.931	0.165	0.148	0.156	0.415	0.410
100	−0.50	0.948	0.932	0.938	0.955	0.944	0.295	0.289	0.289	0.367	0.364
100	0.20	0.951	0.925	0.937	0.943	0.943	0.374	0.377	0.367	0.387	0.383
100	0.85	0.946	0.945	0.937	0.938	0.952	0.112	0.107	0.110	0.296	0.294
250	−0.50	0.956	0.948	0.952	0.948	0.948	0.186	0.186	0.186	0.233	0.232
250	0.20	0.951	0.944	0.947	0.944	0.944	0.238	0.239	0.236	0.245	0.244
250	0.85	0.948	0.950	0.946	0.950	0.954	0.070	0.068	0.069	0.188	0.188

Tabulka 4.3: Intervalové odhady ρ a ρ_K pro nenormované normální rozdělení (4.2).

je obdobná pro všechny tři metody. Pokud je rozsah výběru $n = 250$, tak všechny metody intervalového odhadu pro korelační koeficient dosahují stejného výkonu. Při odhadu Kendallova korelačního koeficientu mají metody KN a KIFEL nízkou pravděpodobnost pokrytí pouze pro rozsah výběru $n = 50$, navíc metoda KN má problém s nedostatečnou pravděpodobností pokrytí pro $\rho = 0.85$ i pro vyšší rozsahy výběru. Pro $n = 100$ a 250má KIFEL vyšší nebo rovnou pravděpodobnost pokrytí než KN, i když průměrná délka intervalů je trochu nižší.

Při prvním pohledu na tabulku intervalových odhadů pro směs rozdělení (4.4) nás jistě zarazí, že pravděpodobnosti pokrytí ve většině kombinací nejsou blízké 95 %. Míra odchýlení od této pravděpodobnosti odpovídá absolutní hodnotě korelačního koeficientu rozdělení zastoupeného 90 % ve směsi, tedy vyšší korelace náhodných veličin znamená vyšší citlivost intervalového odhadu na odlehlá pozorování neboli nižší pravděpodobnost pokrytí teoretické hodnoty. Pravděpodobnost pokrytí se rovněž snižuje s rostoucím rozsahem výběru, jelikož se intervalový odhad zpřesňuje a zkracuje i na základě vyššího počtu odlehlých pozorování. Může se zdát, že intervalové odhady pro Kendallův korelační koeficient se s tímto problémem vyrovnaly o hodně lépe, ale je třeba mít na paměti, že $\rho_K = 0.65$ pro $\rho = 0.85$.

Část tabulky intervalových odhadů pro kvadratické rozdělení 4.5 týkající se metod pro korelační koeficient ρ koresponduje s tabulkami pro normální rozdělení 4.2 a 4.3. Hlavním rozdílem je zde velmi vysoká pravděpodobnost pokrytí pro $\rho =$ 0.85 Fisherovou z-transformací. Pravděpodobnost pokrytí odhadů pro Kendallův korelační koeficient se pro obě metody pohybuje okolo 95 % ve scénářích,

		Pravděpodobnost pokrytí					Průměrná délka intervalu				
n	ρ	FZ	PEL	IFEL	KN	KIFEL	FZ	PEL	IFEL	KN	KIFEL
50	−0.50	0.918	0.931	0.894	0.911	0.920	0.454	0.442	0.434	0.523	0.514
50	0.20	0.943	0.900	0.919	0.949	0.949	0.531	0.531	0.504	0.545	0.535
50	0.85	0.417	0.613	0.427	0.725	0.727	0.271	0.273	0.283	0.449	0.442
100	−0.50	0.852	0.900	0.838	0.902	0.868	0.321	0.321	0.318	0.372	0.369
100	0.20	0.940	0.924	0.927	0.936	0.936	0.378	0.384	0.371	0.387	0.384
100	0.85	0.140	0.270	0.170	0.510	0.414	0.187	0.198	0.203	0.319	0.317
250	−0.50	0.715	0.787	0.714	0.817	0.781	0.202	0.204	0.204	0.236	0.235
250	0.20	0.927	0.929	0.921	0.909	0.909	0.240	0.243	0.239	0.246	0.245
250	0.85	0.002	0.009	0.004	0.137	0.106	0.117	0.127	0.129	0.203	0.202

Tabulka 4.4: Intervalové odhady ρ a ρ_K pro složené dvojrozměrné rozdělení (4.3).

		Pravděpodobnost pokrytí					Průměrná délka intervalu				
n	ρ	FZ	PEL	IFEL	KN	KIFEL	FZ	PEL	IFEL	KN	KIFEL
50	−0.50	0.956	0.921	0.931	0.926	0.952	0.421	0.393	0.394	0.509	0.501
50	0.20	0.944	0.909	0.928	0.934	0.955	0.526	0.520	0.503	0.543	0.533
50	0.85	0.969	0.952	0.928	0.922	0.958	0.163	0.131	0.139	0.374	0.372
100	−0.50	0.958	0.937	0.946	0.941	0.945	0.295	0.280	0.281	0.363	0.359
100	0.20	0.950	0.932	0.942	0.956	0.944	0.374	0.373	0.366	0.386	0.383
100	0.85	0.976	0.954	0.945	0.939	0.955	0.112	0.094	0.097	0.268	0.267
250	−0.50	0.954	0.939	0.944	0.947	0.950	0.186	0.179	0.179	0.230	0.229
250	0.20	0.952	0.948	0.951	0.942	0.952	0.238	0.237	0.235	0.245	0.244
250	0.85	0.978	0.958	0.956	0.944	0.953	0.070	0.060	0.061	0.170	0.170

Tabulka 4.5: Intervalové odhady ρ a ρ_K pro kvadratické rozdělení (4.4).

kdy je rozsha výběru vyšší než 50. Největší problém zaznamenáváme u metody KN, když $n = 50$. Metoda KIFEL v porovnání s KN má vyšší pravděpodobnost pokrytí a mírně nižší průměrnou délku intervalu.

Závěr

V bakalářské práci jsme nejprve zadefinovali základní pojmy týkající se Pearsonova korelačního koeficientu a Kendallova korelačního koeficientu. Udělali jsme si představu o vzhledu sdružené hustoty dvojrozměrného rozdělení a hodnotách obou korelačních koeficientů pro různé vztahy náhodných veličin. Dále jsme si ukázali odvození intervalu spolehlivosti pomocí z-transformace z tvrzení o asymptotickém rozdělení Pearsonova výběrového korelačního koeficientu.

Navázali jsme definicí neparametrické věrohodnosti a důkazem její vlastnosti pro empirickou distribuční funkci. Uvedli jsme spojitost mezi neparametrickou věrohodností a plug-in empirickou věrohodností pro korelační koeficient a pomocí metody Lagrangeových multiplikátorů jsme doplnili výpočet tvaru plug-in empirického log-věrohodnostního poměru, z jehož asymptotického rozdělení jsme získali další intervalový odhad korelačního koeficientu. Analogicky jsme také představili metodu věrohodnosti založené na funkci vlivu pro korelační koeficient.

Třetí kapitola se zabývala metodami intervalových odhadů pro Kendallův korelační koeficient, jejichž základ je postavený na funkci vlivu. Velká část této kapitoly byla zaměřená na podrobné doplnění důkazů vyskytujících se v článku (Huang a Qin, 2022).

Na závěr jsme v simulační studii zkoumali pravděpodobnost pokrytí a průměrnou délku intervalových odhadů odvozených v jednotlivých metodách, když uvažujeme tři možnosti konečného rozsahy výběru, čtyři různá rozdělení a odlišné síly závislosti mezi dvěma náhodnými veličinami.

Seznam použité literatury

Anděl, J. (2011). *Základy matematické statistiky*. Vydání třetí. Matfyzpress, Praha. ISBN 978-80-7378-162-0.

Farda, M. (2021). *Intervaly spolehlivosti pro korelační koeficient*. Bakalářská práce, Univerzita Karlova.

Hu, X., Jung, A. a Qin, G. (2020). Interval estimation for the correlation coefficient. *The American Statistician*, **74**(1), 29–36. doi: 10.1080/00031305.2018.1437077.

Huang, Z. a Qin, G. (2022). Influence function-based confidence intervals for the kendall rank correlation coefficient - computational statistics. *SpringerLink*.

Lazar, N. A. (2021). A review of empirical likelihood. *Annual Review of Statistics and Its Application*, **8**(1), 329–344. doi: https://doi.org/10.1146/annurev-statistics-040720-024710.

Owen, A. (1990). Empirical Likelihood Ratio Confidence Regions. *The Annals of Statistics*, **18**(1), 90 – 120. doi: 10.1214/aos/1176347494.

Owen, A. B. (2001). *Empirical Likelihood*. Chapman & Hall/CRC. ISBN 1584880716. doi: https://doi.org/10.1201/9781420036152.