VYSOKÉ UČENÍ TECHNICKÉ V BRNĚ
BRNO UNIVERSITY OF TECHNOLOGY

FAKULTA STROJNÍHO INŽENÝRSTVÍ
FACULTY OF MECHANICAL ENGINEERING

ÚSTAV MATEMATIKY
INSTITUTE OF MATHEMATICS

POROVNÁNÍ TESTŮ NULOVOSTI KORELAČNÍHO KOEFICIENTU DVOU NORMÁLNÍCH NÁHODNÝCH VELIČIN

COMPARISON OF TESTS ON NULL CORRELATION BETWEEN TWO NORMAL RANDOM VARIABLES

BAKALÁŘSKÁ PRÁCE
BACHELOR'S THESIS

AUTOR PRÁCE
AUTHOR

Vít Kalenský

VEDOUCÍ PRÁCE
SUPERVISOR

doc. Mgr. Zuzana Hübnerová, Ph.D.

BRNO 2016

VYSOKÉ UČENÍ TECHNICKÉ V BRNĚ
FAKULTA STROJNÍHO INŽENÝRSTVÍ

Zadání bakalářské práce

Ústav:	Ústav matematiky
Student:	**Vít Kalenský**
Studijní program:	Aplikované vědy v inženýrství
Studijní obor:	Matematické inženýrství
Vedoucí práce:	**doc. Mgr. Zuzana Hübnerová, Ph.D.**
Akademický rok:	2015/16

Ředitel ústavu Vám v souladu se zákonem č.111/1998 o vysokých školách a se Studijním a zkušebním řádem VUT v Brně určuje následující téma bakalářské práce:

Porovnání testů nulovosti korelačního koeficientu dvou normálních náhodných veličin

Stručná charakteristika problematiky úkolu:

V základním kurzu statistiky se studenti setkávají s několika přístupy k testováni nulovosti korelačního koeficientu dvou normálně rozdělených náhodných veličin.

Cíle bakalářské práce:

1. Zavedení potřebných pojmů
2. Odvození testů
3. Porovnání testů

Seznam literatury:

Anděl, J. Základy matematické statistiky. Praha: MatfyzPress, 2011

Termín odevzdání bakalářské práce je stanoven časovým plánem akademického roku 2015/16

V Brně, dne

L. S.

prof. RNDr. Josef Šlapal, CSc. ředitel ústavu	doc. Ing. Jaroslav Katolický, Ph.D. děkan fakulty

Abstrakt
Práce se zabývá srovnáním testů nulovosti korelačního koeficientu dvou normálních náhodných veličin pomocí T statistiky, Fisherovy transformace, Hotellingovy transformace a Haddad-Provostovy statistiky. Obsahuje odvození testů a jejich silofunkcí, které následně porovnává vzájemně mezi sebou a s hodnotami získanými z nasimulovaných náhodných výběrů v programu MATLAB. Na závěr práce jsou vysvětleny a popsány výsledky. Nechybí zde také teoretický úvod k dané problematice.

Summary
This thesis deals with the comparison of the tests of the correlation coefficient of two normal random variables using T statistics, Fisher's transformation, Hotelling transformation and Haddad-Provost's statistics. It contains the derivation of the tests and their power function, which then compares each other and with the values obtained from the simulated random samples in program MATLAB. Finally the thesis depicts and describe the results. There is also the theoretical introduction to the topic.

Klíčová slova
korelační koeficient, Fisherova transformace,Hotellingova transformace, silofunkce

Keywords
correlation cofficient, Fisher's transformation,Hotelling's transformation, power function

KALENSKÝ, V.*Porovnání testů nulovosti korelačního koeficientu dvou normálních náhodných veličin*. Brno: Vysoké učení technické v Brně, Fakulta strojního inženýrství, 2016. 37 s. Vedoucí bakalářské práce doc. Mgr. Zuzana Hübnerová, Ph.D.

Prohlašuji, že svou bakalářskou práci na téma „Porovnání testů nulovosti korelačního koeficientu dvou normálních náhodných veličin“ jsem zpracoval samostatně pod vedením vedoucího bakalářské práce a že jsem uvedl všechny použité prameny a literaturu, ze kterých jsem čerpal.

Vít Kalenský

Děkuji vedoucí mé bakalářské práce doc. Mgr. Zuzaně Hübnerové, Ph.D. za její rady, ochotný přístup a velkou trpělivost při vedení mé bakalářské práce.

Vít Kalenský

Obsah

1. Úvod

Korelační koeficient vyjadřuje lineární závislost dvou veličin. Jeho nulovost respektive nulovost lineární složky regresního modelu můžeme testovat pomocí T statistiky tak, že zkoumáme, zda T patří do nějakého kritického oboru Studentova rozdělení. Pro testování pravděpodobnosti i jiných hodnot korelačního koeficientu navrhl Sir Ronald Aylmer Fisher takzvanou Z-transformaci. Tato ale bohužel ztrácí funkční sílu pro menší hodnoty rozsahu náhodného výběru ze dvou normálních náhodných veličin, Harold Hotelling proto vytvořil úpravu Fisherovy transformace, která tento nežádaný efekt minimalizuje. Další metodu, díky které se dá testovat nulovost korelačního koeficientu, objevili John N. Haddad a Serge B. Provost, pracovně ji proto nazývejme Haddan-Provostova statistika F. Cílem mé práce je porovnání těchto čtyř testovacích metod.
Bakalářská práce se dělí na tři kapitoly.
V první kapitole uvádím definice některých základních pojmů, které je třeba pro zkoumání této problematiky znát, jako například číselné charakteristiky náhodných veličin, náhodných výběrů a základy teorie k testování hypotéz.
Druhá kapitola nabízí odvození jednotlivých testů a jejich silofunkcí.
Ve třetí kapitole se nachází praktická část práce, ve které uvádím srovnání silofunkcí testů s hodnotami relativního počtu zamítnutí nulové hypotézy. Data ke srovnání jsem získal aplikací testovacích procedur na nasimulované výběry. K simulaci jsem používal software MATLAB R2015a.

2. Zavedení potřebných pojmů

V následující kapitole budeme pracovat s pojmy uvedenými v [1].

2.1. Číselné charakteristiky dvou náhodných veličin

Definice 2.1.1. Buď náhodný vektor $\mathbf{X} = (X_1, ..., X_n)'$ se střední hodnotou $E\mathbf{X} = (EX_1, ..., EX_n)'$. Jestliže $EX_k^2 < \infty$ pro k=1,...,n, pak říkáme, že $\mathbf{X}$ má konečné druhé momenty a definujeme **kovarianci** $\mathrm{cov}(X_i, X_j)$ vztahem

$$\mathrm{cov}(X_i, X_j) = E(X_i - EX_i)(X_j - EX_j).$$

Snadnou úpravou pak dostaneme vzorec

$$\mathrm{cov}(X_i, X_j) = EX_iX_j - EX_iEX_j,$$

který bývá vhodnější pro praktické výpočty.

Definice 2.1.2. Nechť X a Y jsou náhodné veličiny s kladnými rozptyly a konečnými druhými momenty. Lineární závislost těchto dvou veličin na sobě můžeme interpretovat pomocí **korelačního koeficientu**

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

Někdy značíme ρ jako $\rho_{X,Y}$, abychom vyznačili, o které dvě veličiny se jedná. Je zřejmé že $\rho_{X,Y} = \rho_{Y,X}$.

Věta 2.1.1. *Korelační koeficient nabývá hodnot* $-1 \leq \rho \leq 1$. *Rovnost* $\rho_{X,Y} = 1$ *nastává právě tehdy, když* $Y = a + bX$ *s pravděpodobností 1, přičemž* $b > 0$. *Analogicky platí rovnost* $\rho_{X,Y} = -1$ *právě tehdy, když* $Y = a + bX$ *s pravděpodobností 1, přičemž* $b < 0$.

Důkaz. Důkaz plyne ze Schwarzovy nerovnosti

$$|E(X - EX)(Y - EY)| \leq \sqrt{E(X - EX)^2 E(Y - EY)^2}$$

dle našeho značení

$$|\mathrm{cov}(X, Y)| \leq \sqrt{D(X)D(Y)}$$

Platí li $X - EX = 0$ a nebo $Y - EY = 0$ skoro jistě, pak ve Schwarzově nerovnosti nastává rovnost, tento případ jsme ale vyloučili podmínkou $D(X) > 0$ a $D(Y) > 0$ z 2.1.2. Rovnost nastává také v případě, kdy skoro jistě platí $Y - EY = b(X - EX)$ pro $b \neq 0$.

2.2. Náhodný výběr a výběrové statistiky

Definice 2.2.1. Nechť (Ω, A, P_θ) je pravděpodobnostní prostor, $\mathbf{X}_1, ..., \mathbf{X}_n$ jsou nezávislé k-rozměrné náhodné vektory na (Ω, A, P_θ) s distribuční funkcí F_θ z třídy distribučních funkcí $\{F_\theta, \theta \in \Theta\}$. Pak $\mathbf{X}_1, ..., \mathbf{X}_n$ je **náhodný výběr** rozsahu n z k-rozměrného rozdělení s distribuční funkcí F_θ.

Definice 2.2.2. Položme

$$\begin{aligned} \bar{X} &= \frac{1}{n}\sum_{i=1}^{n} X_i, \\ M_2 &= \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2, \\ S^2 &= \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2 \end{aligned}$$

Veličina $\bar{X}$ se nazývá **výběrový průměr** a veličinu M_2 budeme nazývat **momentový rozptyl**. Hodnotě $\sqrt{M_2}$ budeme říkat **výběrová směrodatná odchylka**. Veličina S^2 je definována jen pro $n \geq 2$ a často se používá místo M_2. Mnozí autoři právě S^2 nazývají výběrový rozptyl. Dále zaveďme

$$\begin{aligned} M_3 &= \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^3, \\ M_4 &= \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^4, \\ A_3 &= \frac{M_3}{\sqrt{M_2^3}}, \\ A_4 &= \frac{M_4}{M_2^2}. \end{aligned}$$

Veličina A_3 se nazývá výběrová šikmost a veličina A_4 výběrová špičatost.

Definice 2.2.3. Mějme náhodný výběr

$$(X_1, Y_1)', ..., (X_n, Y_n)'$$

z nějakého dvojrozměrného rozdělení. Označme $\bar{X}$ a S_X^2 charakteristiky výběru $X_1, ..., X_n$ a podobně $\bar{Y}$ a S_Y^2 charakteristiky výběru $Y_1, ..., Y_n$.
Dále definujeme **výběrovou kovarianci**

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

Definice 2.2.4. Korelační koeficient r definujeme vzorcem

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

za podmínky, že $S_X^2 > 0$ a $S_Y^2 > 0$. Někdy se místo r píše $r_{X,Y}$.

2.3. Testování hypotéz

Definice 2.3.1. Nechť nějaký náhodný vektor $\mathbf{X} = (X_1, ..., X_n)'$ má rozdělení, které závisí na paramentu $\underline{\theta} = (\theta_1, ..., \theta_k)'$. Tento parametr patří do množiny Θ, která se nazývá **parametrický prostor**. Předpokládejme, že tento prostor má alespoň dva prvky.

Definice 2.3.2. Tvrzení, že θ patří do nějaké neprázdné vlastní podmnožiny ω parametrického prostoru Θ nazýváme **nulová hypotéza**. Stručně označujeme nulovou hypotézu jako $H_0 : \underline{\theta} \in \omega$.

Definice 2.3.3. $H_1 : \underline{\theta} \notin \omega$ se nazývá **alternativní hypotéza**.

Definice 2.3.4. Zvolíme vhodnou množinu $W \in \mathcal{B}_n$, které budeme říkat **kritický obor**. Dále mějme **testovou statistiku** $T = T(X_1, ..., X_n)$. Jestliže $T \in W$ hypotézu zamítáme. Doplněk W v $\mathbf{R}_n$ nazvěme V. Když $T \in V$ hypotézu nezamítáme.

Definice 2.3.5. Při testování hypotéz se můžeme dopouštět dvou různých chyb. **Chyba I. druhu** spočívá v zamítnutí hypotézy, která je ve skutečnosti platná. Pravděpodobnost chyby tohoto druhu nazýváme **hladina významnosti** a značíme

$$\alpha = P(T \in W|H_0).$$

Definice 2.3.6. Chyba II. druhu znamená nezamítnutí hypotézy, která ve skutečnosti není pravdivá. Pravděpodobnost chyby II. druhu značíme

$$\beta = P(T \notin W|H_1).$$

Poznámka. Znázornění chyb I. a II. druhu můžeme vidět v tabulce 2.1.

Tabulka 2.1: Chyby testování

	Rozhodnutí	
Skutečnost	H0 nezamítáme	H0 zamítáme
H0 platí	správně	Chyba 1. druhu
H0 neplatí	Chyba 2. druhu	správně

Definice 2.3.7. Číslo $\gamma(\theta_0) = 1-\beta, \theta_0 \in \Theta$ se nazývá **síla testu** a udává pravděpodobnost zamítnutí hypotézy v případě, že tato hypotéza ve skutečnosti neplatí, tedy

$$\gamma(\theta_0) = P(T \in W|\theta_0).$$

Definice 2.3.8. Funkce vyjadřující závislost síly testu při různých hodnotách parametru θ, dané hladině významnosti α a daném rozsahu výběru n se nazývá **silofunkce** a značí se $\gamma(\theta), \theta \in \Theta$.

3. Testy nulovosti korelačního koeficientu

Máme náhodný vektor s dvourozměrným normálním rozdělením $N_2(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, kde

$$\boldsymbol{\Sigma} = \begin{bmatrix} \sigma_X^2 & \rho\sigma_X\sigma_Y \\ \rho\sigma_X\sigma_Y & \sigma_Y^2 \end{bmatrix} \tag{3.0.1}$$

je jeho varianční matice.
Máme k dispozici náhodný výběr z rozdělení o rozsahu n, na jehož základě testujeme hypotézu $H_0 : \rho = 0$ oproti oboustranné alternativě $H_1 : \rho \in < -1, 1 > -\{0\}$.
Následně se zabýváme srovnáním více různých testů. Testovat nulovost korelačního koeficientu budeme čtyřmi různými testy založenými na:

1. T statistice viz [1], str. 94
2. Fisherově transformaci viz [1], str. 95
3. Hotellingově transformaci viz [1], str. 97
4. Haddad-Provostově statistice viz [3]

V následujícím textu uvádím odvození jednotlivých testů a jejich silofunkcí.

3.1. Odvození testu založeného na T statistice

Definice 3.1.1. Mějme náhodné veličiny $Y_1, ..., Y_n$ a matici čísel $\mathbf{X} = (x_{ij})$ typu $n \times k$, kde $k < n$, takzvanou **matici plánu**. Předpokládejme, že sloupce matice plánu jsou lineárně nezávislé, takže hodnost matice je k. Předpokládejme také, že pro náhodný vektor $\mathbf{Y} = (Y_1, ..., Y_n)'$ platí

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{e},$$

kde $\boldsymbol{\beta} = (\beta_1, ..., \beta_k)'$ je vektor neznámých parametrů a $\mathbf{e} = (e_1, ..., e_n)'$ je náhodný vektor splňující podmínky

$$E\mathbf{e} = 0 \qquad a \qquad \operatorname{var}\mathbf{e} = \sigma^2\mathbf{I}.$$

Přitom $\sigma^2 > 0$ je rovněž neznámý parametr. Tento model budeme nazývat **regresní model**. Můžeme hovořit o lineárním regresním modelu, protože $\mathbf{Y}$ závisí na $\boldsymbol{\beta}$ lineárně. Je třeba si uvědomit, že vektor $\mathbf{X}\boldsymbol{\beta}$ je nenáhodný a proto platí

$$E\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} \qquad a \qquad \operatorname{var}\mathbf{Y} = \sigma^2\mathbf{I}.$$

V našem případě budeme používat model, kde $k = 2$ a

$$Y_i = \beta_0 + \beta_1 x_i + e_i, i = 1, ..., n. \tag{3.1.1}$$

Matice plánu $\mathbf{X}$ je typu $n \times 2$ a je tvořena sloupcem jedniček a sloupcem $(x_1, ..., x_n)'$.

Definice 3.1.2. Parametry $\beta_1, ..., \beta_k$ se odhadují **metodou nejmenších čtverců** tak, že výraz $(\mathbf{Y}-\mathbf{X}\boldsymbol{\beta})'(\mathbf{Y}-\mathbf{X}\boldsymbol{\beta})$ minimalizujeme podle $\boldsymbol{\beta}$. Označme tyto odhady $\mathbf{b} = (b_1, ..., b_k)'$.

Věta 3.1.1. *Odhady metodou nejmenších čtverců jsou* $\boldsymbol{b} = (\boldsymbol{X}'\boldsymbol{X})^{-1}(\boldsymbol{X}'\boldsymbol{Y})$.

Věta 3.1.2. ***Reziduální součet čtverců*** *je minimum kritéria metody nejmenších čtverců a lze vyjádřit vzorcem*

$$s_e = \boldsymbol{Y}'\boldsymbol{Y} - \boldsymbol{b}'\boldsymbol{X}'\boldsymbol{Y}$$

Věta 3.1.3. *V lineárním regresním modelu je*

$$s^2 = \frac{s_e}{n-2}$$

nestranný odhad σ^2.
Platí $E\boldsymbol{b} = \boldsymbol{\beta}$, *var* $\boldsymbol{b} = \sigma^2(\boldsymbol{X}'\boldsymbol{X})^{-1}$ *a* $\boldsymbol{b} \sim N_k(\boldsymbol{\beta}, \sigma^2(\boldsymbol{X}'\boldsymbol{X})^{-1})$. *Dále platí* $\frac{s_e}{\sigma^2} \sim \chi^2(n-k)$. *Potom vektor* $\boldsymbol{b}$ *a veličina* s^2 *jsou nezávislé. Nechť* $T_i = (b_i - \beta_i)/\sqrt{s^2 v_{ii}}$, *kde* v_{ii} *je i-tý prvek matice* $(\boldsymbol{X}'\boldsymbol{X})^{-1}$ *na diagonále. Pak pro každé* $i = 1, ..., k$ *platí* $T_i \sim t_{n-k}$.

Důkaz. Protože $\mathbf{Y}$ má normální rozdělení a $\mathbf{b}$ vzniká z $\mathbf{Y}$ lineární transformací, vektor $\mathbf{b}$ má rovněž normální rozdělení.
Dále z věty 3.1.3 plyne, že $\frac{b_i-\beta_i}{\sqrt{\sigma^2 v_{ii}}} \sim N(0,1)$. Protože $\frac{S_e}{\sigma^2} \sim \chi^2(n-k)$ a protože $\mathbf{b}$ a s^2 jsou nezávislé, má veličina

$$\frac{\frac{b_i-\beta_i}{\sqrt{\sigma^2 v_{ii}}}\sqrt{n-k}}{\sqrt{\frac{S_e}{\sigma^2}}} = \frac{b_i - \beta_i}{\sqrt{s^2 v_{ii}}} = T_i$$

Studentovo rozdělení t_{n-k}.
Další části důkazu věty 3.1.3 jsou uvedeny viz [1] str. 81, 82, 83.

□

Důsledek 3.1.1 Z věty o odhadech [1], str. 87 vyplývá, že odhady b_0, b_1 z metody nejmenších čtverců jsou dány vzorci

$$b_1 = \frac{\sum (x_i - \bar{x})(Y_i - \bar{Y})}{\sum (x_i - \bar{x})^2}, \qquad b_0 = \bar{Y} - b_1\bar{x}.$$

Důsledek 3.1.2 V modelu 3.1.1 má s^2 tvar

$$s^2 = \frac{\sum Y_i^2 - b_0 \sum Y_i - b_1 \sum x_i Y_i}{n-2},$$

což je

$$s^2 = \frac{\sum (Y_i - \bar{Y})^2 - b_1 \sum (x_i - \bar{x})(Y_i - \bar{Y})}{n-2},$$

kde b_0, b_1 jsou odhady nejmenších čtverců.

Poznámka. Dále budeme značit $\mathbf{Z} = (\mathbf{X}', \mathbf{Y}')' \sim N_2(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, kde $\boldsymbol{\mu}$ je vektor středních hodnot a $\boldsymbol{\Sigma}$ varianční matice

$$\boldsymbol{\mu} = \begin{bmatrix} \mu_X \\ \mu_Y \end{bmatrix} \qquad \boldsymbol{\Sigma} = \begin{bmatrix} \sigma_X^2 & \sigma_X\sigma_Y\rho \\ \sigma_X\sigma_Y\rho & \sigma_Y^2 \end{bmatrix}$$

Věta 3.1.4. *Nechť* $\boldsymbol{Z}$ *má regulární rozdělení* $N_2(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. *Pak podmíněné rozdělení* $\boldsymbol{X}$ *při daném* $\boldsymbol{Y} = y$ *je regulární*

$$N_2(\mu_X + \boldsymbol{\Sigma}_{XY}\boldsymbol{\Sigma}_{YY}^{-1}(y - \mu_Y), \boldsymbol{\Sigma}_{XX} - \boldsymbol{\Sigma}_{XY}\boldsymbol{\Sigma}_{YY}^{-1}\boldsymbol{\Sigma}_{YX}). \tag{3.1.2}$$

Důkaz. [1], str. 67

Věta 3.1.5. *Nechť* $Z_1, ..., Z_n$ *je náhodný výběr z dvojrozměrného normálního rozdělení, který má kladné rozptyly, korelační koeficient* $\rho = 0$ *a rozsah* $n \geq 3$. *Potom*

$$T = \frac{r}{\sqrt{1 - r^2}}\sqrt{n-2} \sim t_{n-2}.$$

Důkaz. Označme střední hodnoty $Z_i = (X_i, Y_i)$ μ_X a μ_Y a rozptyly X a Y jako σ_X^2, σ_Y^2. Korelační koeficient mezi X_i a Y_i je $\rho \in (-1, 1)$, pak z věty 3.1.4 vyplývá, že podmíněné rozdělení Y_i, při daných $X_1 = x_1, ..., X_n = x_n$, je normální

$$N\left(\mu_Y + \rho\frac{\sigma_Y}{\sigma_X}(x_i - \mu_X), \sigma_Y^2(1 - \rho^2)\right) \qquad i = 1, ..., n.$$

Položme nyní

$$\rho\frac{\sigma_Y}{\sigma_X} = \beta_1, \qquad \mu_Y - \beta_1\mu_X = \beta_0.$$

Za předpokladu $\rho = 0$ lze podmíněné rozdělení veličin $Y_1, ..., Y_n$ popsat modelem

$$Y_i = \beta_0 + \beta_1 x_i + e_i \qquad pro \quad i = 1, ..., n,$$

kde $\beta_1 = 0$ a $e_1, ..., e_n$ jsou nezávislé náhodné veličiny s rozdělením $N(0, \sigma_Y^2)$. Rovnice je tedy modelem lineární regrese.
Úpravou vzorců odhadů nejmenších čtverců z důsledku 3.1.1 dostáváme

$$b_1 = r\sqrt{\frac{\sum (Y_i - \bar{Y})^2}{\sum (x_i - \bar{x})^2}}, \qquad s^2 = \frac{1 - r^2}{n-2}\sum (Y_i - \bar{Y})^2.$$

Proto

$$T = \frac{b_1}{s}\sqrt{\sum (x_i - \bar{x})^2} = \frac{r}{\sqrt{1 - r^2}}\sqrt{n-2}.$$

Podle věty 3.1.3 platí, že $T \sim t(n-2)$. Jelikož toto rozdělení nezávisí na podmínce $X_1 = x_1, ..., X_n = x_n$, je totožné s obyčejným nepodmíněným rozdělením veličiny T.

□

3.2. Odvození testu založeného na Fisherově transformaci

Věta 3.2.1. *Nechť* $\boldsymbol{Z}_1, ..., \boldsymbol{Z}_n$ *je náhodný výběr z dvojrozměrného normálního rozdělení, který má kladné rozptyly a rozsah* $n \geq 3$. *Pro* ***Fisherovu z-transformaci*** *tvaru*

$$Z = \frac{1}{2}\ln\frac{1 + \rho}{1 - \rho},$$

platí

$$Z \overset{as}{\sim} N\left(\frac{1}{2}\ln\frac{1 + \rho}{1 - \rho}, \frac{1}{n-3}\right).$$

Důkaz. [2]

Poznámka. V praxi se tato aproximace využívá už při rozsahu $n \geq 10$ není-li ρ blízké -1 nebo 1. Fisherovu Z-transformaci lze užít pro test $H_0 : \rho = \rho_0 \in (-1, 1)$.

Věta 3.2.2. *Uvažujme náhodnou veličinu X jejíž, rozdělení závisí na nějakém parametru θ. Předpokládejme, že střední hodnotou X je právě θ. Jelikož na θ závisí většinou i rozptyl, budeme psát* $\operatorname{var} X = \sigma^2(\theta)$. *Pak transformace stabilizující rozptyl je*

$$g(\theta) = c \int \frac{d\theta}{\sigma(\theta)}. \tag{3.2.1}$$

Konstanta c se volí tak, aby funkce g vypočtená podle vzorce 3.2.1 měla vhodný tvar. Pro takovou transformaci dostaneme veličinu $g(X)$ s rozptylem var $g(X) \doteq c^2$.

Důkaz. Hledejme funkci g, která se nerovná konstantě, tak aby veličina $Y = g(X)$ měla rozptyl nezávislý na θ. Takovou funkci je obtížné analyticky stanovit, takže se pokusíme alespoň o aproximaci. Je-li funkce g dostatečně hladká, pak z Taylorova polynomu

$$g(X) \doteq g(\theta) + (X - \theta)g'(\theta).$$

Proto

$$Eg(X) \doteq g(\theta), \qquad var\ g(X) = [g'(\theta)]^2\sigma^2(\theta). \tag{3.2.2}$$

Platí-li rovnice

$$g'(\theta)\sigma(\theta) = c,$$

kde c je konstanta, výraz $[g'(\theta)]^2\sigma^2(\theta)$ nebude záviset na θ.

□

Věta 3.2.3. *Pro Výběrový korelační koeficient r, počítaný z náhodného výběru o rozsahu n z regulárního dvojrozměrného normálního rozdělení s korelačním koeficientem ρ, platí*

$$Er \doteq \rho, \qquad var\ r = \frac{(1 - \rho^2)^2}{n}.$$

Důsledek 3.2.1 Nechť $\boldsymbol{Z}_1, ..., \boldsymbol{Z}_n$ je náhodný výběr z dvojrozměrného normálního rozdělení, který má kladné rozptyly a rozsah $n \geq 3$. Pak transformace stabilizující rozptyl r je tvaru

$$g(r) = \frac{1}{2} \ln \frac{1 + r}{1 - r}.$$

Důkaz. Jelikož $\sigma(\rho) \doteq (1 - \rho^2)/\sqrt{n}$, máme

$$g(\rho) = c\sqrt{n}\frac{1}{2} \ln \frac{1 + \rho}{1 - \rho}.$$

Položme

$$c = 1/\sqrt{n},$$

pak

$$g(\rho) = arctgh\ \rho,$$

což je tabelovaná funkce. Když dosadíme do vzorce 3.2.2, tak dostaneme

$$EZ \doteq \frac{1}{2} \ln \frac{1 + \rho}{1 - \rho}, \qquad var\ Z = \frac{1}{n}.$$

□

Věta 3.2.4. *Nechť $\boldsymbol{Z}_1, ..., \boldsymbol{Z}_n$ je náhodný výběr z dvojrozměrného normálního rozdělení. Pak šikmost a špičatost Z z věty 3.2.3 je*

$$\alpha_3 \doteq \frac{\rho^6}{(n-1)^3}, \qquad \alpha_4 \doteq 3 + \frac{2}{n-1} + \frac{4+2\rho^2-3\rho^4}{(n-1)^2}.$$

Důsledek 3.2.2 S rostoucím rozsahem n se šikmost Z blíží nule a špičatost trojce. Proto je vedlejším efektem Fisherovy Z transformace normalita. Viz též [2].

Důsledek 3.2.3 Položme $z_0 = \frac{1}{2}\ln\frac{1+\rho_0}{1-\rho_0}$, kde $\rho_0 = 0$, takže $z_0 = \frac{1}{2}\ln(1) = 0$. Pak za platnosti $H_0 : \rho = 0$ má veličina Z přibližně rozdělení $N(0, \frac{1}{n-3})$. Tudíž

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

Tedy, platí-li $|U| \in W = (-\infty, -u_{1-\alpha/2)} \cup (u_{1-\alpha/2}, \infty)$, zamítneme hypotézu na hladině, významnosti α.

3.3. Odvození testu založeného na Hotellingově transformaci

Definice 3.3.1. Test pomocí Fisherovy transformace má pro rozsahy $n < 25$ malou sílu. Pro tyto výběry někteří autoři doporučují použít raději Hotellingovu transformaci.

$$Z^* = Z - \frac{3Z + tgh\, Z}{4(n-1)}, kde \qquad tgh\, x = \frac{e^x - e^{-x}}{e^x + e^{-x}}$$

je hyperbolický tangens.

Věta 3.3.1. *Statistika Z^* má rozdělení*

$$Z^* \overset{as}{\sim} N\left(\frac{1}{2}\ln\frac{1+\rho}{1-\rho} - \frac{\frac{3}{2}\ln\frac{1+\rho}{1-\rho} + \rho}{4(n-1)}, \frac{1}{n-1}\right).$$

Věta 3.3.2. *Hotellingovu statistiku lze napsat ve tvaru*

$$Z^* = Z - \frac{3Z + r}{4(n-1)}$$

Důkaz.

$$tgh\, Z = \frac{e^{\frac{1}{2}ln\frac{1+r}{1-r}} - e^{-\frac{1}{2}ln\frac{1+r}{1-r}}}{e^{\frac{1}{2}ln\frac{1+r}{1-r}} + e^{-\frac{1}{2}ln\frac{1+r}{1-r}}} = \frac{\sqrt{\frac{1+r}{1-r}} - \frac{1}{\sqrt{\frac{1+r}{1-r}}}}{\sqrt{\frac{1+r}{1-r}} + \frac{1}{\sqrt{\frac{1+r}{1-r}}}} = \frac{\frac{1+r}{1-r} - 1}{\frac{1+r}{1-r} + 1} = \frac{2r}{2} = r$$

□

3.4. Odvození testu založeného na Haddad-Provostově statistice

Věta 3.4.1. *Nechť* $\boldsymbol{Z} = (X', Y')' \sim N_2(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, *kde* $\boldsymbol{\mu}$ *je vektor středních hodnot a* $\boldsymbol{\Sigma}$ *varianční matice, přičemž*

$$\boldsymbol{\mu} = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \qquad \boldsymbol{\Sigma} = \begin{bmatrix} \sigma^2 & \rho\sigma^2 \\ \rho\sigma^2 & \sigma^2 \end{bmatrix}. \tag{3.4.1}$$

Pak Haddad-Provostova statistika má rozdělení

$$F = \frac{\frac{\sum_{i=1}^{n}(X_i+Y_i)^2}{2\sigma^2(1+\rho)}}{\frac{\sum_{i=1}^{n}(X_i-Y_i)^2}{2\sigma^2(1-\rho)}} \sim F(n,n).$$

Důkaz.

$$D(X_i + Y_i) = DX_i + DY_i + 2C(X_i, Y_i) = 2\sigma^2 + 2\rho\sigma^2 = 2\sigma^2(1+\rho)$$
$$D(X_i - Y_i) = DX_i + DY_i - 2C(X_i, Y_i) = 2\sigma^2 - 2\rho\sigma^2 = 2\sigma^2(1-\rho).$$

Z toho plyne

$$X_i + Y_i \sim N(0, 2\sigma^2(1+\rho)), \qquad X_i - Y_i \sim N(0, 2\sigma^2(1-\rho))$$

a

$$\frac{\sum_{i=1}^{n}(X_i+Y_i)^2}{2\sigma^2(1+\rho)} \sim \chi^2(n), \qquad \frac{\sum_{i=1}^{n}(X_i-Y_i)^2}{2\sigma^2(1-\rho)} \sim \chi^2(n).$$

Navíc platí , že náhodné veličiny $X_i + Y_i$ a $X_i - Y_i$ pro i=1,…,n jsou nezávislé, protože

$$C(X_i+Y_i, X_i-Y_i) = C(X_i, X_i) - C(X_i, Y_i) + C(Y_i, X_i) - C(Y_i, Y_i) = D(X_i) - D(Y_i) = 0.$$

□

Odkud plyne
Důsledek 3.1.1 Při hypotéze $H_0 : \rho = 0$ platí

$$F = \frac{\sum_{i=1}^{n}(X_i+Y_i)^2}{\sum_{i=1}^{n}(X_i-Y_i)^2} \sim F(n,n),$$

což lze použít při testování nulovosti korelačního koeficientu v mé bakalářské práci.

3.5. Odvození asymptotické silofunkce testu založeného na Fisherově transformaci

Věta 3.5.1. *Asypmtotická silofunkce testu založeného na Fisherově transformaci je tvaru*

$$\gamma(\rho_0) = 2 - \Phi(u_{1-\alpha/2} - \lambda(\rho)) - \Phi(u_{1-\alpha/2} + \lambda(\rho)),$$

kde Φ *je distibuční funkce normálního rozdělení a*

$$\lambda(\rho_0) = \frac{1}{2}\ln\frac{1+\rho_0}{1-\rho_0}\sqrt{n-3}.$$

Důkaz. Silofunkce testu $H_0 : \rho_0 = 0$ proti oboustranné alternativě $H_1 : \rho \neq 0$.

$$\begin{aligned}\gamma(\rho_0) &= P(Z \in W|\rho = \rho_0)\\ &= 1 - P(Z \notin W|\rho = \rho_0) =\\ &= 1 - P\left(-u_{1-\alpha/2} < Z\sqrt{n-3} < u_{1-\alpha/2}\right)\end{aligned}$$

Když označíme

$$Z_0 = \frac{1}{2}\ln\frac{1+\rho_0}{1-\rho_0},$$

lze psát

$$\gamma(\rho_0) = 1 - P\left(-u_{1-\alpha/2} < (Z - Z_0)\sqrt{n-3} + (Z_0)\sqrt{n-3} < u_{1-\alpha/2}\right).$$

Dle důsledku 3.2.1 je $U = (Z - Z_0)\sqrt{n-3} \sim N(0,1)$, pak lze

$$\begin{aligned}\gamma(\rho_0) &= 1 - P\left(-u_{1-\alpha/2} < U + \frac{1}{2}\ln\frac{1+\rho_0}{1-\rho_0}\sqrt{n-3} < u_{1-\alpha/2}\right) =\\ &= 1 - P\left(-u_{1-\alpha/2} < U + \lambda(\rho_0) < u_{1-\alpha/2}\right) =\\ &= 1 - P\left(-u_{1-\alpha/2} - \lambda(\rho_0) < U < u_{1-\alpha/2} - \lambda(\rho_0)\right) =\\ &= 1 - \Phi(u_{1-\alpha/2} - \lambda(\rho_0)) + \Phi(-u_{1-\alpha/2} - \lambda(\rho_0)) =\\ &= 1 - \Phi(u_{1-\alpha/2} - \lambda(\rho_0)) + 1 - \Phi(u_{1-\alpha/2} + \lambda(\rho_0)) =\\ &= 2 - \Phi(u_{1-\alpha/2} - \lambda(\rho_0)) - \Phi(u_{1-\alpha/2} + \lambda(\rho_0))\end{aligned}$$

□

3.6. Odvození asymptotické silofunkce testu založeného na Hotellingově transformaci

Věta 3.6.1. *Asypmtotická silofunkce testu založeného na Hotellingově transformaci je tvaru*

$$\gamma(\rho_0) = 2 - \Phi(u_{1-\alpha/2} - \lambda(\rho)) - \Phi(u_{1-\alpha/2} + \lambda(\rho)),$$

kde Φ *je distibuční funkce normálního rozdělení a*

$$\lambda(\rho_0) = \left(\frac{1}{2}\ln\frac{1+\rho_0}{1-\rho_0} - \frac{\frac{3}{2}\ln\frac{1+\rho_0}{1-\rho_0} + \rho_0}{4(n-1)}\right)\sqrt{n-1}.$$

Důkaz. Silofunkce testu $H_0 : \rho_0 = 0$ proti oboustranné alternativě $H_1 : \rho_0 \neq 0$.

$$\begin{aligned}\gamma(\rho_0) &= P(Z^* \in W|\rho = \rho_0)\\ &= 1 - P(Z^* \notin W|\rho = \rho_0) =\\ &= 1 - P\left(-u_{1-\alpha/2} < Z^*\sqrt{n-1} < u_{1-\alpha/2}\right)\end{aligned}$$

Když označíme

$$Z_0 = \frac{1}{2}\ln\frac{1+\rho_0}{1-\rho_0} - \frac{\frac{3}{2}\ln\frac{1+\rho_0}{1-\rho_0} + \rho_0}{4(n-1)},$$

lze psát

$$\gamma(\rho_0) = 1 - P\left(-u_{1-\alpha/2} < (Z^* - Z_0)\sqrt{n-1} + (Z_0)\sqrt{n-1} < u_{1-\alpha/2}\right).$$

Dle důsledku 3.2.3 je $U = (Z - Z0)\sqrt{n-3} \sim N(0,1)$, pak lze

$$\begin{aligned}
\gamma(\rho_0) &= 1 - P\left(-u_{1-\alpha/2} < U + \left(\frac{1}{2}\ln\frac{1+\rho_0}{1-\rho_0} - \frac{\frac{3}{2}\ln\frac{1+\rho_0}{1-\rho_0} + \rho_0}{4(n-1)}\right)\sqrt{n-1} < u_{1-\alpha/2}\right) = \\
&= 1 - P\left(-u_{1-\alpha/2} < U + \lambda(\rho_0) < u_{1-\alpha/2}\right) = \\
&= 1 - P\left(-u_{1-\alpha/2} - \lambda(\rho_0) < U < u_{1-\alpha/2} - \lambda(\rho_0)\right) = \\
&= 1 - \Phi(u_{1-\alpha/2} - \lambda(\rho_0)) + \Phi(-u_{1-\alpha/2} - \lambda(\rho_0)) = \\
&= 1 - \Phi(u_{1-\alpha/2} - \lambda(\rho_0)) + 1 - \Phi(u_{1-\alpha/2} + \lambda(\rho_0)) = \\
&= 2 - \Phi(u_{1-\alpha/2} - \lambda(\rho_0)) - \Phi(u_{1-\alpha/2} + \lambda(\rho_0))
\end{aligned}$$

□

3.7. Odvození silofunkce testu založeného na Haddad--Provostově statistice

Věta 3.7.1. *Silofunkce testu založeného na Haddad-Provostově statistice je tvaru*

$$\gamma(\rho_0) = 1 - F(f_{1-\alpha/2}(n,n)\lambda(\rho_0)) + F(f_{\alpha/2}(n,n)\lambda(\rho_0)),$$

kde F je distibuční funkce Fisher-Snedecorova rozdělení $F(n,n)$ a

$$\lambda(\rho_0) = \frac{1-\rho_0}{1+\rho_0}.$$

Důkaz. Silofunkce testu $H_0 : \rho_0 = 0$ proti oboustranné alternativě $H_1 : \rho_0 \neq 0$.

$$\begin{aligned}
\gamma(\rho_0) &= P(F \in W|\rho = \rho_0) \\
&= 1 - P(F \notin W|\rho = \rho_0) = \\
&= 1 - P\left(f_{\alpha/2}(n,n) < F < f_{1-\alpha/2}(n,n)\right)
\end{aligned}$$

Když označíme

$$F_0 = \frac{1-\rho}{1+\rho} \qquad a \qquad F\frac{1-\rho}{1+\rho} \sim F(n,n)$$

lze psát

$$\begin{aligned}
\gamma(\rho_0) &= 1 - P\left(f_{\alpha/2}(n,n)\lambda(\rho_0) < F\frac{1-\rho_0}{1+\rho_0} < f_{1-\alpha/2}(n,n)\lambda(\rho_0)\right) = \\
&= 1 - P\left(f_{\alpha/2}(n,n)\lambda(\rho_0) < F\lambda(\rho_0) < f_{1-\alpha/2}(n,n)\lambda(\rho_0)\right) = \\
&= 1 - F(f_{1-\alpha/2}(n,n)\lambda(\rho_0)) + F(f_{\alpha/2}(n,n)\lambda(\rho_0))
\end{aligned}$$

□

4. Porovnání testů nulovosti korelačního koeficientu pomocí simulací

Testy nulovosti korelačního koeficientu náhodného výběru ze dvou normálních náhodných veličin jsem porovnával v programu MATLAB. K simulaci výběru jsem používal příkaz $mvnrnd(\underline{\mu}, \underline{\Sigma}, n)$, kde $\underline{\mu}$ je vektor středních hodnot normálních náhodných veličin. V našem případě byly nastaveny hodnoty

$$\underline{\mu} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$

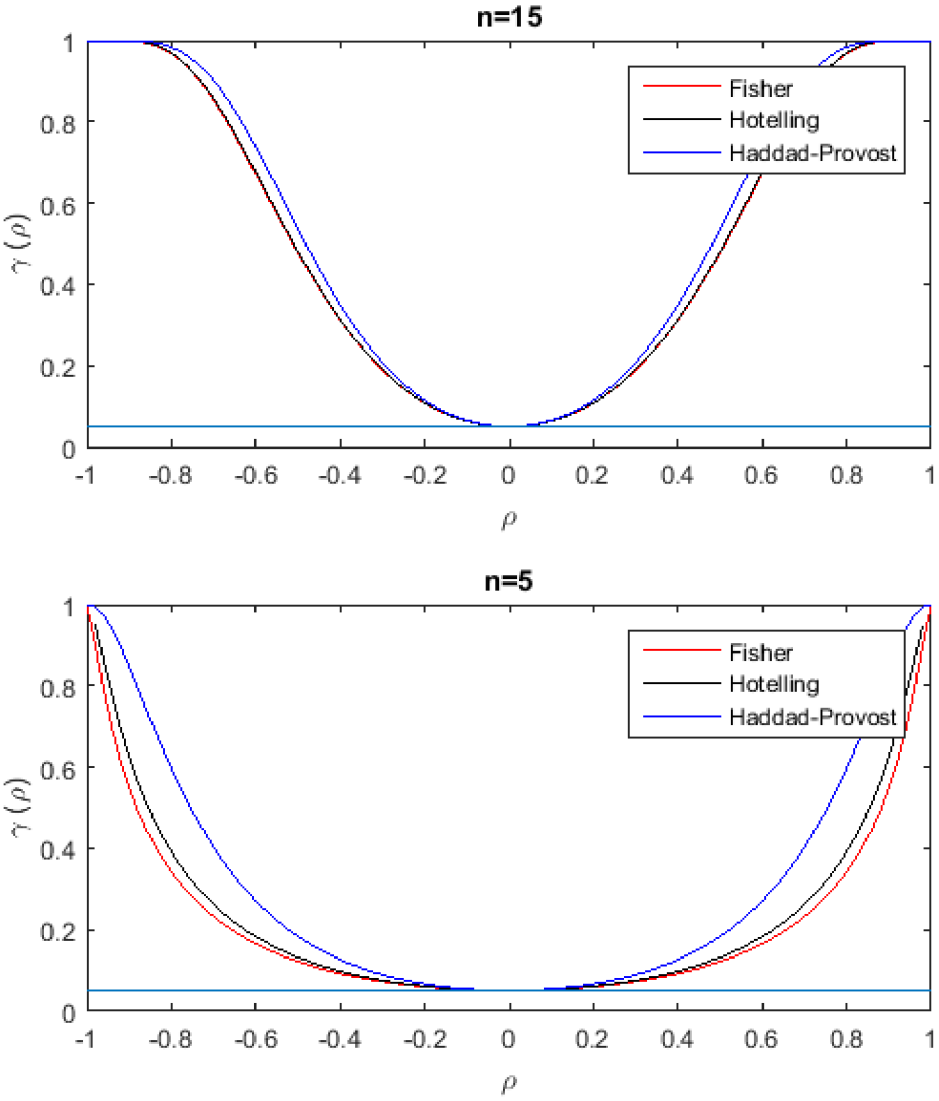
a $\underline{\Sigma}$ jejich varianční matice, jejíž strukturu můžete vidět ve vzorci 3.0.1 v úvodu třetí kapitoly.
V našem případě byly nastaveny hodnoty $\sigma_X^2 = \sigma_Y^2 = 5$, nebo později $\sigma_X^2 = 2$ a $\sigma_Y^2 = 5$, takže

$$\underline{\Sigma} = \begin{bmatrix} 5 & 5\rho \\ 5\rho & 5 \end{bmatrix} \qquad a\,nebo \qquad \underline{\Sigma} = \begin{bmatrix} 2 & \sqrt{2}\sqrt{5}\rho \\ \sqrt{2}\sqrt{5}\rho & 5 \end{bmatrix}$$

Za ρ jsem dosazoval postupně hodnoty $< -1, 1 >$. K získání výběrových průměrů a výběrových rozptylů všech rozsahů simulací a hladin významností jsem použil funkci $normfit(X)$.

4.1. Porovnání asymptotických silofunkcí jednotlivých testů nulovosti korelačního koeficientu

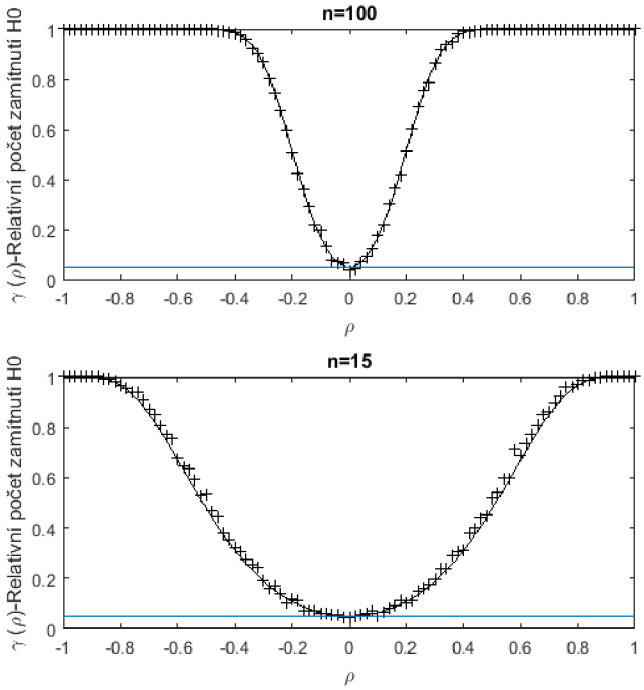
Na obr. 4.1 můžeme vidět dva grafy porovnání silofunkcí testů označených v legendě grafu. Osa x je korelační koeficient veličin X a Y. Osa y zobrazuje hodnoty silofunkce pro daný korelační koeficient $\gamma(\rho)$. Jelikož víme, že síla funkce s rostoucím rozsahem roste, z prvního grafu odvozujeme, že pro hodnoty $n > 15$ už budou silofunkce pomocí Fisherovy a Hotellingovy transformace prakticky totožné. Dále vidíme, že pro hodnoty $n = 5, 15$ je silofunkce testu pomocí Haddad-Provostovy statistiky nejvyšší a předpokládáme, že pro všechny ostatní rozsahy taktéž.



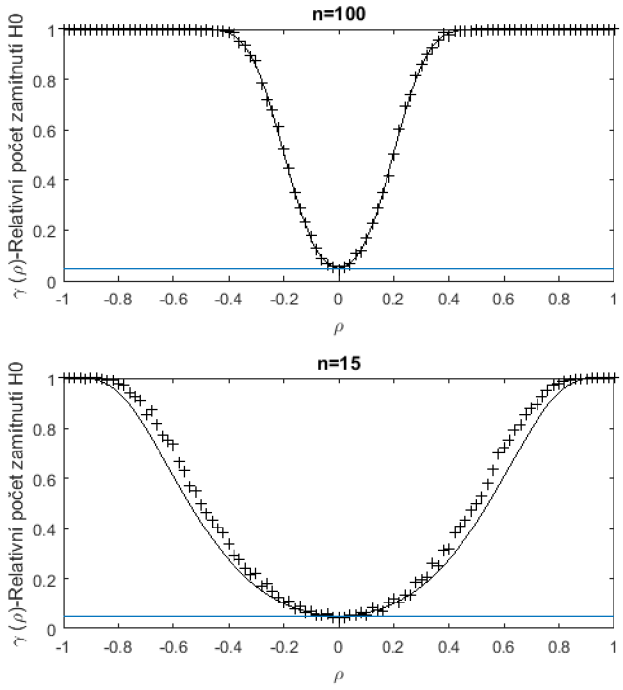
Obrázek 4.1: Porovnání silofunkcí jednotlivých testů nulovosti korelačního koeficientu

4.2. Porovnání silofunkcí a k nim náležitých testů nulovosti korelačního koeficientu pomocí simulací

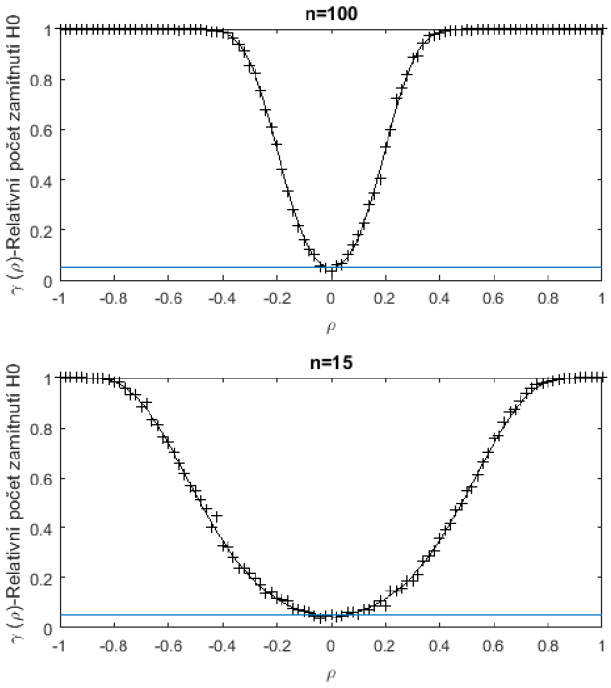
V této kapitole porovnávám křivky silofunkcí daných testů nulovosti korelačního koeficientu s jejich aproximacemi pomocí simulací výběrů. Obrázek 4.2 ukazuje porovnání simulované $\gamma(\rho)$ Fisherovy transformace s její aproximací a obrázek 4.3 Hotellingovy transformace s její aproximací.



Obrázek 4.2: Porovnání silofunkce testu založeného na Fisherově transformaci a její aproximací



Obrázek 4.3: Porovnání silofunkce testu založeného na Hotellingově transformaci a její aproximací



Obrázek 4.4: Porovnání silofunkce testu založeného na Haddad-Provostově statistice a její aproximací

4.3. Porovnání testu nulovosti korelačního koeficientu náhodného výběru ze dvou normálních náhodných veličin se stejným rozptylem pomocí simulací

Pro zjednodušení budu v dalším textu budu používat následující zkrácené výrazy:
T test namísto test korelačního koeficientu založený na T statistice,
Z test namísto test korelačního koeficientu založený na Fisherově transformaci,
Z^* test namísto test korelačního koeficientu založený na Hotellingově transformaci a
F test namísto test korelačního koeficientu založený na Haddad-Provostově statistice.

4.3.1. Vykreslení aproximace silofunkce

U každého testu jsem provedl tisíc opakování pro daný korelační koeficient. Následně jsem spočítal počet zamítnutí a podělil celkovým počtem testů. Tím jsem získal relativní počet zamítnutí hypotézy při daném rozsahu náhodného výběru a daném korelačním koeficientu náhodných veličin.
Výše uvedené testování jsem prováděl pomocí programu matlab postupně pro měnící se ρ od -1 do 1. Na obrázku 4.5 můžete vidět výsledky T testu vykreslené v grafu závislosti relativního zamítnutí testu nulovosti na korelačním koeficientu náhodných veličin. Červené tečky odpovídají síle testu provedeného pro rozsah náhodného výběru $n = 3$, modré

pro $n = 10$ a zelené pro $n = 50$. Modrá čára dole vyjadřuje spodní hranici relativního zamítnutí při hladině významnosti $\alpha = 0,05$.

4.3.2. Barevné rozlišení

V dalších grafech budu označovat testy barevně podle schematu 4.3.2.

černá Test založený na T statistice

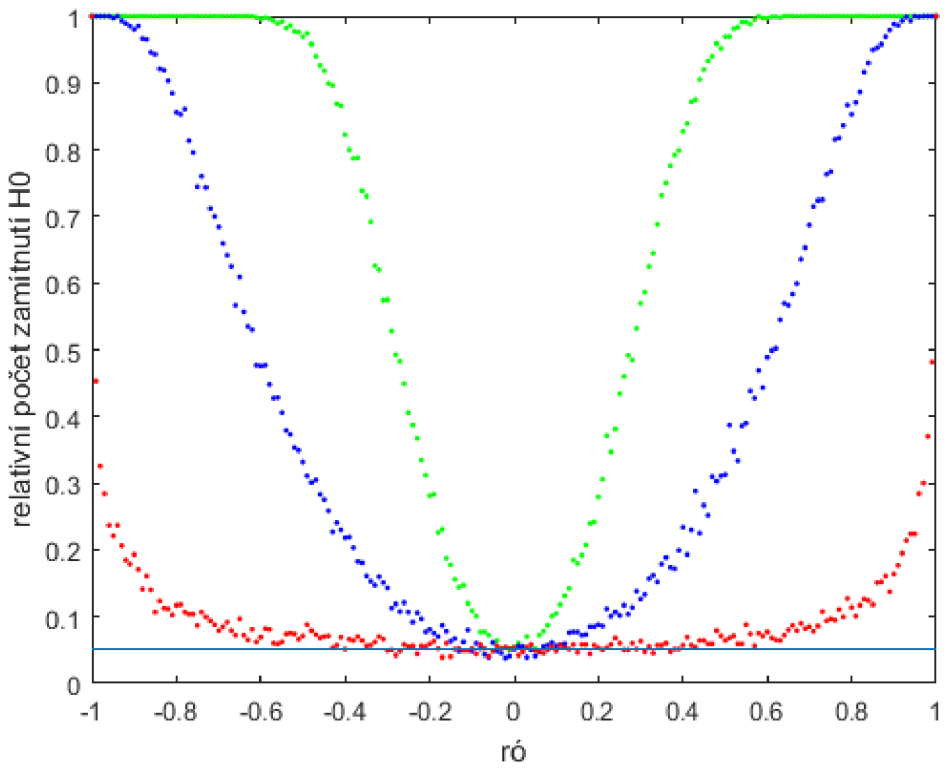
červená Test založený na Fisherově transformaci

zelená Test založený na Hotellingově transformaci

modrá Test založený na Haddad-Provostově statistice

4.3.3. Porovnání testů založených na T statistice při různých rozsazích výběru

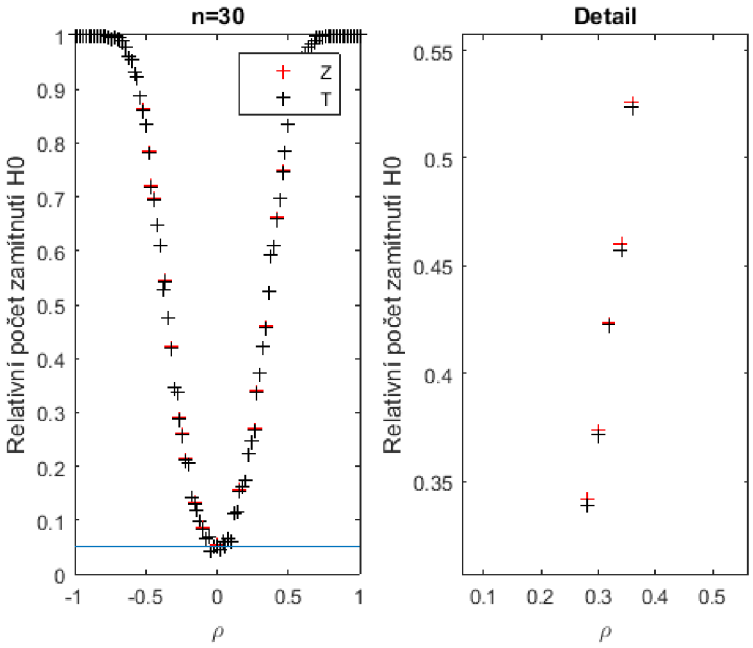
Přímou závislost síly testu na rozsahu simulovaného náhodného výběru pozorujeme v grafu 4.5, na kterém jsou tři aproximace silofunkcí testu založeného na T statistice postupně pro rozsahy $n = 5$ - červená, $n = 30$ - modrá, $n = 100$ - zelená.



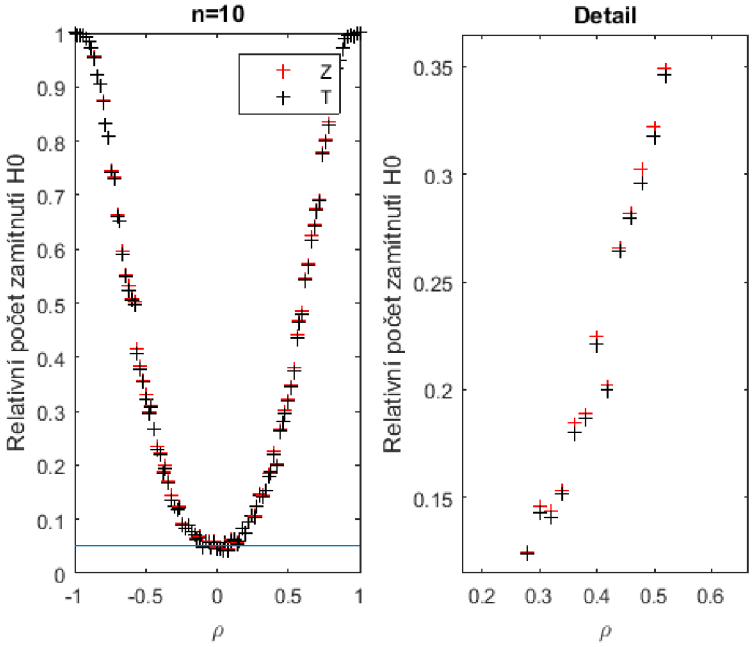
Obrázek 4.5: Relativní zamítnutí testu nulovosti při proměnlivém korelačním koeficientu

4.3.4. Porovnání testů založených na T statistice a Fisherově transformaci

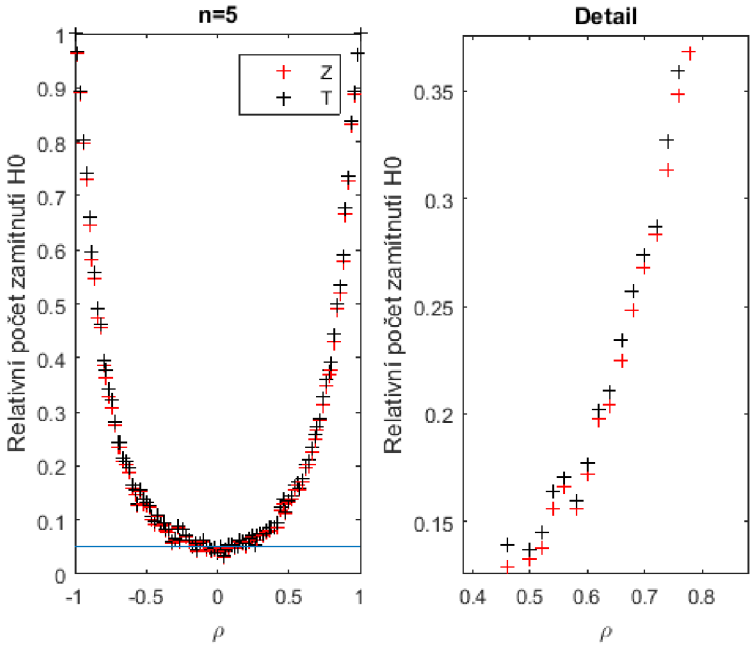
Na grafech 4.6,4.7,4.8 pozorujeme, že Z test má větší sílu pro hodnoty okolo $n = 5$ a T test pro hodnoty okolo $n = 10$, potom se už rozdíly smazávají.



Obrázek 4.6: Porovnání simulace silofunkce T testu a simulace silofunkce Z testu při rozsahu $n = 30$



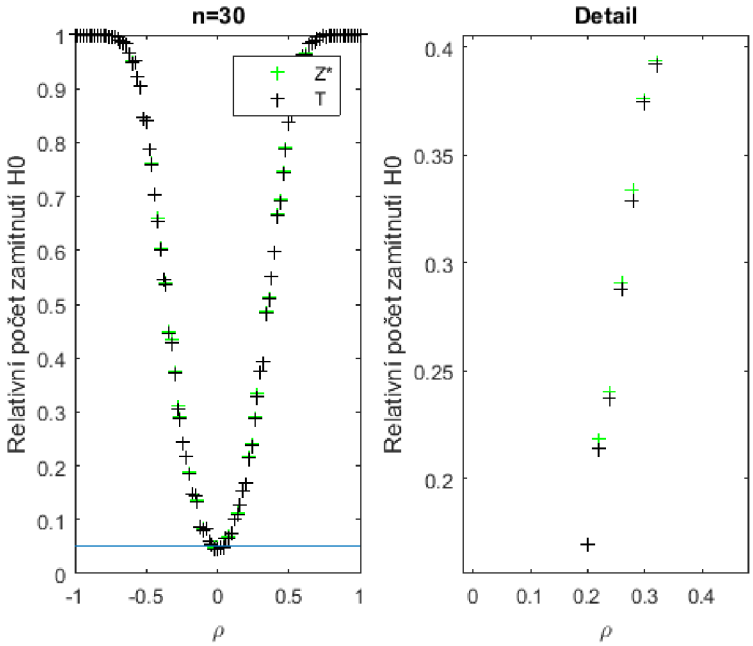
Obrázek 4.7: Porovnání simulace silofunkce T testu a simulace silofunkce Z testu při rozsahu $n = 10$



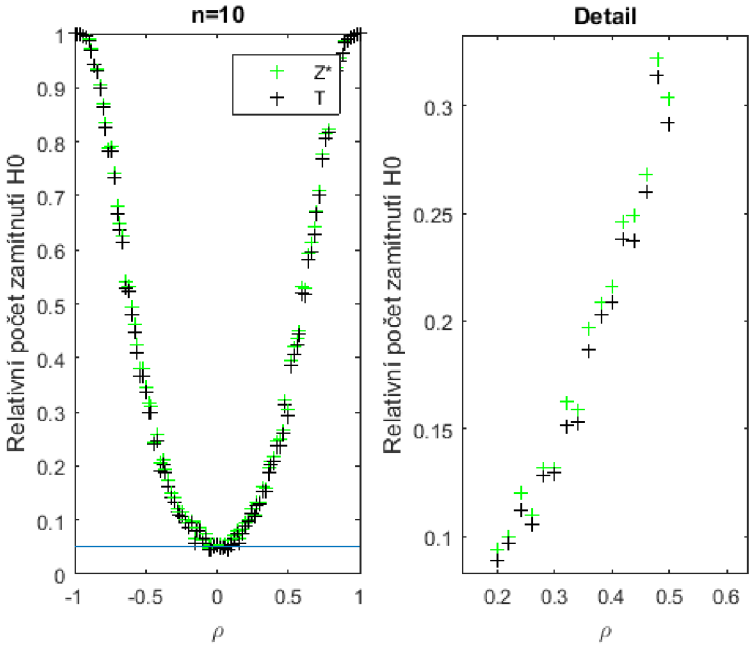
Obrázek 4.8: Porovnání simulace silofunkce T testu a simulace silofunkce Z testu při rozsahu $n = 10$

4.3.5. Porovnání testů založených na T statistice a Hotellingově transformaci

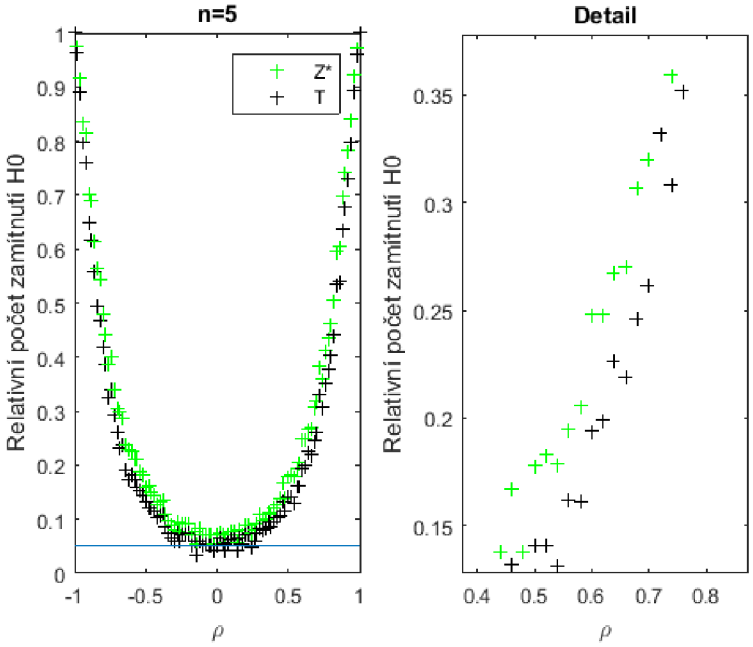
Na grafech 4.9,4.10,4.11 vidíme, že Z^* test má pro všechna n větší sílu než T test.



Obrázek 4.9: Porovnání simulace silofunkce T testu a simulace silofunkce Z^* testu při rozsahu $n = 30$



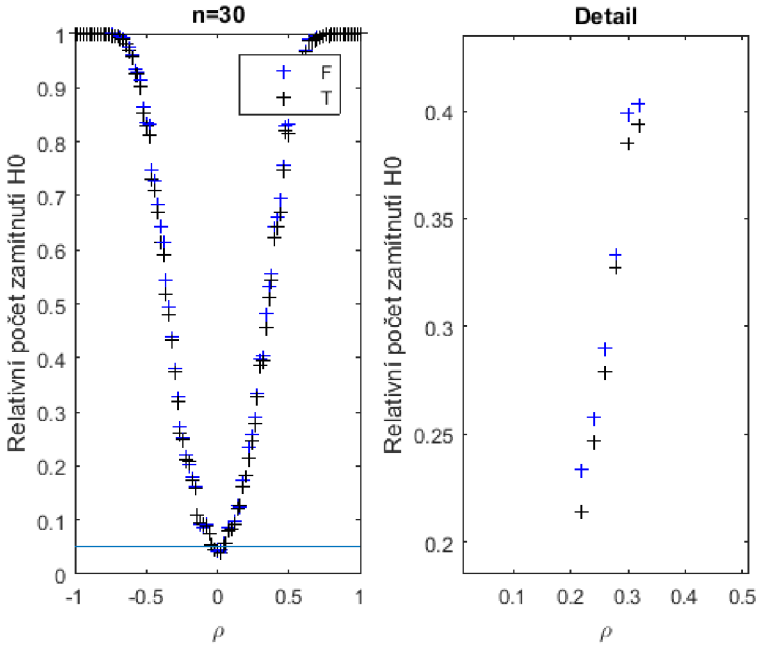
Obrázek 4.10: Porovnání simulace silofunkce T testu a simulace silofunkce Z^* testu při rozsahu $n = 10$



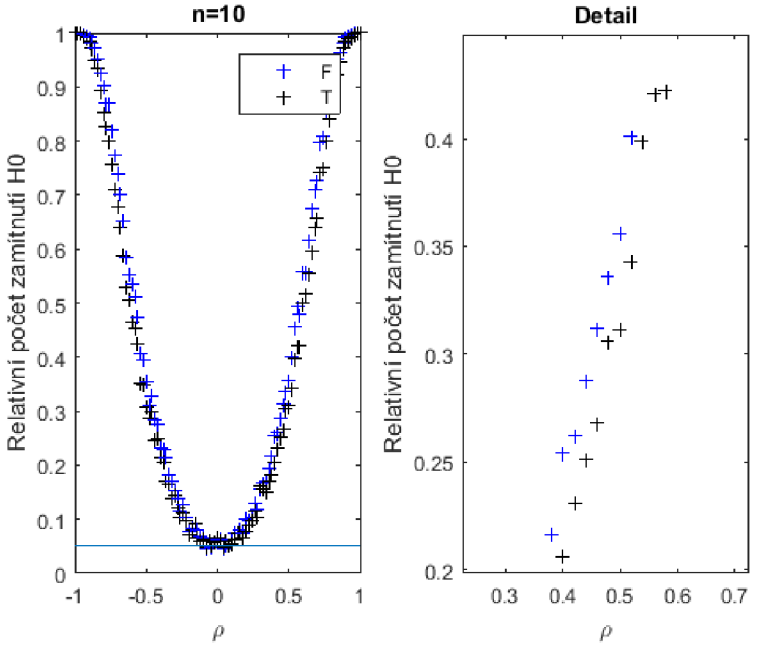
Obrázek 4.11: Porovnání simulace silofunkce T testu a simulace silofunkce Z^* testu při rozsahu $n = 5$

4.3.6. Porovnání testů založených na T statistice a Haddad-Provostově statistice

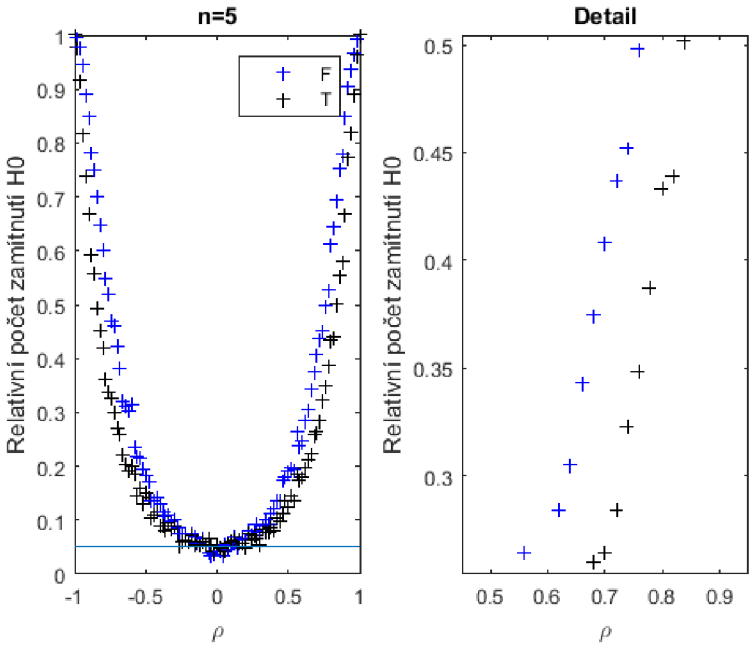
Grafy 4.12,4.13,4.14 zobrazují větší sílu F testu než T test pro všechna n.



Obrázek 4.12: Porovnání simulace silofunkce T testu a simulace silofunkce F testu při rozsahu $n = 30$



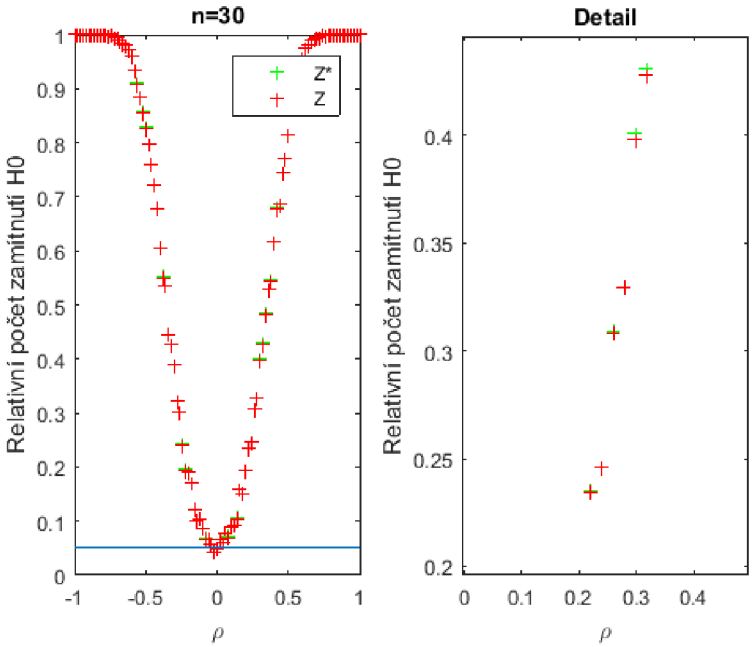
Obrázek 4.13: Porovnání simulace silofunkce T testu a simulace silofunkce F testu při rozsahu $n = 10$



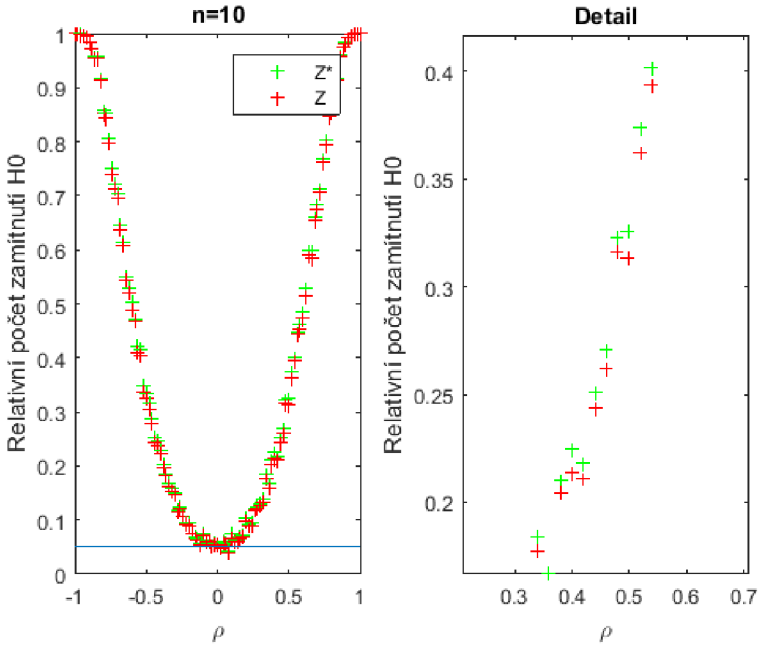
Obrázek 4.14: Porovnání simulace silofunkce T testu a simulace silofunkce F testu při rozsahu $n = 5$

4.3.7. Porovnání testů založených na Fisherově transformaci a Hotellingově transformaci

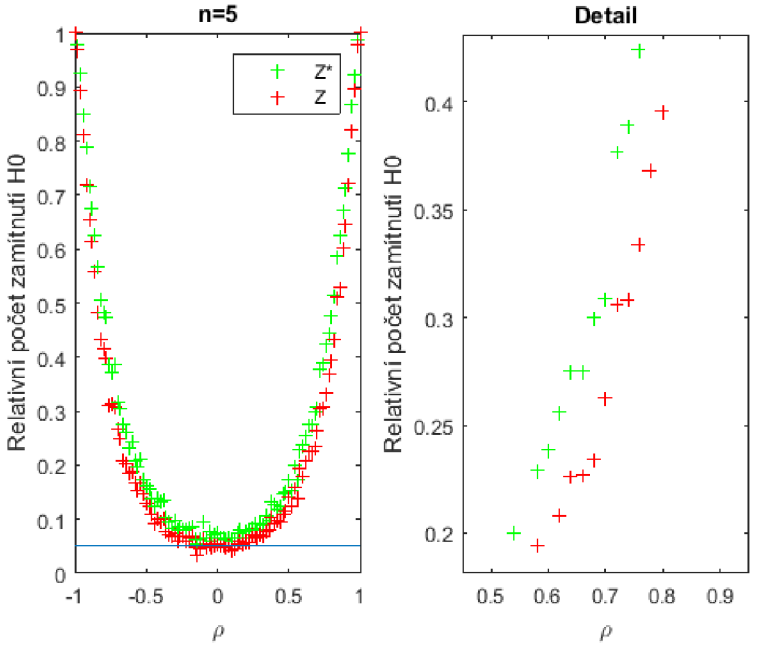
Na grafech 4.15,4.16,4.17 je možné sledovat vyšší sílu Z^* testu pro rozsahy náhodného výběru $n < 25$, což byl účel vzniku Hotellingovy transformace.



Obrázek 4.15: Porovnání simulace silofunkce Z testu a simulace silofunkce Z^* testu při rozsahu $n = 30$



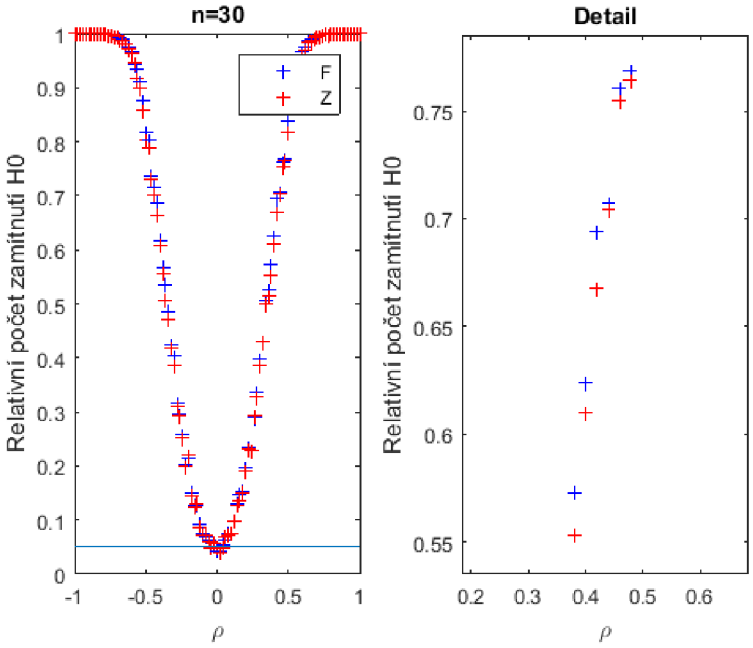
Obrázek 4.16: Porovnání simulace silofunkce Z testu a simulace silofunkce Z^* testu při rozsahu $n = 10$



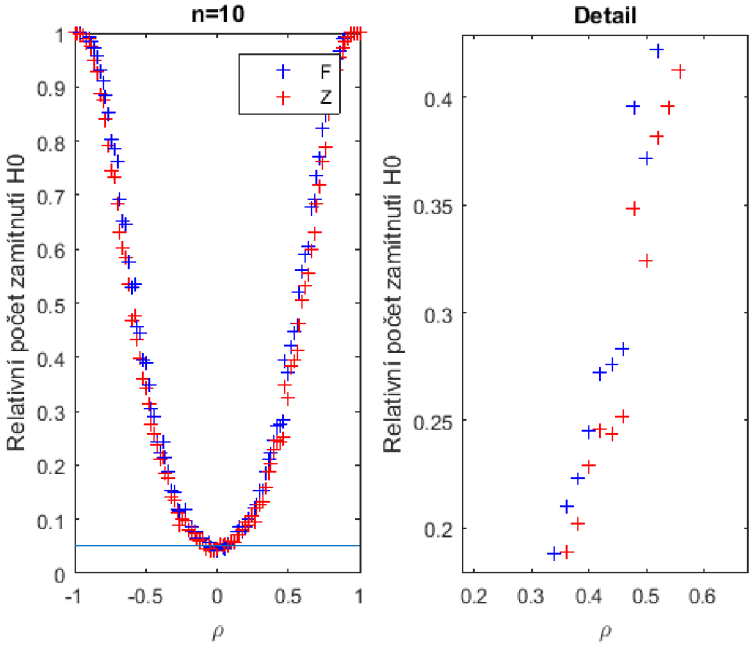
Obrázek 4.17: Porovnání simulace silofunkce Z testu a simulace silofunkce Z^* testu při rozsahu $n = 5$

4.3.8. Porovnání testů založených na Fisherově transformaci a Haddad-Provostově statistice

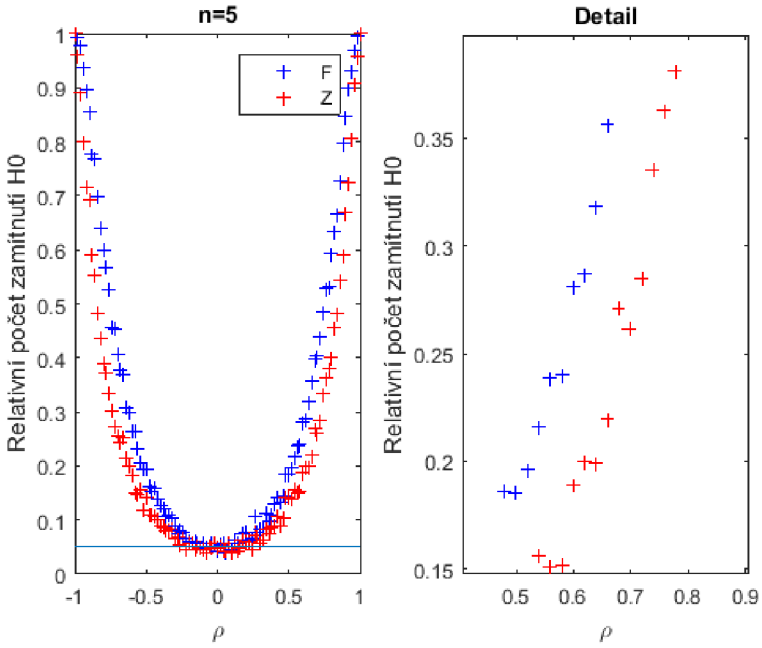
Na grafech 4.18,4.19,4.20 pozorujeme výrazně větší sílu F testu než Z testu.



Obrázek 4.18: Porovnání simulace silofunkce Z testu a simulace silofunkce F testu při rozsahu $n = 30$



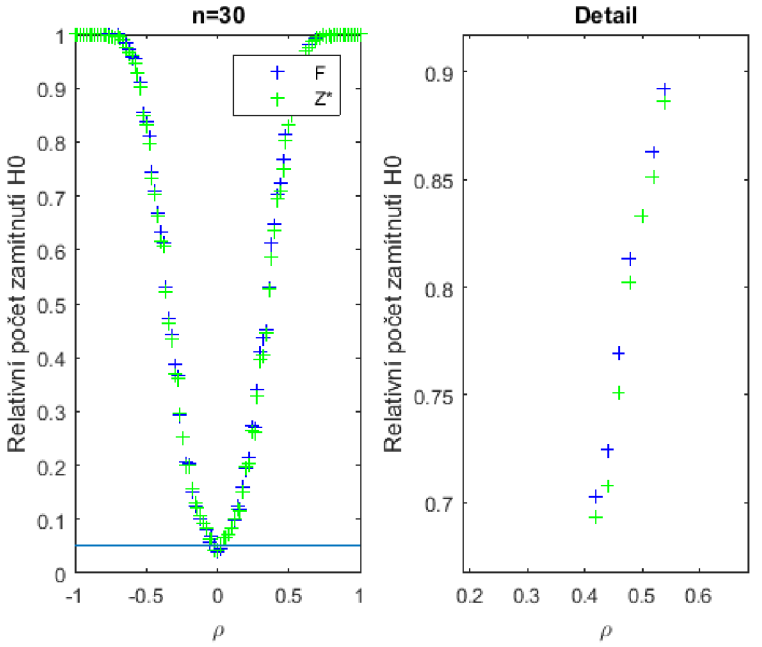
Obrázek 4.19: Porovnání simulace silofunkce Z testu a simulace silofunkce F testu při rozsahu $n = 10$



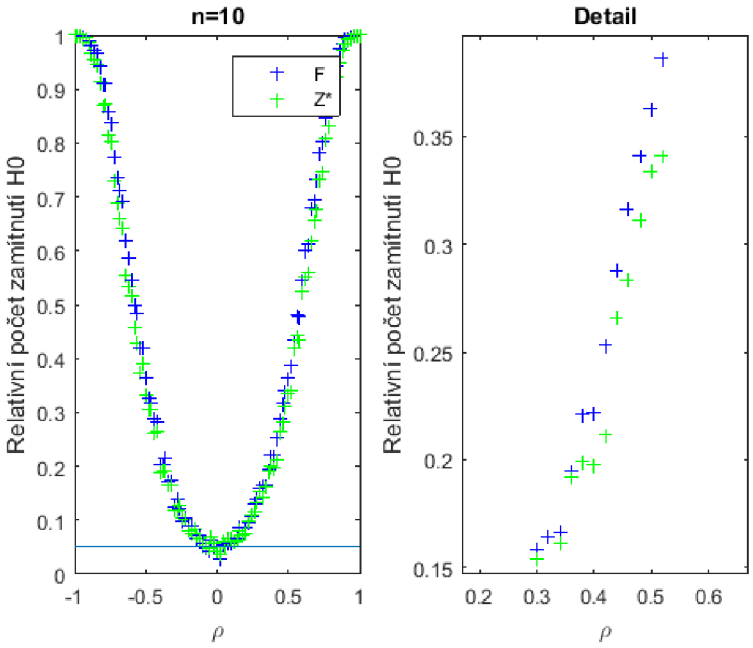
Obrázek 4.20: Porovnání simulace silofunkce Z testu a simulace silofunkce F testu při rozsahu $n = 5$

4.3.9. Porovnání testů založených na Hotellingově transformaci a Haddad-Provostově statistice

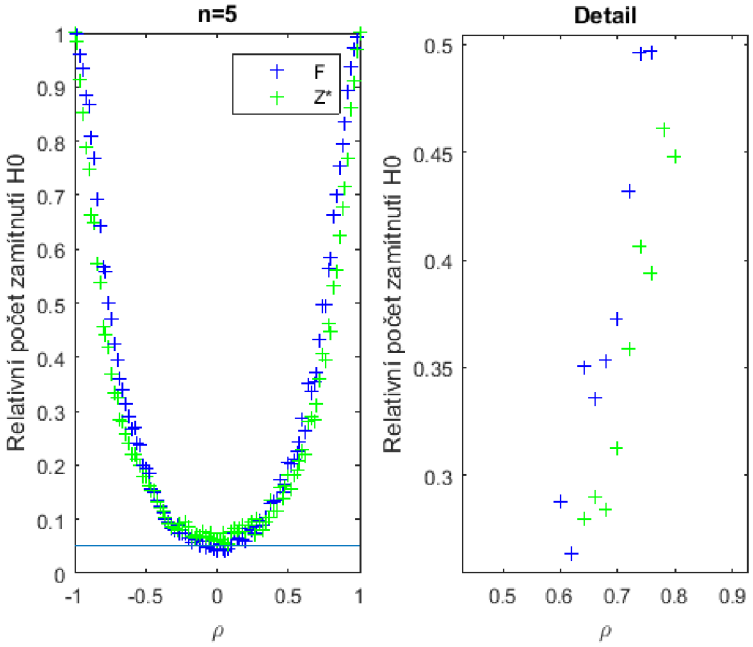
Na grafech 4.21,4.22,4.23 vidíme rozdíl v síle F testu oproti Z^* testu, který je ale výrazně menší než u porovnání F a Z testu.



Obrázek 4.21: Porovnání simulace silofunkce Z^* testu a simulace silofunkce F testu při rozsahu $n = 30$



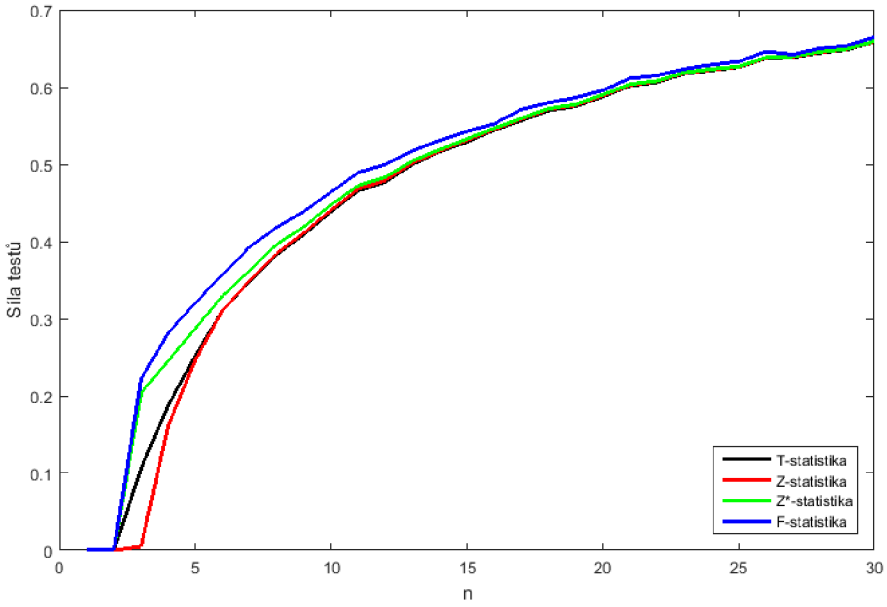
Obrázek 4.22: Porovnání simulace silofunkce Z^* testu a simulace silofunkce F testu při rozsahu $n = 10$



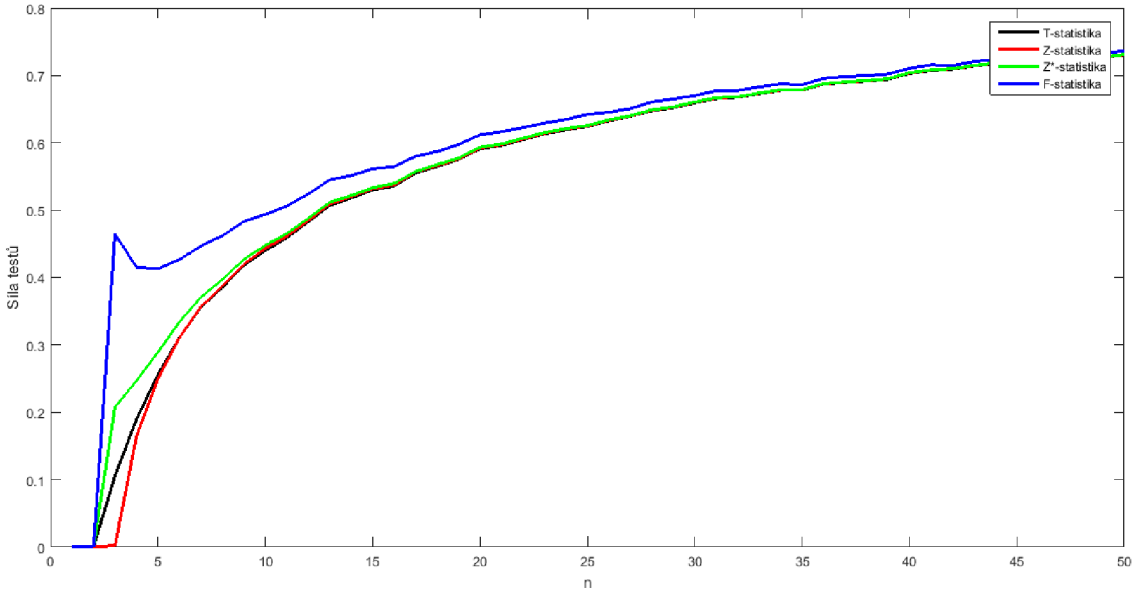
Obrázek 4.23: Porovnání simulace silofunkce Z^* testu a simulace silofunkce F testu při rozsahu $n = 5$

4.3.10. Závislost průměrné hodnoty jednotlivých silofunkcí na rozsahu náhodného výběru

Graf 4.24 vyjadřuje závislost průměrné hodnoty jednotlivých silofunkcí podle $\rho \in (-1, 1)$ na rozsahu náhodného výběru n. Často se vyskytuje efekt, kdy jedna silofunkce je silnější v místech blíže k $\rho = 0$ a jiná zase ve místech $\rho \to \pm 1$. Hlavní význam grafu 4.24 spočívá v tom, že naznačuje, pro jaké rozsahy náhodných výběrů má ještě smysl srovnávat silofunkce a relativní počet zamítnutí hypotézy H_0.
Graf 4.25 vyjadřuje závislost průměrné hodnoty jednotlivých silofunkcí na rozsahu náhodného výběru n pro různé rozptyly σ_X^2 a σ_Y^2. Ukázalo se, že závislost průměrné hodnoty silofunkcí na rozsahu n při $\sigma_X^2 \neq \sigma_Y^2$ se chová obdobně jako při $\sigma_X^2 = \sigma_Y^2$.



Obrázek 4.24: Závislost průměrné hodnoty silofunkcí na rozsahu náhodného výběru



Obrázek 4.25: Závislost průměrné hodnoty silofunkcí na rozsahu náhodného výběru z náhodných veličin s různými rozptyly

5. Závěr

Cílem mé bakalářské práce bylo porovnat testy nulovosti korelačního koeficientu dvou normálních náhodných veličin. Nejprve jsem definoval způsob provedení tohoto testu včetně nastavení potřebných parametrů. Následně jsem odvodil vzorec pro T statistiku a transformaci stabilizující rozptyl potřebnou k výpočtu Fisherovy transformace. Pomocí Fisherovy transformace jsem zavedl Z statistiku. Hotellingovou transformací jsem z ní odvodil Z^* statistiku. Nakonec jsem odvodil F statistiku a ukázal, proč má Fisher-Snedecorovo rozdělení. Následně jsem odvodil jednotlivé silofunkce testů, mimo silofunkce testu pomocí T statistiky, neboť ta by pravděpodobně závisela na daném výběru.
V praktické části práce jsem porovnal výsledné hodnoty testů nulovosti korelačního koeficientu na nasimulovaném náhodném výběru s náležitými vykreslenými silofunkcemi, ve Fisherově a Hotelingově případě jejich aproximacemi. Současně jsem silofunkce a aproximace silofunkcí porovnal mezi sebou. Ukázal jsem vlastnost přímé závislosti síly funkce na rozsahu náhodného výběru při daném ρ a nakonec jsem porovnal výsledné hodnoty náležitých testů nulovosti korelačního koeficientu na nasimulovaném náhodném výběru mezi sebou. Na závěr jsem vypracoval grafy závislosti průměrné hodnoty jednotlivých silofunkcí na rozsahu náhodného výběru, z nichž bylo vidět, že pro hodnoty $n > 30$ už jsou rozdíly testování nulovosti korelačního koeficientu různými způsoby minimální.
Výsledky potvrdily původní hypotézy, tedy zmenšování síly testu na úkor multifunkčnosti Fisherovy Z transformace a minimalizace tohoto efektu při použití Hotellingovy transformace.
Obdobné srovnání jsem provedl i pro výběr veličin s různými rozptyly a středními hodnotami. Výstupy byly natolik podobné, že je nemělo smysl prezentovat samostatně, jsou ovšem uvedeny v příloze.
Vypracování této práce mě obohatilo o mnoho zkušeností z oboru statistiky, práce v programu MATLAB a v neposlední řadě i v tvorbě v sázecím programu LaTeX. Příjemně mě překvapila úroveň přesnosti, jaké lze dosáhnout už při velmi malých rozsazích náhodného výběru. Naopak mě zarazila časová náročnost výpočtu hodnot jednotlivých grafů.

Literatura

[1] ANDĚL, Jiří. *Základy matematické statistiky.* Vyd. 1. Praha: Matfyzpress, 2005. ISBN 978-80-7378-162-0.

[2] WINTERBOTTOM, Alan. *A Note on the Derivation of Fisher's Transformation of the Correlation Coefficient.* The American Statistician, August 1979, Vol. 33, No. 3

[3] HADDAD, John N., PROVOST, Serge B. *Approximations to the Distribution of the Sample Correlation Coefficient*

[4] PROVOST, Serge B. *Closed-Form Representations of the Density Function and Integer Moments of the Sample Correlation Coefficient*

6. Seznam použitých zkratek a symbolů

σ	směrodatná odchylka
X, Y	náhodná veličina
$\boldsymbol{X}, \boldsymbol{Z}$	náhodný vektor
$\boldsymbol{X}$	matice plánu
$N(\mu, \sigma^2)$	normální rozdělení
$u_{1-\frac{\alpha}{2}}$	kvantil normálního rozdělení
$t(n)$	Studentovo rozdělení
$t_{1-\frac{\alpha}{2}}(n)$	kvantil studentova rozdělení
$F(m, n)$	Fisher-Snedecorovo rozdělení
$f_{1-\frac{\alpha}{2}}(n, n)$	kvantil Fisher-snedecorova rozdělení
$\chi^2(n)$	Chí-kvadrát rozdělení

7. Seznam příloh

CD s vloženým souborem programů softwaru MATLAB

silofunkce	Vykreslení jednotlivých silofunkcí
Graf por s sig	Graf porovnání testování výběru z veličin se stejnými rozptyly
Graf por r sig	Graf porovnání testování výběru z veličin s různými rozptyly
tabulka, jentak	Grafy na konci práce.
Silofce aprox	Porovnání silofunkce a nasimulovaných hodnot daného testu