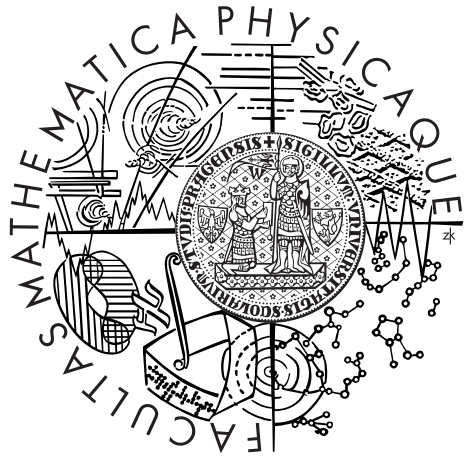
Univerzita Karlova v Praze

Matematicko-fyzikální fakulta

BAKALÁŘSKÁ PRÁCE



Richard Németh

Pearsonův korelační koeficient a jeho využití ve statistice

Katedra pravděpodobnosti a matematické statistiky

Vedoucí bakalářské práce:	Ing. Marek Omelka, PhD.
Studijní program:	Matematika
Studijní obor:	Obecná matematika

Praha 2015

Dovoluji si na tomto místě poděkovat Ing. Markovi Omelkovi, PhD., za jeho nestálou podporu, čas, ochotu i cenné rady při vzniku této bakalářské práce.

Prohlašuji, že jsem tuto bakalářskou práci vypracoval samostatně a výhradně s použitím citovaných pramenů, literatury a dalších odborných zdrojů.

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 v Praze má právo na uzavření licenční smlouvy o užití této práce jako školního díla podle §60 odst. 1 autorského zákona.

V ..... dne ..... Podpis autora

Název práce: Pearsonův korelační koeficient a jeho využití ve statistice

Autor: Richard Németh

Katedra: Katedra pravděpodobnosti a matematické statistiky

Vedoucí bakalářské práce: Ing. Marek Omelka, PhD., Katedra pravděpodobnosti a matematické statistiky MFF UK

Abstrakt: Cílem této práce je určení asymptotického rozdělení výběrového korelačního koeficientu bez předpokladu normality a prozkoumat následné důsledky tohoto rozdělení na běžně užívané statistické testy nezávislosti a intervaly spolehlivosti pro korelační koeficient. Problém je vyřešen pomocí centrální limitní věty a delta metody. Dokázali jsme, že běžně užívané testy nezávislosti v praxi jsou v asymptotickém smyslu v pořádku i bez předpokladu normálního rozdělení. V práci jsou odvozené další varianty statistických testů pro nezávislost náhodných veličín a taky další varianty intervalů spolehlivosti pro korelační koeficient bez předpokladu normality. V závěru pomocí simulací porovnáváme jednotlivé statistické testy nezávislosti a intervaly spolehlivosti pro specifická vícerozměrná rozdělení.

Klíčová slova: korelační koeficient, asymptotické rozdělení, testy nezávislosti

Title: Pearsonův correlation coefficient and its use in statistical inference

Author: Richard Németh

Department: Department of Probability and Mathematical Statistics

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

Abstract: The main objective of this thesis is to determine asymptotic distribution of sample correlation coefficient without the assumption of normal distribution and its effects on commonly used statistical tests of independence and confidence intervals for correlation coefficient. The problem is solved by central limit theorem and delta method. We have shown that the commonly used statistical tests for independence in practice are valid, even without the assumption of normal distribution. We have also derived more versions of statistical tests for independence of random variables and more versions of confidence intervals for correlation coefficient without the assumption of normality. In conclusion we have compared individual statistical tests and confidence intervals for specific multivariate distributions using simulations.

Keywords: correlation coefficient, asymptotic distribution, tests of independence

Obsah

1 Úvod **1**
1.1 Úvod a základní pojmy 1

2 Statistická inference o korelačním koeficientu **5**
2.1 Asymptotické rozdělení 5
2.2 Testy nezávislosti 12
2.3 Fisherova Z-transformace 13
2.4 Obecný přístup k intervalům spolehlivosti 14

3 Simulace **17**
3.1 Hladiny testů 17
3.2 Intervaly spolehlivosti 24

Závěr **28**

Literatura **29**

Seznam obrázků **30**

Seznam tabulek **31**

Kapitola 1

Úvod

1.1 Úvod a základní pojmy

Jednou z nejběžnějších otázek, které se v statistice vyskytují, je když máme data ve formě dvojic, tak zda jsou tyto data na sobě závislá. Existuje několik způsobů jak tuto otázku vyřešit a jedním způsobem je právě užití korelačního koeficientu.

1.1. Definice (Korelační koeficient). Nechť $(X,Y)'$ je náhodný vektor, který splňuje nerovnosti $0 < \mathrm{var}(X) < +\infty$, $0 < \mathrm{var}(Y) < +\infty$, pak **korelačním koeficientem** $\mathrm{corr}(X,Y)$ náhodných veličín X,Y rozumíme

$$\mathrm{corr}(X,Y) = \frac{\mathrm{cov}(X,Y)}{\sqrt{\mathrm{var}(X)}\sqrt{\mathrm{var}(Y)}}.$$

1.2. Tvrzení (Vlastnosti korelačního koeficientu). Nechť X,Y jsou nedegenerované náhodné veličiny[1], pak

1. $|\mathrm{corr}(X,Y)| \leq 1$ a $|\mathrm{corr}(X,Y)| = 1$ právě tehdy, když existují taková $a,b \in \mathbb{R}$:

$$Y = aX + b \text{ skoro jistě.}$$

2. Jsou-li veličiny X,Y nezávislé, pak $\mathrm{corr}(X,Y) = 0$. Opačná implikace ale neplatí.
3. $\forall a,b,c,d \in \mathbb{R} : ac \neq 0$ platí $\mathrm{corr}(aX + b, cY + d) = \mathrm{sgn}(ac)\mathrm{corr}(X,Y)$.

Důkaz. Tvrzení 1. a 3. je dokázáno v Anděl (2007) na str. 39 a tvrzení 2. lze nalézt v Dupač a Hušková (2005) na str. 62. □

[1] Reálná náhodná veličina X je degenerovaná právě tehdy, když existuje $c \in \mathbb{R}$ takové, že $\mathrm{P}(X = c) = 1$. Následně pak náhodná veličina je nedegerovaná pokud není degenerovaná.

Tvrzení 1.2 vlastně říká, že korelační koeficient nám udáva míru lineární závislosti mezi náhodnými veličinami. V praxi se setkáváme s dvojicemi dat a může nás zajímat, zda některá data na sebe závisí. Proto zavádíme výběrový korelační koeficient.

1.3. Definice (Výběrový korelační koeficient). Nechť $(X_1,Y_1)',\ldots,(X_n,Y_n)'$ je náhodný výběr z dvojrozměrného rozdělení, pak **výběrovým korelačním koeficientem** r_n rozumíme náhodnou veličinu definovanou předpisem

$$r_n = \frac{s_{\mathbf{XY}}}{s_{\mathbf{X}} s_{\mathbf{Y}}}, \tag{1.1}$$

kde

$$s_{\mathbf{XY}} = \frac{1}{n-1}\sum_{i=1}^{n}\left(X_i - \overline{X}_n\right)\left(Y_i - \overline{Y}_n\right),$$

$$s^2_{\mathbf{X}} = \frac{1}{n-1}\sum_{i=1}^{n}\left(X_i - \overline{X}_n\right)^2, \qquad s^2_{\mathbf{Y}} = \frac{1}{n-1}\sum_{i=1}^{n}\left(Y_i - \overline{Y}_n\right)^2,$$

$$\overline{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i, \qquad \overline{Y}_n = \frac{1}{n}\sum_{i=1}^{n} Y_i.$$

1.4. Věta. Při výběru o rozsahu alespoň 2 ze spojitého rozdělení je výběrový korelační koeficient definován s pravděpodobností 1.

Důkaz. Viz Anděl (2007) str. 93. □

1.5. Definice (Normální rozdělení). Nechť $k \in \mathbb{N}$, $\boldsymbol{\mu} \in \mathbb{R}^k$ a $\mathbf{V} \in \mathbb{R}^{k\times k}$ je positivně semidefinitní reálná symetrická matice. Řekneme, že reálný k-rozměrný náhodný vektor $\mathbf{X}$ má **k-rozměrné normální rozdělení** s parametry $\boldsymbol{\mu}$,$\mathbf{V}$, jestliže $\forall \boldsymbol{c} \in \mathbb{R}^k : \boldsymbol{c}'\mathbf{X} \sim \mathcal{N}(\boldsymbol{c}'\boldsymbol{\mu},\boldsymbol{c}'\mathbf{V}\boldsymbol{c})$, kde rozdělením $\mathcal{N}(\boldsymbol{c}'\boldsymbol{\mu},0)$ rozumíme Diracovu míru v bodě $\boldsymbol{c}'\boldsymbol{\mu}$. Vícerozměrné normální rozdělení budeme značit $\mathcal{N}_k(\boldsymbol{\mu},\mathbf{V})$.

Protože definice výběrového korelačního koeficientu vychází z jeho teoretického protějšku, je poměrně snadné odvodit analogické vlastnosti z Tvrzení 1.2, ale pro výběrový korelační koeficient.

1.6. Tvrzení (Vlastnosti výběrového korelačního koeficientu). Nechť $(X_1,Y_1)',\ldots,(X_n,Y_n)'$ je náhodný výběr z dvojrozměrného rozdělení. Označme $r_n(\mathbf{X},\mathbf{Y})$ jako výběrový korelační koeficient náhodných vektorů $\mathbf{X} = (X_1,\ldots,X_m)'$, $\mathbf{Y} = (Y_1,\ldots,Y_m)'$, pak platí následující tvrzení:

1. $r_n(a\mathbf{X}+\boldsymbol{b},c\mathbf{Y}+\boldsymbol{d}) = \operatorname{sgn}(ac) r_n(\mathbf{X},\mathbf{Y})$, pro libovolné $a,c \in \mathbb{R}$ a $\boldsymbol{b},\boldsymbol{d} \in \mathbb{R}^m$,
2. $|r_n(\mathbf{X},\mathbf{Y})| \leq 1$. Navíc $|r_n(\mathbf{X},\mathbf{Y})| = 1$ skoro jistě právě tehdy, když existují $a \in \mathbb{R}$ a $\boldsymbol{b} \in \mathbb{R}^m : \mathbf{X} = a\mathbf{Y} + \boldsymbol{b}$.

Důkaz. Viz Anděl (2007) str. 93. □

Víme, že obecně $\text{corr}(X,Y) = 0$ neimplikuje nezávislost náhodných veličín X,Y. Má-li ale náhodný vektor $(X,Y)'$ normální dvojrozměrné rozdělení, pak nulová korelace již charakterizuje nezávislost.

1.7. Věta. Nechť $(X,Y)' \sim \mathcal{N}_2\left(\begin{pmatrix} \mu_X \\ \mu_Y \end{pmatrix}, \begin{pmatrix} \sigma_X^2 & \rho\sigma_X\sigma_Y \\ \rho\sigma_X\sigma_Y & \sigma_Y^2 \end{pmatrix}\right)$, pak $\rho = 0$ právě tehdy, když X,Y jsou nezávislé.

Důkaz. Implikace ($\Leftarrow$) plyne okamžitě z Tvrzení 1.2, pro opačnou implikaci ($\Rightarrow$) bez újmy na obecnosti předpokládejme, že $\mu_X = \mu_Y = 0$ a $\sigma_X = \sigma_Y = 1$, jinak použijeme tvrzení na vektor $((X-\mu_X)/\sigma_X,(Y-\mu_Y)/\sigma_Y)$. Pak marginální hustoty f_X,f_Y se shodují s hustotou rozdělení $\mathcal{N}(0,1)$. Čili

$$f_X(x)f_Y(y) = \frac{1}{2\pi}\exp\left\{-\frac{x^2+y^2}{2}\right\} = f_{(X,Y)'}(x,y),\ \forall (x,y)' \in \mathbb{R}^2,$$

proto jsou veličiny X,Y nezávislé. □

1.8. Věta. Nechť $(X_1,Y_1)',\ldots,(X_n,Y_n)'$ je náhodný výběr z dvojrozměrného normálního rozdělení s kladnými rozptyly a nulovým korelačním koeficientem. Nechť $n \geq 3$, pak

$$T_n = \frac{r_n}{\sqrt{1-r_n^2}}\sqrt{n-2} \sim t_{n-2}, \tag{1.2}$$

kde t_{n-2} je Studentovo t-rozdělení o $n-2$ stupních volnosti.

Důkaz. Viz Anděl (2007) str. 94. □

Pomocí Věty 1.8 lze zkonstruovat test nezávislosti.

1.9. Test nezávislosti. Nechť $(X_1,Y_1)',\ldots,(X_n,Y_n)'$ je náhodný výběr z dvojrozměrného normálního rozdělení. Budeme testovat hypotézu, zda jsou náhodné veličiny X_1 a Y_1 nezávislé. Dle Věty 1.7 lze položit nulovou a alternativní hypotézu ve tvaru:

$$H_0 : \text{corr}(X,Y) = 0 \quad \text{vs.} \quad H_1 : \text{corr}(X,Y) \neq 0.$$

Testovací statistikou bude náhodná veličina T_n ze vztahu (1.2), která má za platnosti nulové hypotézy Studentovo t rozdělení o $n-2$ stupněch volnosti. Hypotézu H_0 proto zamítáme na hladině $\alpha \in (0,1)$ právě tehdy, když

$$|T_n| \geq u_{1-\alpha/2}.$$

Pro hledání intervalů spolehlivosti nebo testování hypotéz pro nenulový korelační koeficient nám pomůže *Fisherova Z-transformace*.

1.10. Věta (Fisherova Z-transformace). Nechť $(X_1,Y_1)',\ldots,(X_n,Y_n)'$ je náhodný výběr z dvojrozměrného normálního rozdělení s kladnými rozptyly a korelačním koeficientem ρ. Nechť $n \geq 4$, pak

$$\lim_{n\to\infty} \mathrm{P}\left(\sqrt{n-3}\left(\frac{1}{2}\log\frac{1+r_n}{1-r_n} - \frac{1}{2}\log\frac{1+\rho}{1-\rho}\right) \leq x\right) = \Phi(x), \quad \forall x \in \mathbb{R},$$

kde Φ je distribuční funkce rozdělení $\mathcal{N}(0,1)$.

Důkaz. Viz Příklad 2.10 na str. 10. □

Kapitola 2

Statistická inference o korelačním koeficientu

2.1 Asymptotické rozdělení

Všechny věty v předchozí částí předpokládaly, že náhodné výběry měli dvojrozměrné normální rozdělení. Přirozenou otázkou je, zda jsou veškeré intervaly spolehlivosti a kritické obory v jistém smyslu *v pořádku* i když výběr nepochází z normálního rozdělení? Prvním krokem k zodpovězení této otázky je určení asymptotického rozdělení veličiny r_n, což je tématem této sekce.

2.1. Definice (Konvergence v distribuci). Nechť $k \in \mathbb{N}$ a $\{\mathbf{X}_n\}_{n=1}^{\infty}$ je posloupnost k-rozměrných reálných náhodných vektorů a nechť $\mathbf{X}$ je reálný k-rozměrný náhodný vektor. Řekneme, že posloupnost $\mathbf{X}_n$ **konverguje v distribuci** k $\mathbf{X}$ právě tehdy, když pro každou spojitou omezenou funkci $f : \mathbb{R}^k \to \mathbb{R}$ platí:

$$\lim_{n\to\infty} \mathbb{E}[f(\mathbf{X}_n)] = \mathbb{E}[f(\mathbf{X})].$$

Konvergenci v distribuci budeme značit symbolicky $\mathbf{X}_n \xrightarrow{d} \mathbf{X}$ pro $n \to \infty$. Má-li speciálně $\mathbf{X} \sim \mathcal{N}_k(\boldsymbol{\mu},\mathbf{V})$, pak budeme rovnou psát $\mathbf{X}_n \xrightarrow{d} \mathcal{N}_k(\boldsymbol{\mu},\mathbf{V})$.

2.2. Věta. Nechť $k,l \in \mathbb{N}$ a $\{\mathbf{X}_n\}_{n=1}^{\infty}$ je posloupnost k-rozměrných reálných náhodných vektorů a nechť $\mathbf{X}$ je reálný k-rozměrný náhodný vektor splňující

$$\mathbf{X}_n \xrightarrow{d} \mathbf{X} \text{ pro } n \to \infty.$$

Dále nechť $g : (\mathbb{R}^k,\mathcal{B}_k) \to (\mathbb{R}^l,\mathcal{B}_l)$ je spojité měřitelné zobrazení, kde $\mathcal{B}_k$ je k-rozměrná borelovská σ-algebra, pak

$$g(\mathbf{X}_n) \xrightarrow{d} g(\mathbf{X}) \text{ pro } n \to \infty.$$

Důkaz. Buď $f : \mathbb{R}^l \to \mathbb{R}$ spojitá omezená funkce, pak funkce $f \circ g$ je spojitá a omezená, tedy platí:

$$\lim_{n\to\infty} \mathbb{E}[f(g(\mathbf{X}_n))] = \lim_{n\to\infty} \mathbb{E}[f \circ g(\mathbf{X}_n)] = \mathbb{E}[f \circ g(\mathbf{X})] = \mathbb{E}[f(g(\mathbf{X}))].$$

□

Vztah mezi konvergencí náhodných veličín v distribuci a bodovou konvergencí jejich distribučních funkcí popisuje následující věta.

2.3. Věta. Nechť $\{\mathbf{X}_n\}_{n=1}^{\infty}$ je posloupnost k-rozměrných reálných náhodných vektorů a nechť pro každé $n \in \mathbb{N}$ je $F_{\mathbf{X}_n}$ distribuční funkce náhodného vektoru $\mathbf{X}_n$. Je-li $\mathbf{X}$ reálný k-rozměrný náhodný vektor a $F_{\mathbf{X}}$ jeho distribuční funkce, pak platí:

$$\mathbf{X}_n \xrightarrow{d} \mathbf{X} \text{ pro } n \to \infty \Leftrightarrow F_{\mathbf{X}_n}(\boldsymbol{x}) \overset{n\to\infty}{\to} F_{\mathbf{X}}(\boldsymbol{x}),\ \boldsymbol{x} \in C_{\mathbf{X}},$$

kde $C_{\mathbf{X}} = \{\boldsymbol{x} \in \mathbb{R}^k; F_{\mathbf{X}} \text{ je spojitá v bodě } \boldsymbol{x}\}$.

Důkaz. Viz Van der Vaart (1998) str. 6, Lemma 2.2 (Portmanteau). □

Nejdůležitejším nástrojem ve studiu asymptotických statistik jsou centrální limitní věty (v dalším textu jenom zkráceně CLV). Uvedeme si Lévy-Lindebergovu CLV.

2.4. Věta (Lévy-Lindebergova k-rozměrná centrální limitní věta). Nechť $k \in \mathbb{N}$ a $\{\mathbf{X}_n\}_{n=1}^{\infty}$ je posloupnost nezávislých a stejně rozdělených k-rozměrných reálných náhodných vektorů. Nechť všechny složky vektoru $\boldsymbol{\mu} = \left(\mathbb{E}X_1^{(1)}, \ldots, \mathbb{E}X_1^{(k)}\right)'$ a matice $\Sigma = \left(\mathrm{cov}(X_1^{(i)}, X_1^{(j)})\right)_{i,j=1}^{k}$ jsou konečné. Pak

$$\frac{1}{\sqrt{n}} \sum_{i=1}^{n} (\mathbf{X}_i - \boldsymbol{\mu}) \xrightarrow{d} \mathcal{N}_k(\mathbf{0}, \Sigma) \text{ pro } n \to \infty.$$

Důkaz. Viz Lehmann (1999) str. 313. □

K určení asymptotického rozdělení výběrového korelačního koeficientu nám pomůže delta metoda.

2.5. Věta (k-rozměrná delta metoda). Nechť $k \in \mathbb{N}$ a $\{\mathbf{X}_n\}_{n=1}^{\infty}$ je posloupnost náhodných k-rozměrných reálných vektorů splňujících

$$\sqrt{n}\,(\mathbf{X}_n - \boldsymbol{\mu}) \xrightarrow{d} \mathcal{N}_k(\mathbf{0}, \Sigma) \text{ pro } n \to \infty,$$

kde $\boldsymbol{\mu} \in \mathbb{R}^k$ a $(\sigma_{pq})_{p,q=1}^{k} = \Sigma \in \mathbb{R}^{k\times k}$ je varianční matice. Dále buď $\phi : (\mathbb{R}^k, \mathcal{B}_k) \to (\mathbb{R}, \mathcal{B})$ měřitelné zobrazení, které má totální diferenciál v bodě $\boldsymbol{\mu}$. Pak

$$\sqrt{n}\,(\phi(\mathbf{X}_n) - \phi(\boldsymbol{\mu})) \xrightarrow{d} \mathcal{N}(0, \sigma^2) \text{ pro } n \to \infty,$$

kde

$$\sigma^2 = \sum_{p=1}^{k}\sum_{q=1}^{k}\sigma_{pq}\frac{\partial\phi}{\partial x_p}(\boldsymbol{\mu})\frac{\partial\phi}{\partial x_q}(\boldsymbol{\mu}).$$

Označíme-li symbolem $\nabla\phi(\boldsymbol{\mu})$ gradient funkce ϕ v bodě $\boldsymbol{\mu}$, tj. $\nabla\phi(\boldsymbol{\mu}) = \left(\frac{\partial\phi}{\partial x_1}(\boldsymbol{\mu}),\ldots,\frac{\partial\phi}{\partial x_k}(\boldsymbol{\mu})\right)'$, pak σ^2 má tvar

$$\sigma^2 = \nabla\phi(\boldsymbol{\mu})'\cdot\Sigma\cdot\nabla\phi(\boldsymbol{\mu}).$$

Důkaz. Viz Lehmann (1999) str.315. □

2.6. Poznámka. K odvození asymptotického rozdělení náhodné veličiny r_n využijeme obě předchozí věty. Předtím si ale upravíme vyjádření výběrových rozptylů a výběrové kovariance, tj.

$$s_{\mathbf{XY}} = \frac{1}{n-1}\sum_{i=1}^{n}(X_i-\overline{X}_n)(Y_i-\overline{Y}_n) = \frac{1}{n-1}\left(\sum_{i=1}^{n}X_iY_i - n\overline{X}_n\overline{Y}_n\right),$$
$$s^2_{\mathbf{X}} = \frac{1}{n-1}\sum_{i=1}^{n}\left(X_i-\overline{X}_n\right)^2 = \frac{1}{n-1}\left(\sum_{i=1}^{n}X_i^2 - n\overline{X}_n^2\right) \text{ a}$$
$$s^2_{\mathbf{Y}} = \frac{1}{n-1}\sum_{i=1}^{n}\left(Y_i-\overline{Y}_n\right)^2 = \frac{1}{n-1}\left(\sum_{i=1}^{n}Y_i^2 - n\overline{Y}_n^2\right).$$

Dosadíme-li vztahy do definice (1.1) výběrového korelačního koeficientu, získame alternativní vyjádření ve tvaru

$$r_n = \frac{\frac{n}{n-1}\left(\frac{1}{n}\sum_{i=1}^{n}X_iY_i - \overline{X}_n\overline{Y}_n\right)}{\sqrt{\frac{n}{n-1}\left(\frac{1}{n}\sum_{i=1}^{n}X_i^2-\overline{X}_n^2\right)}\sqrt{\frac{n}{n-1}\left(\frac{1}{n}\sum_{i=1}^{n}Y_i^2-\overline{Y}_n^2\right)}}$$
$$= \frac{\frac{1}{n}\sum_{i=1}^{n}X_iY_i - \frac{1}{n}\sum_{i=1}^{n}X_i\frac{1}{n}\sum_{i=1}^{n}Y_i}{\sqrt{\frac{1}{n}\sum_{i=1}^{n}X_i^2-\left(\frac{1}{n}\sum_{i=1}^{n}X_i\right)^2}\sqrt{\frac{1}{n}\sum_{i=1}^{n}Y_i^2-\left(\frac{1}{n}\sum_{i=1}^{n}Y_i\right)^2}}.$$

Označíme-li $\overline{X}_n = \frac{1}{n}\sum_{i=1}^{n}X_i$, $\overline{Y}_n = \frac{1}{n}\sum_{i=1}^{n}Y_i$, $\overline{X_n^2} = \frac{1}{n}\sum_{i=1}^{n}X_i^2$, $\overline{Y_n^2} = \frac{1}{n}\sum_{i=1}^{n}Y_i^2$ a $\overline{XY_n} = \frac{1}{n}\sum_{i=1}^{n}X_iY_i$, dostaneme jednodušší tvar výběrového korelačního koeficientu:

$$r_n = \frac{\overline{XY_n} - \overline{X}_n\overline{Y}_n}{\sqrt{\overline{X_n^2}-\overline{X}_n^2}\sqrt{\overline{Y_n^2}-\overline{Y}_n^2}}. \tag{2.1}$$

2.7. Věta (o asymptotickém rozdělení výběrového korelačního koeficientu). Nechť $n\in\mathbb{N}$ a $(X_1,Y_1)',\ldots,(X_n,Y_n)'$ je náhodný výběr z libovolného dvojrozměrného rozdělení. Nechť $\mathbb{E}|X_1|^4<+\infty$ a $\mathbb{E}|Y_1|^4<+\infty$, $\mathrm{corr}(X_1,Y_1)=\rho$ a nechť $|\rho|<1$. Pak

$$\sqrt{n}(r_n-\rho)\xrightarrow{d}\mathcal{N}\left(0,(\sigma^*)^2\right) \text{ pro } n\to\infty,$$

kde $(\sigma^*)^2$ splňuje

$$(\sigma^*)^2 = v' \begin{pmatrix} \mathrm{cov}(\tilde{X}^2,\tilde{X}^2) & \mathrm{cov}(\tilde{X}^2,\tilde{Y}^2) & \mathrm{cov}(\tilde{X}^2,\tilde{X}\tilde{Y}) \\ \mathrm{cov}(\tilde{Y}^2,\tilde{X}^2) & \mathrm{cov}(\tilde{Y}^2,\tilde{Y}^2) & \mathrm{cov}(\tilde{Y}^2,\tilde{X}\tilde{Y}) \\ \mathrm{cov}(\tilde{X}\tilde{Y},\tilde{X}^2) & \mathrm{cov}(\tilde{X}\tilde{Y},\tilde{Y}^2) & \mathrm{cov}(\tilde{X}\tilde{Y},\tilde{X}\tilde{Y}) \end{pmatrix} v, \qquad (2.2)$$

$$v = \left(-\frac{\rho}{2}, -\frac{\rho}{2}, 1\right)',$$

kde $\tilde{X} = \dfrac{X_1 - \mathbb{E}X_1}{\sqrt{\mathrm{var}(X_1)}}$ a $\tilde{Y} = \dfrac{Y_1 - \mathbb{E}Y_1}{\sqrt{\mathrm{var}(Y_1)}}$.

Důkaz. Protože vektory $(X_1,Y_1)', \ldots, (X_n,Y_n)'$ jsou stejně rozdělené, tak pro zjednodušení zápisu budeme psát místo X_1 a Y_1 jenom X a Y. Bez újmy na obecnosti předpokládejme, že $\mathbb{E}X = \mathbb{E}Y = 0$ a $\mathrm{var}(X) = \mathrm{var}(Y) = 1$. Kdyby tomu tak nebylo, pak stačí uvažovat náhodné veličiny $\tilde{X}$ a $\tilde{Y}$ místo X a Y, neboť $\mathbb{E}\tilde{X} = \mathbb{E}\tilde{Y} = 0$, $\mathrm{var}(\tilde{X}) = \mathrm{var}(\tilde{Y}) = 1$, $\mathbb{E}|\tilde{X}|^4 < +\infty$, $\mathbb{E}|\tilde{Y}|^4 < +\infty$, z Tvrzení 1.2 na str. 1 platí

$$\mathrm{corr}\left(\tilde{X},\tilde{Y}\right) = \mathrm{sgn}\left(\frac{1}{\sqrt{\mathrm{var}(X)\mathrm{var}(Y)}}\right)\mathrm{corr}(X,Y) = \rho,$$

a z Tvrzení 1.6 na str. 2 platí

$$r_n(\tilde{\mathbf{X}},\tilde{\mathbf{Y}}) = r_n\left(\frac{\mathbf{X} - \mathbb{E}(\mathbf{X})}{\sqrt{\mathrm{var}(X)}}, \frac{\mathbf{Y} - \mathbb{E}(\mathbf{Y})}{\sqrt{\mathrm{var}(Y)}}\right) = r_n(\mathbf{X},\mathbf{Y}).$$

Položme $\overline{\mathbf{Z}}_n = \left(\overline{X}_n, \overline{Y}_n, \overline{X_n^2}, \overline{Y_n^2}, \overline{XY_n}\right)'$ dle Poznámky 2.6. Platí:

$$\begin{aligned} \mathbb{E}X &= 0, & \mathbb{E}Y &= 0, \\ \mathbb{E}X^2 &= \mathrm{var}(X) + (\mathbb{E}X)^2 = 1, & \mathbb{E}Y^2 &= 1, \\ \mathbb{E}XY &= \mathrm{cov}(X,Y) + \mathbb{E}X\mathbb{E}Y = \rho. \end{aligned}$$

Dále položme $\boldsymbol{\mu} = (0,0,1,1,\rho)'$. Protože jsou náhodné veličiny $X_1, \ldots, X_n$ nezávislé a stejně rozdělené, pak jsou taky veličiny $X_1^2, \ldots, X_n^2$ nezávislé a stejně rozdělené. Analogicky pro veličiny $Y_1^2, \ldots, Y_n^2$ a $X_1Y_1, \ldots, X_nY_n$ a proto jsou náhodné vektory $\mathbf{Z}_1, \ldots, \mathbf{Z}_n$ nezávislé a stejně rozdělené, kde

$$\mathbf{Z}_i = \left(X_i, Y_i, X_i^2, Y_i^2, X_iY_i\right)'.$$

Pak dle Věty 2.4 plyne, že

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n}(\mathbf{Z}_i - \boldsymbol{\mu}) = \sqrt{n}\left(\overline{\mathbf{Z}}_n - \boldsymbol{\mu}\right) \xrightarrow{d} \mathcal{N}_5(\mathbf{0},\Sigma) \text{ pro } n \to \infty,$$

kde Σ je matice tvaru:

$$\Sigma = \begin{pmatrix} \mathrm{cov}(X,X) & \mathrm{cov}(X,Y) & \mathrm{cov}(X,X^2) & \mathrm{cov}(X,Y^2) & \mathrm{cov}(X,XY) \\ \mathrm{cov}(Y,X) & \mathrm{cov}(Y,Y) & \mathrm{cov}(Y,X^2) & \mathrm{cov}(Y,Y^2) & \mathrm{cov}(Y,XY) \\ \mathrm{cov}(X^2,X) & \mathrm{cov}(X^2,Y) & \mathrm{cov}(X^2,X^2) & \mathrm{cov}(X^2,Y^2) & \mathrm{cov}(X^2,XY) \\ \mathrm{cov}(Y^2,X) & \mathrm{cov}(Y^2,Y) & \mathrm{cov}(Y^2,X^2) & \mathrm{cov}(Y^2,Y^2) & \mathrm{cov}(Y^2,XY) \\ \mathrm{cov}(XY,X) & \mathrm{cov}(XY,Y) & \mathrm{cov}(XY,X^2) & \mathrm{cov}(XY,Y^2) & \mathrm{cov}(XY,XY) \end{pmatrix}.$$

Položme $\phi : (\mathbb{R}^5,\mathcal{B}_5) \to (\mathbb{R},\mathcal{B})$ měřitelné zobrazení předpisem

$$\phi(\mathbf{x}) = \frac{x_5 - x_1x_2}{\sqrt{x_3 - x_1^2}\sqrt{x_4 - x_2^2}}, \text{ kde } \mathbf{x} = (x_1,x_2,x_3,x_4,x_5)' \in \mathbb{R}^5.$$

Pak jednoduchým dosazením dostávame, že

$$\phi(\overline{\mathbf{Z}}_n) = \frac{\overline{XY}_n - \overline{X}_n\overline{Y}_n}{\sqrt{\overline{X_n^2} - \overline{X}_n^2}\sqrt{\overline{Y_n^2} - \overline{Y}_n^2}} = r_n \text{ dle (2.1) a } \phi(\boldsymbol{\mu}) = \frac{\rho}{\sqrt{1}\sqrt{1}} = \rho.$$

Zobrazení ϕ má spojité všechny parciální derivace v bodě $\boldsymbol{\mu}$ a tudíž má totální diferenciál v bodě $\boldsymbol{\mu}$ a platí

$$\begin{aligned}
\frac{\partial\phi}{\partial x_1}(\boldsymbol{\mu}) &= \left.\frac{x_1x_5 - x_2x_3}{(x_3 - x_1^2)^{\frac{3}{2}}\sqrt{x_4 - x_2^2}}\right|_{\mathbf{x}=\boldsymbol{\mu}} = 0, \\
\frac{\partial\phi}{\partial x_2}(\boldsymbol{\mu}) &= \left.\frac{x_2x_5 - x_1x_4}{(x_4 - x_2^2)^{\frac{3}{2}}\sqrt{x_1 - x_3^2}}\right|_{\mathbf{x}=\boldsymbol{\mu}} = 0, \\
\frac{\partial\phi}{\partial x_3}(\boldsymbol{\mu}) &= \left.\frac{x_5 - x_1x_2}{2(x_3 - x_1^2)^{\frac{3}{2}}\sqrt{x_4 - x_2^2}}\right|_{\mathbf{x}=\boldsymbol{\mu}} = -\frac{\rho}{2}, \\
\frac{\partial\phi}{\partial x_4}(\boldsymbol{\mu}) &= \left.\frac{x_5 - x_1x_2}{2(x_4 - x_2^2)^{\frac{3}{2}}\sqrt{x_3 - x_1^2}}\right|_{\mathbf{x}=\boldsymbol{\mu}} = -\frac{\rho}{2}, \\
\frac{\partial\phi}{\partial x_5}(\boldsymbol{\mu}) &= \left.\frac{1}{\sqrt{x_3 - x_1^2}\sqrt{x_4 - x_2^2}}\right|_{\mathbf{x}=\boldsymbol{\mu}} = 1.
\end{aligned}$$

Konečně z Věty 2.5 dostaneme

$$\sqrt{n}(\phi(\overline{\mathbf{Z}}_n) - \phi(\boldsymbol{\mu})) = \sqrt{n}(r_n - \rho) \xrightarrow{d} \mathcal{N}\left(0,(\sigma^*)^2\right) \text{ pro } n \to \infty,$$

kde $(\sigma^*)^2 = \nabla\phi(\boldsymbol{\mu})' \cdot \Sigma \cdot \nabla\phi(\boldsymbol{\mu})$. Protože gradient $\nabla\phi(\boldsymbol{\mu})$ má první dvě složky nulové, $(\sigma^*)^2$ pak nezávisí na prvních dvou řádcích a sloupcích matice Σ. Tím pádem konečně dostávame vztah (2.2). □

2.8. Důsledek. Nechť $(X_1,Y_1)',\ldots,(X_n,Y_n)'$ je náhodný výběr a nechť jsou X_1 a Y_1 nezávislé náhodné veličiny s konečnými čtvrtými momenty. Pak

$$\sqrt{n}r_n \xrightarrow{d} \mathcal{N}(0,1), \text{ pro } n \to \infty.$$

Důkaz. Protože jsou složky vektorů nezávislé, pak $\rho = 0$ dle Tvrzení 1.2 na str. 1. Dále dle Věty 2.7 platí

$$\sqrt{n}r_n \xrightarrow{d} \mathcal{N}(0,\text{var}(\tilde{X}_1\tilde{Y}_1)) \text{ pro } n \to \infty, \text{ kde } \tilde{X} = \frac{X_1 - \mathbb{E}X_1}{\sqrt{\text{var}(X_1)}},\ \tilde{Y} = \frac{Y_1 - \mathbb{E}Y_1}{\sqrt{\text{var}(Y_1)}}.$$

Opět užitím nezávislosti dostaneme $\text{var}(\tilde{X}_1\tilde{Y}_1) = \mathbb{E}\tilde{X}_1^2\tilde{Y}_1^2 - (\mathbb{E}\tilde{X}_1\tilde{Y}_1)^2 = \mathbb{E}\tilde{X}_1^2\mathbb{E}\tilde{Y}_1^2 = 1$. □

2.9. Poznámka. Předpoklad $|\rho| < 1$ ve Větě 2.7. není extrémně omezující, jenom vylučuje nepraktické případy. Kdyby totiž $|\rho| = 1$, pak dle Tvrzení 1.2 na str. 1 najdeme $a,b \in \mathbb{R}$: $Y = aX + b$ skoro jistě. Pak ale z Tvrzení 1.6 na str. 2 plyne:

$$\begin{aligned} r_n(\mathbf{X},a\mathbf{X}+b) &= \text{sgn}(a)r_n(\mathbf{X},\mathbf{X}) = \text{sgn}(a)\frac{\frac{1}{n-1}\sum_{i=1}^n (X_i - \overline{X}_n)^2}{\sqrt{\frac{1}{n-1}\sum_{i=1}^n (X_i - \overline{X}_n)^2}\sqrt{\frac{1}{n-1}\sum_{i=1}^n (X_i - \overline{X}_n)^2}} \\ &= \text{sgn}(a). \end{aligned}$$

Pak nemá vůbec smysl hledat asymptotické rozdělení r_n, proto tento případ neuvažujeme.

V následujících paragrafech najdeme asymptotické rozdělení korelačního koeficientu pro specifická dvojrozměrná rozdělení.

2.10. Příklad. Podíváme se na případ, kdy budeme mít náhodný výběr z normálního dvojrozměrného rozdělení. Předpokládejme, že $(X_1,Y_1)',\ldots,(X_n,Y_n)'$ je náhodný výběr z rozdělení $\mathcal{N}_2\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}\right)$. Pak $\mathbb{E}X^4 = \mathbb{E}Y^4 = 3$, $\mathbb{E}X^2Y^2 = 2\rho^2 + 1$ a $\mathbb{E}XY^3 = \mathbb{E}X^3Y = 3\rho$ a dle Věty 2.7. platí

$$\sqrt{n}(r_n - \rho) \xrightarrow{d} \mathcal{N}\left(0,(\sigma^*)^2\right) \text{ pro } n \to \infty,$$

kde

$$(\sigma^*)^2 = \left(-\frac{\rho}{2},-\frac{\rho}{2},1\right)\begin{pmatrix} 2 & 2\rho^2 & 2\rho \\ 2\rho^2 & 2 & 2\rho \\ 2\rho & 2\rho & \rho^2+1 \end{pmatrix}\begin{pmatrix} -\frac{\rho}{2} \\ -\frac{\rho}{2} \\ 1 \end{pmatrix} = \left(1-\rho^2\right)^2.$$

Teď chceme najít transformaci, která stabilizuje rozptyl, tj. hledáme takovou měřitelnou funkci $g : (\mathbb{R},\mathcal{B}) \to (\mathbb{R},\mathcal{B})$, aby měla vlastní derivaci v bodě ρ a aby $g'(\rho)^2(1-\rho^2)^2 = 1$. My totiž chceme aby výslední limitní rozdělení bylo $\mathcal{N}(0,1)$, což nám tato transformace zaručí. Jedna z možností je aby platilo $g'(\rho) = \frac{1}{1-\rho^2}$. Spočteme tedy primitivní funkci rozkladem na parciální zlomky:

$$\int \frac{1}{1-\rho^2}\mathrm{d}\rho = \int \left(\frac{1/2}{1-\rho} + \frac{1/2}{1+\rho}\right)\mathrm{d}\rho = \frac{1}{2}\log\frac{1+\rho}{1-\rho} + C,\ \rho \in (-1,1).$$

Položme $g(\rho) = \frac{1}{2}\log\frac{1+\rho}{1-\rho}$, pak $g'(\rho)^2 = (1-\rho^2)^{-2}$, proto $g'(\rho)^2(1-\rho^2)^2 = 1$ a užitím Věty 2.5 dostaneme

$$\sqrt{n}(g(r_n) - g(\rho)) = \sqrt{n}\left(\frac{1}{2}\log\frac{1+r_n}{1-r_n} - \frac{1}{2}\log\frac{1+\rho}{1-\rho}\right) \xrightarrow{d} \mathcal{N}(0,1) \text{ pro } n \to \infty.$$

Tím jsme taky dokázali Větu 1.10 na str. 4, neboť $\sqrt{n-3}/\sqrt{n} \to 1$ pro $n \to \infty$.

2.11. Definice (Studentovo t-rozdělení). Nechť $k \in \mathbb{N}$, $\mathbf{X} \sim \mathcal{N}_k(0,\mathbf{V})$, nechť dále $\boldsymbol{\mu} \in \mathbb{R}^k$ a $Y \sim \chi^2_n$ kde $n \in \mathbb{N}$ přičemž $\mathbf{X}$ a Y jsou nezávislé. Pak rozdělení náhodné veličiny tvaru

$$\boldsymbol{\mu} + \frac{\mathbf{X}}{\sqrt{Y}}\sqrt{n}$$

nazýváme **k-rozměrné Studentovo t-rozdělení s n stupni volnosti** s parametry $\boldsymbol{\mu}$,$\mathbf{V}$. Vícerozměrné Studentovo t-rozdělení budeme značit $t_k(\boldsymbol{\mu},\mathbf{V},n)$.

2.12. Věta (hustota Studentova t-rozdělení). Nechť $\mathbf{Z} \sim t_k(\boldsymbol{\mu},\mathbf{V},n)$, pak hustota náhodného vektoru $\mathbf{Z}$ má tvar

$$f_{\mathbf{Z}}(\mathbf{z}) = \frac{\Gamma\left(\frac{n+k}{2}\right)}{\Gamma\left(\frac{n}{2}\right)(n\pi)^{k/2}\det(\mathbf{V})^{1/2}}\left(1 + \frac{1}{n}(\mathbf{z}-\boldsymbol{\mu})'\mathbf{V}^{-1}(\mathbf{z}-\boldsymbol{\mu})\right)^{-\frac{n+k}{2}}.$$

Důkaz. Viz Kotz a Nadarajah (2004). □

2.13. Příklad. Nechť $(X_1,Y_1)', \ldots, (X_n,Y_n)'$ je náhodný výběr z rozdělení $t_2(\mathbf{0},\mathbf{V},m)$, kde $m \in \mathbb{N} : m \geq 5$ a $\mathbf{V} = \frac{m-2}{m}\begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}$, pak platí:

$$\mathbb{E}X^4 = \mathbb{E}Y^4 = \frac{3(m-2)}{m-4},\ \mathbb{E}X^2Y^2 = \frac{(m-2)(2\rho^2+1)}{m-4} \text{ a } \mathbb{E}XY^3 = \mathbb{E}X^3Y = \frac{3(m-2)\rho}{m-4}.$$

Spočteme rozptyl asymptotického rozdělení výběrového korelačního koeficientu podle Věty 2.7:

$$(\sigma^*)^2 = \left(-\frac{\rho}{2}, -\frac{\rho}{2}, 1\right)\begin{pmatrix} \frac{2(m-1)}{m-4} & \frac{2(m-2)\rho^2+2}{m-4} & \frac{2(m-1)\rho}{m-4} \\ \frac{2(m-2)\rho^2+2}{m-4} & \frac{2(m-1)}{m-4} & \frac{2(m-1)\rho}{m-4} \\ \frac{2(m-1)\rho}{m-4} & \frac{2(m-1)\rho}{m-4} & \frac{m\rho^2+m-2}{m-4} \end{pmatrix}\begin{pmatrix} -\frac{\rho}{2} \\ -\frac{\rho}{2} \\ 1 \end{pmatrix} = \frac{(m-2)\left(\rho^2-1\right)^2}{m-4}.$$

Máme-li náhodný výběr z dvojrozměrného Studentova rozdělení s m-stupni volnosti, pak $\sqrt{n}r_n$ má asymptoticky rozdělení $\mathcal{N}\left(\rho, \frac{(m-2)\left(1-\rho^2\right)^2}{m-4}\right)$. Ještě poznamenejme, že pro $m \to \infty$ se asymptotické rozdělení shoduje s rozdělením v Příkladu 2.10. Tento fakt bylo možné

očekávat, neboť víme, že pro vysoký stupeň volnosti lze Studentovo t-rozdělení aproximovat normálním rozdělením.

2.14. Příklad Nechť $(X_1,Y_1)',\ldots,(X_n,Y_n)'$ je náhodný výběr z dvojrozměrného Poissonova rozdělení s parametry $\lambda,\lambda_1,\lambda_2 > 0$, tj. z rozdělení s pravděpodobnostní funkcí tvaru

$$\mathrm{P}(X = k_1,Y = k_2) = e^{-\lambda-\lambda_1-\lambda_2}\frac{\lambda_1^{k_1}}{k_1!}\frac{\lambda_2^{k_2}}{k_2!}\sum_{j=0}^{\min(k_1,k_2)}\binom{k_1}{j}\binom{k_2}{j}j!\left(\frac{\lambda}{\lambda_1\lambda_2}\right)^j,\ k_1,k_2 \in \mathbb{N}\cup\{0\}.$$

Pak $\mathbb{E}X = \mathrm{var}(X) = \lambda+\lambda_1$ a $\mathbb{E}Y = \mathrm{var}(Y) = \lambda+\lambda_2$ a budeme uvažovat náhodné veličiny $\frac{X_i-\lambda-\lambda_1}{\sqrt{\lambda+\lambda_1}},\frac{Y_i-\lambda-\lambda_2}{\sqrt{\lambda+\lambda_2}}$ pro $i \in \{1,\ldots,n\}$. Pak výpočtem dostaneme:

$$\begin{aligned}(\sigma^*)^2 &= \begin{pmatrix} -\frac{\rho}{2} \\ -\frac{\rho}{2} \\ 1 \end{pmatrix}\begin{pmatrix} \frac{2\lambda+2\lambda_1+1}{\lambda+\lambda_1} & \frac{\lambda(2\lambda+1)}{(\lambda+\lambda_1)(\lambda+\lambda_2)} & \frac{\lambda(2\lambda+2\lambda_1+1)}{(\lambda+\lambda_1)^{3/2}\sqrt{\lambda+\lambda_2}} \\ \frac{\lambda(2\lambda+1)}{(\lambda+\lambda_1)(\lambda+\lambda_2)} & \frac{2\lambda+2\lambda_2+1}{\lambda+\lambda_2} & \frac{\lambda(2\lambda+2\lambda_2+1)}{\sqrt{\lambda+\lambda_1}(\lambda+\lambda_2)^{3/2}} \\ \frac{\lambda(2\lambda+2\lambda_1+1)}{(\lambda+\lambda_1)^{3/2}\sqrt{\lambda+\lambda_2}} & \frac{\lambda(2\lambda+2\lambda_2+1)}{\sqrt{\lambda+\lambda_1}(\lambda+\lambda_2)^{3/2}} & \frac{\lambda_1(\lambda+\lambda_2)+\lambda(2\lambda+\lambda_2+1)}{(\lambda+\lambda_1)(\lambda+\lambda_2)} \end{pmatrix}\left(-\frac{\rho}{2},-\frac{\rho}{2},1\right) \\ &= \frac{\lambda_2\left(4\lambda_2+1\right)\lambda^2+\lambda_1\left(8\lambda_2^2+(8\lambda+4)\lambda_2+\lambda\right)\lambda+4\lambda_1^2\left(\lambda+\lambda_2\right)^2}{4\left(\lambda+\lambda_1\right)^2\left(\lambda+\lambda_2\right)^2},\end{aligned}$$

kde jsme využili, že $\rho = \frac{\lambda}{\sqrt{\lambda+\lambda_1}\sqrt{\lambda+\lambda_2}}$. Speciálně pro případ $\lambda = \lambda_1 = \lambda_2$ máme $(\sigma^*)^2 = \frac{3(6\lambda+1)}{32\lambda}$. Více o dvojrozměrném Poissonově rozdělení lze nalézt např. v Kocherlakota a Kocherlakota (2006).

2.2 Testy nezávislosti

Podíváme se, co se stane s testem pro nezávislost náhodných veličín z Věty 1.8 na str. 3, bude-li porušen předpoklad normality.

2.15. Věta (o porušení I). Nechť $(X_1,Y_1)',\ldots,(X_n,Y_n)'$ je náhodný výběr z libovolného dvojrozměrného rozdělení s rozsahem alespoň 3 a nechť X_1,Y_1 jsou nezávislé náhodné veličiny s konečnými čtvrtými momenty. Pak

$$\mathrm{P}\left(\left|\frac{r_n}{\sqrt{1-r_n^2}}\sqrt{n-2}\right| \geq t_{n-2}(1-\alpha/2)\right) \to \alpha \text{ pro } n \to \infty.$$

Důkaz. Z Důsledku 2.8 víme, že

$$\sqrt{n}r_n \xrightarrow{d} \mathcal{N}(0,1) \text{ pro } n \to \infty.$$

Dále z Věty 2.2 víme, že

$$\frac{\sqrt{1-\rho^2}}{\sqrt{1-r_n^2}} = \frac{1}{\sqrt{1-r_n^2}} \xrightarrow{d} 1 \Rightarrow \frac{1}{\sqrt{1-r_n^2}} \xrightarrow{p} 1 \text{ pro } n \to \infty.$$

Užitím Cramér-Sluckého věty dostaneme:

$$\sqrt{n-2}\frac{r_n}{\sqrt{1-r_n^2}} = \underbrace{\sqrt{n}r_n}_{\xrightarrow{d}\mathcal{N}(0,1)} \underbrace{\frac{1}{\sqrt{1-r_n^2}}}_{\xrightarrow{p}1} \frac{\sqrt{n-2}}{\sqrt{n}} \xrightarrow{d} \mathcal{N}(0,1), \text{ pro } n\to\infty.$$

Z Věty 2.3 pak víme, že platí

$$\mathrm{P}\left(\left|\sqrt{n-2}\frac{r_n}{\sqrt{1-r_n^2}}\right| \le u_{1-\alpha/2}\right) \overset{n\to\infty}{\to} \Phi(u_{1-\alpha/2}) - \Phi(-u_{1-\alpha/2}) = 1-\alpha$$

a pro kvantily Studentova t-rozdělení platí $t_{n-2}(1-\alpha/2) \overset{n\to\infty}{\to} u_{1-\alpha/2}$. Tím konečně dostaneme výsledné tvrzení. □

Věta 1.8 na str. 3 sice tvrdí, že výsledné rozdělení je Studentovo t-rozdělení s $n-2$ stupni volnosti, ale pro dostatečně vysoký stupeň volnosti lze z CLV Studentovo rozdělení aproximovat normálním rozdělením $\mathcal{N}(0,1)$ tak, jak to bylo uvedeno v předchozím důkazu. Tím pádem Věta 2.15 nám napovídá, že pokud je předpoklad normality u dat porušen, ale náhodné veličiny jsou nezávislé, pak zůstava hladina asymptotického testu nezávislosti z paragrafu 1.9 na str. 3 zachována.

2.3 Fisherova Z-transformace

2.16. Poznámka (o porušení II). Uvažujme náhodný výběr $(X_1,Y_1)',\ldots,(X_n,Y_n)'$ z libovolného dvojrozměrného rozdělení a nechť je rozsah výběru alespoň 4. Z Věty 1.10 na str. 4 víme, že náhodná veličina

$$Fr_n = \frac{1}{2}\log\frac{1+r_n}{1-r_n}.$$

má za předpokladu normality asymptoticky normální rozdělení s jednotkovým rozptylem. Zjistíme, zda to platí obecně, tedy bez předpokladu normálního rozdělení. Z Věty 2.7. víme, že

$$\sqrt{n}(r_n-\rho) \xrightarrow{d} \mathcal{N}\left(0,[\sigma^*(\rho)]^2\right) \text{ pro } n\to\infty.$$

Pak z Věty 2.5. dostaneme

$$\sqrt{n}(Fr_n - F\rho) \xrightarrow{d} \mathcal{N}\left(0, \left(\frac{\sigma^*(\rho)}{1-\rho^2}\right)^2\right) \text{ pro } n\to\infty.$$

Obecně jistě nemůžeme tvrdit, že $\sigma^*(\rho) = 1-\rho^2$, čili taky nemůžeme tvrdit, že $\sqrt{n}(Fr_n - F\rho)$ má asymptoticky rozdělení $\mathcal{N}(0,1)$. Závěrem tedy je, že Věta 1.10 na str. 4 bez předpokladu normality neplatí. Tento výsledek bylo možné očekávat, neboť odvození Fisherove Z-transformace se odvíjelo od tvaru $(\sigma^*(\rho))^2$ v Příkladu 2.10, který závisí od rozdělení náhodného vektoru $(X_1,Y_1)'$.

2.4 Obecný přístup k intervalům spolehlivosti

2.17. Úkol. Již v předchozí sekci jsme zjistili, že veličina r_n se asymptoticky blíží k veličině s normálním rozdělením. Cílem této sekce bude vytvořit obecný asymptotický intervalový odhad r_n. Základní myšlenkou je nalézt odhad ξ_n parametru σ^* a to tak, aby

$$\frac{\xi_n}{\sigma^*} \xrightarrow{p} 1, \text{ pro } n \to \infty.$$

2.18. Věta (o konzistenci odhadu ξ_n). Nechť $(X_1,Y_1)', \ldots, (X_n,Y_n)'$ je náhodný výběr z dvojrozměrného rozdělení s konečnými čtvrtými momenty. Nechť $\xi_n^2 = \overline{v}_n' \cdot \overline{\Sigma}_n \cdot \overline{v}_n$, kde $\overline{\Sigma}_n$ je matice tvaru

$$\overline{\Sigma}_n = \begin{pmatrix} s_{\tilde{\mathbf{X}}^2\tilde{\mathbf{X}}^2} & s_{\tilde{\mathbf{X}}^2\tilde{\mathbf{Y}}^2} & s_{\tilde{\mathbf{X}}^2\tilde{\mathbf{X}}\tilde{\mathbf{Y}}} \\ s_{\tilde{\mathbf{Y}}^2\tilde{\mathbf{X}}^2} & s_{\tilde{\mathbf{Y}}^2\tilde{\mathbf{Y}}^2} & s_{\tilde{\mathbf{Y}}^2\tilde{\mathbf{X}}\tilde{\mathbf{Y}}} \\ s_{\tilde{\mathbf{X}}\tilde{\mathbf{Y}}\tilde{\mathbf{X}}^2} & s_{\tilde{\mathbf{X}}\tilde{\mathbf{Y}}\tilde{\mathbf{Y}}^2} & s_{\tilde{\mathbf{X}}\tilde{\mathbf{Y}}\tilde{\mathbf{X}}\tilde{\mathbf{Y}}} \end{pmatrix},$$

kde

$$s_{\tilde{\mathbf{X}}^2\tilde{\mathbf{Y}}^2} = \frac{1}{n-1}\sum_{i=1}^{n}\left(\tilde{X}_i^2 - \frac{1}{n}\sum_{j=1}^{n}\tilde{X}_j^2\right)\left(\tilde{Y}_i^2 - \frac{1}{n}\sum_{j=1}^{n}\tilde{Y}_j^2\right),\ \tilde{X}_i = \frac{X_i - \overline{X}_n}{s_{\mathbf{X}}},\ \tilde{Y}_i = \frac{Y_i - \overline{Y}_n}{s_{\mathbf{Y}}}$$

a zbylé výrazy jsou definovány analogicky s tím, že $\mathbf{\tilde{X}\tilde{Y}} = (\tilde{X}_i\tilde{Y}_i)_{i=1}^n$ a $\overline{v}_n = \left(-\frac{r_n}{2}, -\frac{r_n}{2}, 1\right)$. Pak $\xi_n^2 \xrightarrow{p} (\sigma^*)^2$ pro $n \to \infty$, kde $(\sigma^*)^2$ splňuje rovnici (2.2) z Věty 2.7.

Důkaz. Stačí ukázat, že $s_{\tilde{\mathbf{X}}\tilde{\mathbf{Y}}}$ je konzistentní odhad parametrické funkce $\frac{\mathbb{E}(X_1Y_1) - \mathbb{E}(X_1)\mathbb{E}(Y_1)}{\sqrt{\text{var}(X_1)}\sqrt{\text{var}(Y_1)}}$. Pro zbylé členy to lze provést analogicky. Tvrzení pak plyne z Věty 2.7. Platí:

$$\begin{aligned} s_{\tilde{\mathbf{X}}\tilde{\mathbf{Y}}} &= \frac{1}{n-1}\sum_{i=1}^{n}\left(\tilde{X}_i - \frac{1}{n}\sum_{j=1}^{n}\tilde{X}_j\right)\left(\tilde{Y}_i - \frac{1}{n}\sum_{j=1}^{n}\tilde{Y}_j\right) \\ &= \frac{1}{n-1}\sum_{i=1}^{n}\left(\frac{X_i - \overline{X}_n}{s_{\mathbf{X}}} - \frac{1}{n}\sum_{j=1}^{n}\frac{X_j - \overline{X}_n}{s_{\mathbf{X}}}\right)\left(\frac{Y_i - \overline{Y}_n}{s_{\mathbf{Y}}} - \frac{1}{n}\sum_{j=1}^{n}\frac{Y_j - \overline{Y}_n}{s_{\mathbf{Y}}}\right) \\ &= \frac{\frac{1}{n-1}\sum_{i=1}^{n}\left(X_iY_i - X_i\overline{Y}_n - Y_i\overline{X}_n + \overline{X}_n\overline{Y}_n\right)}{s_{\mathbf{X}}s_{\mathbf{Y}}} \\ &= \frac{1}{s_{\mathbf{X}}s_{\mathbf{Y}}}\left(\frac{1}{n-1}\sum_{i=1}^{n}X_iY_i - \frac{n}{n-1}\overline{X}_n\overline{Y}_n\right). \end{aligned}$$

Z Poznámky 2.6 víme, že $s_{\mathbf{X}}^2 = \frac{1}{n-1}\sum_{i=1}^{n}X_i^2 - \frac{n}{n-1}\overline{X}_n^2$ a proto ze silného zákona velkých čísel víme, že $\overline{X}_n \xrightarrow{p} \mathbb{E}(X_1)$, $s_{\mathbf{X}} \xrightarrow{p} \sqrt{\mathbb{E}(X_1^2) - (\mathbb{E}X_1)^2} = \sqrt{\text{var}(X_1)}$ a

$$\frac{1}{n-1}\sum_{i=1}^{n}X_iY_i \xrightarrow{p} \mathbb{E}(X_1Y_1) \text{ pro } n \to \infty.$$

Proto platí

$$s_{\tilde{\mathbf{X}}\tilde{\mathbf{Y}}} \xrightarrow{p} \frac{\mathbb{E}(X_1Y_1) - \mathbb{E}(X_1)\mathbb{E}(Y_1)}{\sqrt{\mathrm{var}(X_1)}\sqrt{\mathrm{var}(Y_1)}} = \mathrm{cov}\left(\frac{X_1 - \mathbb{E}(X_1)}{\sqrt{\mathrm{var}(X_1)}}, \frac{Y_1 - \mathbb{E}(Y_1)}{\sqrt{\mathrm{var}(Y_1)}}\right) \text{ pro } n \to \infty.$$

□

2.19. Věta (o asymptotickém intervalovém odhadu ρ). Nechť $(X_1,Y_1)', \ldots, (X_n,Y_n)'$ jsou nezávislé a stejně rozdělené náhodné vektory z dvojrozměrného rozdělení s konečnými čtvrtými momenty. Nechť ξ_n je odhad parametru σ^* z Věty 2.18 a r_n je výběrový korelační koeficient definován v (1.1). Buď $\alpha \in (0,1)$ a $\rho = \mathrm{corr}(X_1,Y_1)$, pak asymptotický intervalový odhad parametru ρ o spolehlivosti $1-\alpha$ je tvaru

$$\left(r_n - u_{1-\alpha/2}\frac{\xi_n}{\sqrt{n}}, r_n + u_{1-\alpha/2}\frac{\xi_n}{\sqrt{n}}\right), \tag{2.3}$$

kde u_β je β-kvantil $\mathcal{N}(0,1)$. Dále asymptotický dolní intervalový odhad parametru ρ o spolehlivosti $1-\alpha$ je tvaru

$$\left(r_n - u_{1-\alpha}\frac{\xi_n}{\sqrt{n}}, +\infty\right)$$

a asymptotický horní intervalový odhad parametru ρ o spolehlivosti $1-\alpha$ je ve tvaru

$$\left(-\infty, r_n + u_{1-\alpha}\frac{\xi_n}{\sqrt{n}}\right).$$

Důkaz. Z Věty 2.18. víme, že platí $\frac{\sigma^*}{\xi_n} \xrightarrow{p} 1$ pro $n \to \infty$. Tím pádem užitím Cramér-Sluckého Věty dostaneme:

$$\frac{\sqrt{n}(r_n - \rho)}{\xi_n} = \frac{\sqrt{n}(r_n - \rho)}{\sigma^*} \cdot \frac{\sigma^*}{\xi_n} \xrightarrow{d} \mathcal{N}(0,1).$$

Rozepsáním výše uvedené limity pomocí Věty 2.3 platí

$$\lim_{n\to\infty} \mathrm{P}\left(u_{\alpha/2} \le \frac{\sqrt{n}(r_n - \rho)}{\xi_n} \le u_{1-\alpha/2}\right) = 1 - \alpha$$

a úpravou nerovností dostaneme tvar (2.3). Analogickým postupem dostaneme vzorečky pro horní a dolní asymptotický intervalový odhad. □

Kombinací Fisherove Z-transformace a intervalového odhadu z Věty 2.19. sestrojíme ještě jednu variantu intervalu spolehlivosti pro korelační koeficient.

2.20. Věta (o kombinaci intervalových odhadů). Nechť $(X_1,Y_1)', \ldots, (X_n,Y_n)'$ je náhodný výběr z dvojrozměrného rozdělení s konečnými čtvrtými momenty. Je-li ρ korelační koeficient náhodných veličin X_1 a Y_1, pak

$$\left(\frac{1}{2}\log\left(\frac{1+r_n}{1-r_n}\right) - \frac{u_{1-\alpha/2}}{\sqrt{n}} \cdot \frac{\xi_n}{1-r_n^2}, \frac{1}{2}\log\left(\frac{1+r_n}{1-r_n}\right) + \frac{u_{1-\alpha/2}}{\sqrt{n}} \cdot \frac{\xi_n}{1-r_n^2}\right) \tag{2.4}$$

je asymptotický intervalový odhad funkce $\frac{1}{2}\log(\frac{1+\rho}{1-\rho})$ o spolehlivosti $1-\alpha$.

Důkaz. Označíme-li $F\rho = \frac{1}{2}\log(\frac{1+\rho}{1-\rho})$, pak z Poznámky 2.16 o porušení II víme, že

$$\sqrt{n}(Fr_n - F\rho) \xrightarrow{d} \mathcal{N}\left(0, \left(\frac{\sigma^*}{1-\rho^2}\right)^2\right) \text{ pro } n \to \infty.$$

Užitím Cramér-Sluckého Věty tedy dostávame

$$\sqrt{n}\frac{Fr_n - F\rho}{\frac{\xi_n}{1-r_n^2}} = \underbrace{\sqrt{n}\frac{Fr_n - F\rho}{\frac{\sigma^*}{1-\rho^2}}}_{\xrightarrow{d}\mathcal{N}(0,1)} \cdot \underbrace{\frac{\frac{\sigma^*}{1-\rho^2}}{\frac{\xi_n}{1-r_n^2}}}_{\xrightarrow{p}1} \xrightarrow{d} \mathcal{N}(0,1).$$

Výslední intervalový odhad vznikne analogickým postupem jako v důkazu Věty 2.19. □

Kapitola 3

Simulace

3.1 Hladiny testů

3.1. Úvod. V předchozí části jsme vyšetřovali asymptotické chování výběrového korelačního koeficientu, tj. zjišťovali jsme, jaké má asymptotické rozdělení, jestliže jsme měli náhodný výběr z nespecifikovaného rozdělení. Pak jsme se podívali na výsledné rozdělení, měli jsme-li náhodný výběr vektorů s nezávislými složkami a výsledky byly shrnuty ve Větě 2.15 na str. 12 a v Poznámce 2.16 na str. 13. Dospěli jsme k závěru, že za předpokladu nezávislosti složek náhodných vektorů, test z Věty 1.8 na str. 3 zachová předepsanou hladinu. Na druhé straně neznáme skutečnou hladinu testu. Obecně ji bohužel nelze spočíst, můžeme jí ale odhadnout pomocí simulací, což je tématem této sekce.

3.2. Hypotéza. Buď $(X_1,Y_1),\ldots,(X_n,Y_n)$ náhodný výběr z rozdělení s distribuční funkcí F. Položíme

$$H_0 : X_1,Y_1 \text{ jsou nezávislé vs. } H_1 : \text{neplatí } H_0.$$

Označme dále Q statistiku z Věty 1.8 na str. 3 a U statistiku z Důsledku 2.8 na str. 10, tj.

$$Q = \frac{r_n}{\sqrt{1-r_n^2}}\sqrt{n-2},$$
$$U = \sqrt{n}r_n.$$

Vytvoříme 3 různé testy. Test $\mathcal{Q}_1$ bude testovat hypotézu pomocí statistiky Q a kritický obor nalezne pomocí Studentova t-rozdělení o $n-2$ stupněch volnosti, tj. $Q \overset{H_0}{\sim} t_{n-2}$. Druhý test $\mathcal{Q}_2$ bude testovat asymptoticky pomocí statistiky Q, ale kritický obor nalezne pomocí normálního rozdělení, tj. $Q \overset{H_0,d}{\to} \mathcal{N}(0,1)$. Konečně posledný test $\mathcal{U}$ bude testovat opět asymptoticky pomocí statistiky U a kritický obor nalezne přes normální rozdělení, tj. $U \overset{H_0,d}{\to} \mathcal{N}(0,1)$. Shrnutí všech tří testů lze nalézt v Tabulce 3.1.

Test	Testovací statistika	Rozdělení
$\mathcal{Q}_1$	$Q = \frac{r_n}{\sqrt{1-r_n^2}}\sqrt{n-2}$	t_{n-2}
$\mathcal{Q}_2$	$Q = \frac{r_n}{\sqrt{1-r_n^2}}\sqrt{n-2}$	$\mathcal{N}(0,1)$
$\mathcal{U}$	$U = \sqrt{n}r_n$	$\mathcal{N}(0,1)$

Tabulka 3.1: Shrnutí testů

3.3. Poznámka. Test $\mathcal{Q}_1$ je běžně užívaný test v praxi, test $\mathcal{Q}_2$ je asymptotická verze testu $\mathcal{Q}1$. Poslední test se dá lehce odvodit a je taky užíván statistickým software, proto lze předpokládat, že se často objevuje v praxi. Z těchto důvodů nás bude u těchto testů zajímat skutečná hladina.

3.4. Definice (hladina testu). Buď $X_1,\ldots,X_n$ náhodný výběr z rozdělení s distribuční funkcí F závislou na $\boldsymbol{\theta} \in \mathbb{R}^k$ a buď $G = G(X_1,\ldots,X_n)$ statistika s kritickým oborem W. Nechť nulová hypotéza je tvaru $H_0 : \boldsymbol{\theta} \in \Theta_0$ a alternativní $H_1 : \boldsymbol{\theta} \in \Theta_1$ pak hodnotu

$$\alpha = \sup_{\boldsymbol{\theta}\in\Theta_0} \mathrm{P}(G \in W),$$

nazýváme **hladina testu**.

3.5. Průběh simulace. Simulace budou mít několik kroků:

1. Nejprve si předem zvolíme dané $\alpha \in (0,1)$ což bude odrážet předpokládanou hladinu testu. Běžná hladina je 5%, proto zvolíme $\alpha = 5\%$.
2. Zvolíme distribuční funkci F, resp. zvolíme si rozdělení, z jakého bude pocházet náhodný výběr.
3. Nasimulujeme náhodný výběr 10 000-krát z rozdělení F o rozsahu n, kde $n \in \{5,10, 15,20,30,50\}$. U asymptotických testů nás hlavně budou zajímat malé rozsahy výběrů, abychom viděli, jak moc je narušená hladina. Taky nás může zajímat, zda pro velké rozsahy výběrů je odhadnutá hladina dostatečně blízká předpokládané.
4. Odhadneme skutečnou hladinu jednotlivých testů pomocí vztahů:

$$\mathcal{Q}_1 : \hat{\alpha} = \frac{1}{10\,000}\sum_{i=1}^{10\,000} \mathbf{1}\{|Q_i| \geq t_{n-2}(1-\alpha/2)\},$$

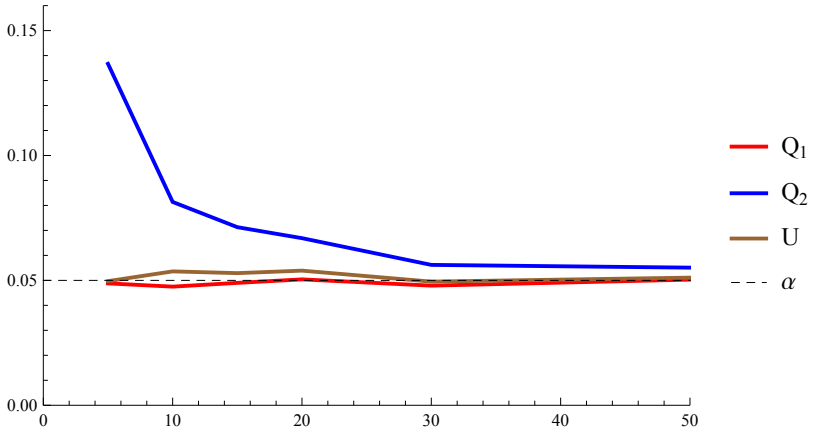
$$\mathcal{Q}_2 : \hat{\alpha} = \frac{1}{10\,000}\sum_{i=1}^{10\,000} \mathbf{1}\{|Q_i| \geq u_{1-\alpha/2}\},$$

$$\mathcal{U} : \hat{\alpha} = \frac{1}{10\,000}\sum_{i=1}^{10\,000} \mathbf{1}\{|U_i| \geq u_{1-\alpha/2}\},$$

kde Q_i je i-tá realizace testové statistiky Q, U_i je i-tá realizace testové statistiky U a $\mathbf{1}$ značí indikátorovou funkci.

$\mathcal{N}$	$\mathcal{Q}_1 : Q \sim t_{n-2}$	$\mathcal{Q}_2 : Q \sim \mathcal{N}(0,1)$	$\mathcal{U} : U \sim \mathcal{N}(0,1)$
$n = 5$	0.0488	0.1367	0.0497
$n = 10$	0.0475	0.0814	0.0536
$n = 15$	0.0490	0.0713	0.0529
$n = 20$	0.0504	0.0669	0.0539
$n = 30$	0.0479	0.0562	0.0495
$n = 50$	0.0503	0.0551	0.0511

Tabulka 3.2: Simulace z normální rozdělení



Obrázek 3.1: Simulace z normálního rozdělení

Simulace provedeme v statistickém software R Core Team (2013) a na grafické znázornění výsledků použijeme Wolfram Research (2012).

3.6. Simulace z normálního rozdělení. Jako první provedeme simulace z normálního rozdělení, tj. budeme vytvářet náhodné výběry z rozdělení $\mathcal{N}_2(\boldsymbol{\mu},\mathbf{V})$, kde $\boldsymbol{\mu} = (0,0)'$ a $\mathbf{V}$ je jednotková matice. Výsledky jsou sepsány v Tabulce 3.2 a graficky znázorněny na Obr. 3.1.

V prvním sloupci tabulky je rozsah výběru a zbylé sloupce reprezentují jednotlivé testy $\mathcal{Q}_1$, $\mathcal{Q}_2$ a $\mathcal{U}$. Simulace pro test $\mathcal{Q}_1$ splňuje všechny předpoklady Věty 1.8 na str. 3 a tudíž výsledný odhad by měl odpovídat předpokládané hladině testu, tj. $\alpha = 5\%$. Je vidět, že tomu tak skutečně je, neboť jak malé, tak i vysoké rozsahy výběru mají odhadnutou hladinu testu blízkou hodnotě α. Test $\mathcal{Q}_2$ je zajímavější. Ten užívá asymptotickou normalitu testové statistiky Q a kritický obor nalezne pomocí kvantilu normálního rozdělení. Je vidět, že pro malé rozsahy výběrů ($n \in \{5,10,15\}$) je hladina vyšší než u simulací s vyššími rozsahy. To znamená, že maximální pravděpodobnost, že testovací statistika padne do kritického oboru za platnosti nulové hypotézy, je při malém rozsahu značně vyšší než předpokládaná i když jde o náhodný výběr z dvojrozměrného normálního rozdělení. Poučení z toho je, že test $\mathcal{Q}_2$ je vhodný jenom pro výběry vyšších rozsahů. Poslední test $\mathcal{U}$ je taky asymptotický,

t_5	$\mathcal{Q}_1 : Q \sim t_{n-2}$	$\mathcal{Q}_2 : Q \sim \mathcal{N}(0,1)$	$\mathcal{U} : U \sim \mathcal{N}(0,1)$
$n = 5$	0.0490	0.1492	0.0500
$n = 10$	0.0503	0.0869	0.0576
$n = 15$	0.0537	0.0738	0.0582
$n = 20$	0.0538	0.0696	0.0570
$n = 30$	0.0538	0.0648	0.0563
$n = 50$	0.0556	0.0596	0.0564

Tabulka 3.3: Simulace ze Studentovo t-rozdělení o 5 stupních volnosti

t_{10}	$\mathcal{Q}_1 : Q \sim t_{n-2}$	$\mathcal{Q}_2 : Q \sim \mathcal{N}(0,1)$	$\mathcal{U} : U \sim \mathcal{N}(0,1)$
$n = 5$	0.0509	0.1460	0.0516
$n = 10$	0.0452	0.0831	0.0511
$n = 15$	0.0541	0.0746	0.0591
$n = 20$	0.0510	0.0661	0.0543
$n = 30$	0.0520	0.0611	0.0538
$n = 50$	0.0481	0.0552	0.0498

Tabulka 3.4: Simulace ze Studentovo t-rozdělení o 10 stupních volnosti

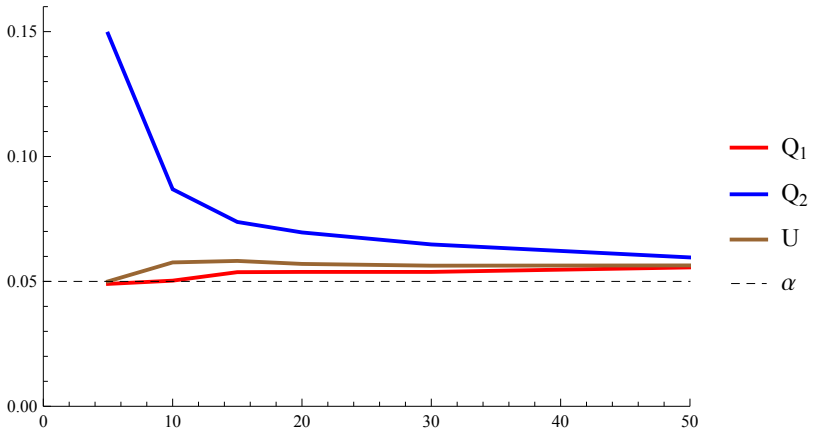
ale odhadnutá hladina se chová přibližně stejně jako u testu $\mathcal{Q}_1$. Dá se tedy říct, že testy $\mathcal{Q}_1$ a $\mathcal{U}$ mají v případě normality přibližně stejné hladiny.

3.7. Simulace ze Studentova t-rozdělení. Stejným způsobem provedeme simulace ze Studentova t-rozdělení s m stupni volnosti pro $m \in \{5{,}10\}$, tj. budeme simulovat náhodné výběry $(X_1{,}Y_1)', \ldots, (X_n{,}Y_n)'$ kde X_1 a Y_1 jsou nezávislé a obojí mají Studentovo t-rozdělení o m stupních volnosti. Pro vyšší stupně volnosti již lze Studentovo rozdělení aproximovat normálním, čili dá se očekávat, že i výsledné odhady skutečné hladiny budou přibližně stejné. Jednotlivé výsledky lze nalézt v Tabulkách 3.3 a 3.4.

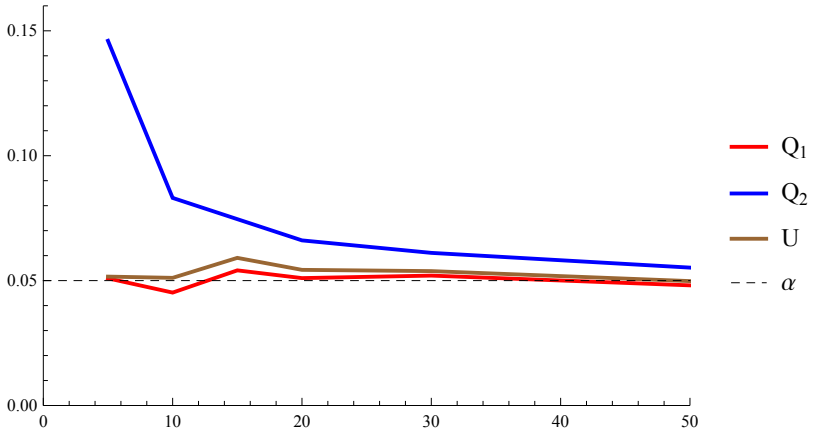
Je vidět, že výsledky jsou velmi podobné simulacím z normálního rozdělení a to dokonce i pro malé stupně volnosti. Grafické znázornění odhadnutých hladin lze nalézt na Obr. 3.2 a Obr. 3.3.

3.9. Simulace z gamma rozdělení. Budeme simulovat gamma rozdělení s parametry $k = 7$ a $a = 599$, tj. $(X_1{,}Y_1)', \ldots, (X_n{,}Y_n)'$ kde X_1 a Y_1 jsou nezávislé a stejně rozdělené s gamma rozdělením $\Gamma(7{,}599)$. Odhadneme skutečnou hladinu jednotlivých testů $\mathcal{Q}_1{,}\mathcal{Q}_2$ a $\mathcal{U}$. Výsledky jsou v Tabulce 3.5 a graficky v Obr. 3.4.

Je vidět, že odhady jsou téměř totožné s předchozími simulacemi. Testy $\mathcal{Q}_1$ a $\mathcal{Q}_2$ mají



Obrázek 3.2: Simulace ze Stundetova t rozdělení o 5 stupních volnosti



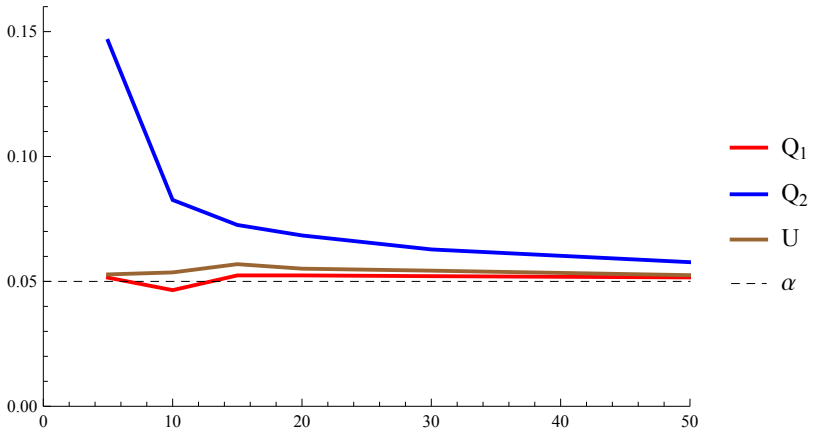
Obrázek 3.3: Simulace ze Stundetova t rozdělení o 10 stupních volnosti

$\Gamma(k,a)$	$\mathcal{Q}_1 : Q \sim t_{n-2}$	$\mathcal{Q}_2 : Q \sim \mathcal{N}(0,1)$	$\mathcal{U} : U \sim \mathcal{N}(0,1)$
$n = 5$	0.0515	0.1463	0.0528
$n = 10$	0.0465	0.0826	0.0536
$n = 15$	0.0524	0.0726	0.0569
$n = 20$	0.0524	0.0684	0.0551
$n = 30$	0.0521	0.0628	0.0543
$n = 50$	0.0516	0.0577	0.0525

Tabulka 3.5: Simulace z Gamma rozdělení

dokonce i při malém rozsahu výběru odhadnutou hladinu téměř totožnou s předvolenou hladinou α.

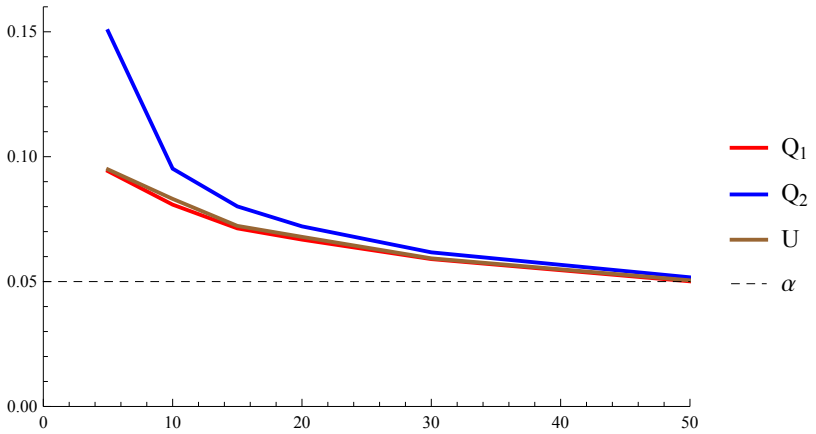
3.10. Simulace z lognormálního rozdělení. Budeme simulovat náhodný výběr $(X_1,Y_1)'$, $\ldots,(X_n,Y_n)'$ z dvojrozměrného normálního rozdělení s parametry $\mu_1 = \mu_2 = 10$, $\sigma_1^2 =$



Obrázek 3.4: Simulace z gamma rozdělení

$\mathcal{LN}$	$\mathcal{Q}_1 : Q \sim t_{n-2}$	$\mathcal{Q}_2 : Q \sim \mathcal{N}(0,1)$	$\mathcal{U} : U \sim \mathcal{N}(0,1)$
$n = 5$	0.0942	0.1503	0.0949
$n = 10$	0.0808	0.0952	0.0831
$n = 15$	0.0713	0.0801	0.0723
$n = 20$	0.0668	0.0721	0.0679
$n = 30$	0.0590	0.0617	0.0593
$n = 50$	0.0501	0.0517	0.0505

Tabulka 3.6: Simulace z Lognormálního rozdělení

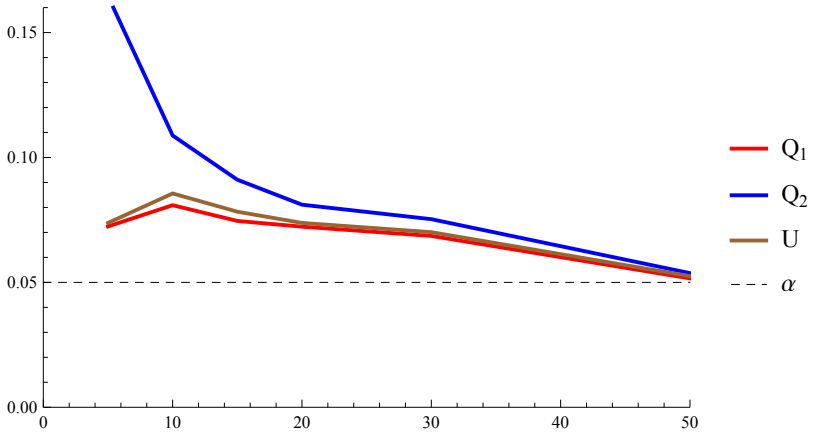


Obrázek 3.5: Simulace z lognormálního rozdělení

$\sigma_2^2 = 5$ a korelačním koeficientem $\rho = 0$. Následně pro transformovaný náhodný výběr $(e^{X_1},e^{Y_1})', \ldots, (e^{X_n},e^{Y_n})'$ budeme obdobným způsobem odhadovat skutečnou hladinu testu. Výsledky jsou v Tabulce 3.6 a graficky jsou znázorněny na Obr. 3.5.

t_1	$\mathcal{Q}_1 : Q \sim t_{n-2}$	$\mathcal{Q}_2 : Q \sim \mathcal{N}(0,1)$	$\mathcal{U} : U \sim \mathcal{N}(0,1)$
$n = 5$	0.0724	0.1646	0.0738
$n = 10$	0.0809	0.1088	0.0856
$n = 15$	0.0746	0.0911	0.0783
$n = 20$	0.0723	0.0811	0.0738
$n = 30$	0.0686	0.0753	0.0701
$n = 50$	0.0516	0.0537	0.0525

Tabulka 3.7: Simulace z Cauchyova rozdělení



Obrázek 3.6: Simulace z Cauchyova rozdělení

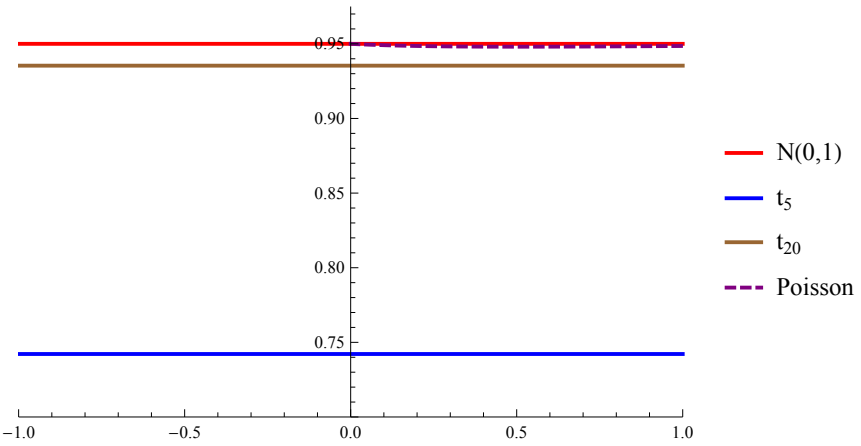
V lognormálním rozdělení je odhad skutečné hladiny všech testů pro malé rozsahy výběrů přibližně dvojnásobný předpokládané hladině. To znamená, že testy nemusí být asymptoticky přesné pro malé rozsahy ($n = 5, \ldots, 15$). Na druhé straně pro vyšší rozsahy výběrů je odhadnutá hladina téměř totožná s hladinou α.

3.11. Simulace z Cauchyova rozdělení. Budeme simulovat náhodný výběr $(X_1, Y_1)', \ldots, (X_n, Y_n)'$ kde X_1 a Y_1 jsou nezávislé a stejně rozdělené z Cauchyova rozdělení, tj. ze spojitého rozdělení s hustotou

$$f(x) = \frac{1}{\pi(1 + x^2)}, \; x \in \mathbb{R}.$$

Výsledky simulací lze nalézt v Tabulce 3.7 a graficky jsou znázorněné na Obr. 3.6.

Cauchyovo rozdělení je specifické tím, že nemá konečný ani první moment. Tím pádem nemůže mít ani konečný rozptyl a korelační koeficient pro dvojici náhodných veličin z Cauchyova rozdělení není definován. Všechny statistické testy $\mathcal{Q}_1$,$\mathcal{Q}_2$ a $\mathcal{U}$ byly odvozeny za předpokladu konečnosti čtvrtých momentů a tedy skutečná hladina testu může být libovolná. To nám ale nebrání v provedení simulací. Překvapivě, odhadnuté hladiny jednotlivých testů jsou pro vysoké rozsahy výběru $n = 50$ téměř shodné s hladinou $\alpha = 5\%$.



Obrázek 3.7: Skutečné pokrytí se spolehlivostí $1-\alpha$ pro $\alpha = 5\%$

3.2 Intervaly spolehlivosti

3.12. Poznámka. Podíváme se na Fisherovu Z-transformaci a její aplikace v intervalových odhadech. Z Poznámky 2.16 na str. 13 víme, že

$$\sqrt{n}(Fr_n - F\rho) \xrightarrow{d} \mathcal{N}\left(0, \left(\frac{\sigma^*(\rho)}{1-\rho^2}\right)^2\right) \text{ pro } n \to \infty.$$

Máme-li náhodný výběr z dvojrozměrného normálního rozdělení, pak $\sigma^*(\rho) = 1-\rho^2$, čili asymptotický interval spolehlivosti se spolehlivostí $1-\alpha$ má tvar $\left(Fr_n \mp \frac{u_{1-\alpha/2}}{\sqrt{n}}\right)$, kde $u_{1-\alpha/2} = \Phi^{-1}(1-\alpha/2)$ kde Φ je distribuční funkce normálního rozdělení $\mathcal{N}(0,1)$. Teď zjistíme pokrytí tohto intervalu pro obecné dvojrozměrné rozdělení. Platí:

$$\mathrm{P}\left(Fr_n - \frac{u_{1-\alpha/2}}{\sqrt{n}} \le F\rho \le Fr_n + \frac{u_{1-\alpha/2}}{\sqrt{n}}\right) = \mathrm{P}\left(-u_{1-\alpha/2} \le \sqrt{n}(Fr_n - F\rho) \le u_{1-\alpha/2}\right) =$$

$$\mathrm{P}\left(-\frac{u_{1-\alpha/2}}{\sigma^*(\rho)/(1-\rho^2)} \le \frac{\sqrt{n}(Fr_n - F\rho)}{\sigma^*(\rho)/(1-\rho^2)} \le \frac{u_{1-\alpha/2}}{\sigma^*(\rho)/(1-\rho^2)}\right) \overset{n\to\infty}{\to} \Phi\left(\frac{u_{1-\alpha/2}}{\sigma^*(\rho)/(1-\rho^2)}\right) -$$

$$-\Phi\left(-\frac{u_{1-\alpha/2}}{\sigma^*(\rho)/(1-\rho^2)}\right) = 2\Phi\left(\frac{u_{1-\alpha/2}}{\sigma^*(\rho)/(1-\rho^2)}\right) - 1,$$

Na obrázku 3.7 jsou znázorněny grafy funkcí pokrytí pro rozdělení $\mathcal{N}_2(\mathbf{0},\mathbf{V}(\rho))$,$t_2(\mathbf{0},\mathbf{V}(\rho),5)$, $t_2(\mathbf{0},\mathbf{V}(\rho),20)$ a dvojrozměrné Poissonovo rozdělení s parametry $\lambda_1 = \lambda_2 = 10$ a $\lambda(\rho)$ v závislosti na korelačním koeficientu ρ. Je vidět, že skutečné pokrytí pro dvojrozměrné Studentovo t-rozdělení s rostoucím stupněm volnosti se blíží k pokrytí normálního rozdělení. V případě Poissonova rozdělení platí $\rho(\lambda) = \dfrac{\lambda}{\lambda+10}$ pro $\lambda > 0$ a tudíž $\rho > 0$. Je ale vidět, že nejvyšší spolehlivost je pro ρ blízké nule. Na druhé straně nejnižší spolehlivost nabývá

pro $\rho = \frac{1}{2}$ s hodnotou 94.81%.

Na závěr ještě provedeme simulace, ve kterých budeme odhadovat spolehlivost intervalových odhadů a jejich délku. Budeme porovnávat 3 různé intervalové odhady korelačního koeficientu:

1. Interval spolehlivosti vytvořen pomocí Fisherovy Z-transformace z Věty 1.10 na str. 4. Intervalový odhad vytvořen touto metodou označíme pro účely následujících simulací symbolem IO_F.
2. Interval spolehlivosti vytvořen pomocí konzistentního odhadu asymptotického rozptylu výběrového korelačního koeficientu z Věty 2.19 na str. 15 tvaru (2.3). Intervalový odhad vytvořen touto metodou označíme symbolem IO_p.
3. Interval spolehlivosti vytvořen pomocí kombinace předchozích dvou způsobů z Věty 2.20 na str. 15 tvaru (2.4). Intervalový odhad vytvořen touto metodou označíme symbolem IO_k.

3.13. Průběh simulace.

1. Zvolíme hodnotu $\alpha \in (0,1)$, kde $1-\alpha$ je předpokládaná spolehlivost intervalového odhadu. V naších simulacích budeme všude volit $\alpha = 5\%$.
2. Zvolíme dvojrozměrné rozdělení s distribuční funkci F, ze kterého budeme simulovat náhodné výběry a určíme korelační koeficient ρ.
3. Budeme simulovat 10 000 náhodných výběrů z rozdělení s distribuční funkcí F a s rozsahem $n = 5,10,15,20,30,50$. Pro každý náhodný výběr najdeme intervalové odhady korelačního koeficientu pomocí všech tří výše zmíněných metod. Výsledkem tedy budou intervaly (D_i,U_i) pro $i \in \{1,\ldots,10000\}$, kde D_i je dolní mez a U_i je horní mez pro každou metodu.
4. Odhadneme spolehlivost intervalového odhadu $\widehat{1-\alpha}$:

$$\widehat{1-\alpha} = \frac{1}{10\,000} \sum_{i=1}^{10\,000} \mathbf{1}\{D_i \leq \rho \leq U_i\},$$

a jeho průměrnou délku $\hat{d}$:

$$\hat{d} = \frac{1}{10\,000} \sum_{i=1}^{n} |U_i - D_i|,$$

pro každou ze tří metod.

Rozsah	IO_F		IO_p		IO_k	
n	$\hat{d}$	$\widehat{1-\alpha}$	$\hat{d}$	$\widehat{1-\alpha}$	$\hat{d}$	$\widehat{1-\alpha}$
5	0.8001	0.9576	0.2649	0.6200	0.2929	0.6714
10	0.3592	0.9551	0.2099	0.7784	0.2337	0.8221
15	0.2525	0.9507	0.1766	0.8311	0.1930	0.8700
20	0.2043	0.9487	0.1561	0.8565	0.1678	0.8897
30	0.1549	0.9522	0.1295	0.8840	0.1365	0.9111
50	0.1137	0.9509	0.1020	0.9110	0.1055	0.9259

Tabulka 3.8: Simulace z normálního rozdělení s $\rho = 0.9$

Rozsah	IO_F		IO_p		IO_k	
n	$\hat{d}$	$\widehat{1-\alpha}$	$\hat{d}$	$\widehat{1-\alpha}$	$\hat{d}$	$\widehat{1-\alpha}$
5	1.4047	0.9626	0.6941	0.5858	0.6750	0.6713
10	0.9157	0.9601	0.6556	0.7453	0.6166	0.8197
15	0.7484	0.9592	0.5655	0.8112	0.5584	0.8637
20	0.6626	0.9673	0.5317	0.8663	0.5250	0.9001
30	0.5765	0.9648	0.5017	0.9194	0.4944	0.9168
50	0.4290	0.9734	0.4067	0.9522	0.4034	0.9552

Tabulka 3.9: Simulace z Poissonova rozdělení s $\rho = 0.5$

3.14. Simulace z normálního rozdělení. Nejprve budeme simulovat náhodné výběry z dvojrozměrného normálního rozdělení s parametry $\boldsymbol{\mu} = (0,0)'$ a $\mathbf{V} = \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}$, kde budeme uvažovat korelační koeficient $\rho = 0.9$. Výsledky jsou v Tabulce 3.8.

Z Tabulky 3.8 je vidět, že intervalový odhad vytvořen metodou Fisherovy Z-transformace udržuje spolehlivost pro jakýkoliv rozsah výběru. Na druhé straně tím zaplatí průměrnou délkou, která je pro malé rozsahy výběru skoro jednotková. Na Obr. 3.7 vidíme, že asymptotická spolehlivost je přesně 0.95, což také potrvzují simulace. Opačným případem je intervalový odhad IO_p, který má průměrní délku menší než IO_F, ale jeho spolehlivost je pro malé rozsahy výběrů třetinově menší než předpokládaná. Intervalový odhad IO_k poskytuje odhady s vyšší spolehlivostí než IO_p, s vyšší průměrní délkou, ale pořád je spolehlivost značně menší než u odhadu IO_F. Pro rozsah $n = 50$ se všechny tři odhady liší už jen minimálně.

3.15. Simulace z Poissonova rozdělení. Dle Obr. 3.7 jsme zjistili, že nejnižší asymptotické pokrytí v Poissonově rozdělení je v případě, když je korelační koeficient roven $\frac{1}{2}$, tj. pokud $\lambda = \lambda_1 = \lambda_2 = 10$. Provedeme simulace pro Poissonovo rozdělení s parametry $\lambda = \lambda_1 = \lambda_2 = 10$ a zjistíme odhad spolehlivosti. Výsledky jsou v Tabulce 3.9.

Rozsah	IO_F		IO_p		IO_k	
n	$\hat{d}$	$\widehat{1-\alpha}$	$\hat{d}$	$\widehat{1-\alpha}$	$\hat{d}$	$\widehat{1-\alpha}$
5	1.5494	0.9319	0.8885	0.5175	0.8342	0.5977
10	1.1296	0.9058	0.9194	0.6938	0.8574	0.7562
15	0.9406	0.8874	0.8638	0.7598	0.8115	0.8153
20	0.8251	0.8741	0.8147	0.8006	0.7700	0.8445
30	0.6839	0.8649	0.7290	0.8351	0.6959	0.8696
50	0.5377	0.8571	0.6242	0.8800	0.6027	0.9006

Tabulka 3.10: Simulace ze Studentova t-rozdělení s $\rho = 0$

Ve všech simulacích odhad IO_F zachováva spolehlivost i pro malé rozsahy výběrů. Problém je vysoká průměrná délka, která je v poslední simulaci dokonce až s hodnotou 1.4047 pro $n = 5$. Na druhé straně intervalový odhad IO_p dělá přesný opak, tj. průměrní délku a spolehlivost má nejnižší ze všech intervalových odhadů. Intervalový odhad IO_k poskytuje kompromis mezi IO_F a IO_p, tj. poskytne odhad s vyšší spolehlivostí a vyšší průměrní délkou než IO_p. Pro vyšší rozsahy výběrů se všechny tři intervalové odhady začínají shodovat ve všech ukazatelích.

3.16. Simulace ze Studentova t-rozdělení. Budeme simulovat náhodné výběry ze Studentova t-rozdělení $t_2(0,I,5)$. Korelační koeficient má hodnotu $\rho = 0$ a výsledky jsou v Tabulce 3.10.

Z Obr. 3.7 víme, že asymptotická spolehlivost pro výběry z $t_2(0,\mathbf{V}(\rho),5)$ je konstantní s přibližnou hodnotou 0.74, proto v simulacích odhadnutá spolehlivost klesá s rostoucím rozsahem výběru. Na druhé straně odhadnuté spolehlivosti intervalových odhadů IO_p a IO_k rostou k předpokládané spolehlivosti $1-\alpha$ a tedy IO_p a IO_k poskytují s rostoucím rozsahem výběru intervalový odhad s vyšší spolehlivostí, než IO_F.

Závěr

V této práci jsme se zabývali výběrovým korelačním koeficientem a jeho asymptotickými vlastnostmi. Uvedli jsme základní vlastnosti korelačního koeficientu a běžně užívaný test nezávislosti společně s intervalovým odhadem. Pak jsme odvodili asymptotické rozdělení výběrového korelačního koeficientu bez předpokladu normality a ukázali jsme, že běžně užívaný test nezávislosti lze užít i bez předpokladu normality. Následně jsme uvažovali 3 různé testy nezávislosti založené na korelačním koeficientu a 3 různé intervalové odhady na základě asymptotického rozdělení výběrového korelačního koeficientu.

V další části jsme prováděli simulace pro jednotlivé testy nezávislosti a intervalové odhady pro specifická rozdělení. Zjistili jsme, že běžně užívaný intervalový odhad vytvořen Fisherovou Z-transformaci zachovává předepsanou spolehlivost pro malé rozsahy výběrů. Pro větší rozsahy výběrů může mít intervalový odhad vytvořen Fisherovou Z-transformací asymptotickou spolehlivost radikálně menší než předvolenou spolehlivost (např. v případě dvojrozměrného Studentova t-rozdělení) a proto je spolehlivější test, který je vytvořen kombinací Fisherovy Z-transformace a konzistentního odhadu asymptotického rozptylu výběrového korelačního koeficientu.

Pomocí simulací v poslední částí této práce se nám podařilo odhadnout skutečnou hladinu jednotlivých testů. Na druhé straně nevíme, jakou mají jednotlivé testy sílu. Bohužel, na zodpovězení této otázky již není v této práci místo. V dalším výzkumu bychom se proto mohli zaměřit na odhad síly a rozsahu výběru u testů nezávislosti pro náhodné výběry z běžně užívaných rozdělení. Výsledky by nám mohly poskytnout lepší nadhled nad jednotlivými testy a společně s výsledky z této práce by tvořily postačující množství údajů o jednotlivých testech pro náhodné výběry ze specifických rozdělení.

Literatura

Anděl, J. (2007). *Základy matematické statistiky*. Druhé opravené vydání. Matfyzpress, Praha. ISBN 80-7378-001-1.

Dupač, V. a Hušková, M. (2005). *Pravděpodobnost a matematická statistika*. Karolinum, Praha. ISBN 978-80-246-0009-3.

Kocherlakota, S. a Kocherlakota, A. (2006). *Bivariate Discrete Distributions*. Encyclopedia of Statistical Sciences. John Wiley and Sons.

Kotz, S. a Nadarajah, S. (2004). *Multivariate t Distributions and Their Applications*. Cambridge University Press, USA. ISBN 0-521-82654-3.

Lehmann, E. L. (1999). *Elements of Large-Sample Theory*. Second Edition. Springer-Verlag, New York. ISBN 0-387-98595-6.

R Core Team (2013). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. URL `http://www.R-project.org/`.

Van der Vaart, A. (1998). *Asymptotic statistics*. Paperback edition. Cambridge University Press, UK. ISBN 0-521-78450-6.

Wolfram Research, I. (2012). *Mathematica*. Champaign, Illinois.

Seznam obrázků

3.1 Simulace z normálního rozdělení 19
3.2 Simulace ze Stundetova t rozdělení o 5 stupních volnosti 21
3.3 Simulace ze Stundetova t rozdělení o 10 stupních volnosti 21
3.4 Simulace z gamma rozdělení 22
3.5 Simulace z lognormálního rozdělení 22
3.6 Simulace z Cauchyova rozdělení 23
3.7 Skutečné pokrytí se spolehlivostí $1 - \alpha$ pro $\alpha = 5\%$ 24

Seznam tabulek

3.1 Shrnutí testů 18
3.2 Simulace z normální rozdělení 19
3.3 Simulace ze Studentovo t-rozdělení o 5 stupních volnosti 20
3.4 Simulace ze Studentovo t-rozdělení o 10 stupních volnosti 20
3.5 Simulace z Gamma rozdělení 21
3.6 Simulace z Lognormálního rozdělení 22
3.7 Simulace z Cauchyova rozdělení 23
3.8 Simulace z normálního rozdělení s $\rho = 0.9$ 26
3.9 Simulace z Poissonova rozdělení s $\rho = 0.5$ 26
3.10 Simulace ze Studentova t-rozdělení s $\rho = 0$ 27