

Univerzita Karlova v Praze  
Matematicko-fyzikální fakulta

# BAKALÁRSKA PRÁCA



Alena Straňáková

## Neymanove testy dobrej zhody

Katedra pravděpodobnosti a matematické statistiky  
Vedúci bakalárskej práce: Mgr. David Kraus

Študijný program: Matematika  
Študijný obor: Obecná matematika

2006

Moje poďakovanie patrí predovšetkým Mgr. Davidovi Krausovi. A potom taktiež Tomášovi Maradovi.

Prehlasujem, že som svoju bakalársku prácu napísala samostatne a výhradne s použitím citovaných prameňov. Súhlasím so zapožičiavaním práce a jej zverejňovaním.

V Prahe dňa 7.8.2006

Alena Straňáková

# Obsah

<b>1</b>	<b>ÚVOD</b>	<b>5</b>
<b>2</b>	<b>TESTY ZALOŽENÉ NA VIEROHODNOSTI</b>	<b>7</b>
2.1	Vierohodnosť, skórový vektor, informačná matica . . . . .	7
2.2	Testy . . . . .	8
<b>3</b>	<b>POUŽITIE NA NEYMANOVE TESTY</b>	<b>11</b>
3.1	Formulácia hypotézy a alternatívy . . . . .	11
3.2	Štatistika skórového testu . . . . .	12
3.3	Asymptotické rozdelenie štatistiky skórového testu . . . . .	14
3.4	Test na hladine $\alpha$ . . . . .	14
3.5	Voľba bázových funkcií $\varphi_1, \dots, \varphi_k$ . . . . .	15
<b>4</b>	<b>SIMULÁCIE</b>	<b>16</b>
4.1	Chovanie za platnosti nulovej hypotézy . . . . .	16
4.2	Chovanie za platnosti alternatívnej hypotézy . . . . .	17
4.3	Zhrnutie . . . . .	24

Názov práce: Neymanove testy dobrej zhody

Autor: Alena Straňáková

Katedra (ústav): Katedra pravdepodobnosti a matematické statistiky

Vedúci bakalárskej práce: Mgr. David Kraus

e-mail vedúceho: kraus@karlin.mff.cuni.cz

Abstrakt: Práca sa zaoberá testovaním dobrej zhody. Študujeme jednoduchú hypotézu, že rozdelenie dát má hustotu  $f_o$  proti alternatíve, že dáta sú z nejakého iného rozdelenia. Hlavnou témou je Neymanov prístup, ktorý spočíva v nahradení obecnej alternatívy  $k$ -rozmernou alternatívnou hypotézou popísanou hladkými funkciami.

Kľúčové slová: vierohodnosť, skórový vektor, skórová štatistika, Fisherova informačná matica, ortonormálne funkcie

Title: Neyman's Smooth Tests of Fit

Author: Alena Straňáková

Department: Department of Probability and Mathematical Statistics

Supervisor: Mgr. David Kraus

Supervisor's e-mail address: kraus@karlin.mff.cuni.cz

Abstract: This work deals with testing goodness-of-fit. We study the simple hypothesis that the density of the data have a different distribution. The main idea is Neyman's approach which is based on replacement of the general alternative with a  $k$ -dimensional alternative hypothesis described by smooth functions.

Keywords: likelihood, score vector, score statistic, Fisher information matrix, orthogonal functions

# Kapitola 1

## ÚVOD

Uvažujme náhodný výber  $X_1, \dots, X_n$  z rozdelenia s hustotou  $f$ . V tejto práci sa zaoberáme testami dobrej zhody, teda testami hypotézy, že rozdelenie dát má nejakú danú hustotu  $f_o$ .

Najznámejšie sú testy založené na empirickej distribučnej funkcii  $\hat{F}(x) = \frac{1}{n} \sum_{i=1}^n I[X_i \leq x]$ , napríklad test Kolmogorov–Smirnov s testovou štatistikou  $\sup_{x \in \mathbb{R}} |\hat{F}(x) - F_o(x)|$ . Testy založené na empirickej distribučnej funkcii sú testy proti obecnej alternatíve, že združená hustota  $f$  je iná než  $f_o$ . My sa budeme zaoberať Neymanovým prístupom k tomuto problému, ktorý spočíva v nahradení testovej úlohy

$$H_o : f = f_o \quad \text{proti} \quad A : f \neq f_o$$

úlohou

$$H_o : f = f_o \quad \text{proti} \quad H_k : f = f_k,$$

pričom  $k$ -rozmerná alternatíva je daná vzťahom

$$f_k(x, \theta) = f_o(x) \exp\{\theta^T \varphi(F_o(x)) - c_k(\theta)\},$$

kde  $\varphi = (\varphi_1, \dots, \varphi_k)^T$  je vektor nejakých bázových funkcií popisujúci rozdelenie za alternatívy. Testuje sa hypotéza  $H_o : \theta = 0$  proti  $H_1 : \theta \neq 0$ . K riešeniu tohto problému budeme používať skórový test.

Kapitola 2 podáva prehľad o hlavných výsledkoch z teórie maximálnej vierohodnosti. Obsahuje definície základných pojmov, dôležité tvrdenia a poskytuje informácie o testových štatistikách používaných v metóde maximálnej vierohodnosti.

Kapitola 3 popisuje konštrukciu Neymanových testov. Venuje sa charakterizácii skórového testu, ktorý sa pri konštrukcii používa. Na konci kapitoly sú zmienené možnosti voľby bázových funkcií  $\varphi_1, \dots, \varphi_k$ .

V kapitole 4 zrovnáme na základe simulácií Neymanove testy a test Kolmogorov-Smirnov. Výsledky testov sú uvedené v tabuľkách a rozdiely medzi testami sú znázornené graficky na obrázkoch.

## Kapitola 2

# TESTY ZALOŽENÉ NA VIEROHODNOSTI

### 2.1 Vierohodnosť, skórový vektor, informačná matica

Uvažujme náhodný výber  $\mathbf{X} = (X_1, \dots, X_n)^T$ , ktorý má hustotu  $f(x, \theta)$  vzhľadom k Lebesguovej miere  $\mu$ , pričom  $\theta = (\theta_1, \dots, \theta_k)^T \in \mathbb{R}^k$  je  $k$ -rozmerný parameter. Najprv si objasníme teóriu maximálne vierohodných odhadov. Budeme odhadovať celý parameter  $\theta \in \Omega$  na základe náhodného vektoru  $\mathbf{X}$ .

*Vierohodnostnú funkciu* definujeme vzťahom

$$L_k(\theta) = \prod_{i=1}^n f(X_i, \theta) \quad (2.1)$$

ako funkciu  $\theta$  pri pevnom  $X_i$ .

Metóda maximálnej vierohodnosti používa *maximálne vierohodný odhad* parametru  $\theta$ . Je to hodnota  $\hat{\theta}$  parametru  $\theta$ , ktorá maximalizuje vierohodnostnú funkciu  $L_k(\theta)$ , teda platí

$$\max_{\theta \in \Omega} \prod_{i=1}^n f(X_i, \theta) = \prod_{i=1}^n f(X_i, \hat{\theta}) \quad (2.2)$$

Vektor  $\hat{\theta}_k \in \Omega$  je maximálne vierohodným odhadom parametru  $\theta$ , práve keď pre každé  $\theta \in \Omega$  platí  $L_k(\hat{\theta}_k) \geq L_k(\theta)$ .

Definujme  $l_k(\theta) = \log L_k(\theta)$ . Funkciu  $l_k(\theta)$  nazývame *logaritmickej vierohodnostná funkcia*.

Za predpokladu, že  $f(x, \theta)$  je dostatočne hladká funkcia  $\theta$  v nejakom okolí skutočnej hodnoty parametru  $\theta = \theta_o$ , môžeme definovať:

skórový vektor ako vektor parciálnych derivácií logaritmickej vierohodnostnej funkcie podľa parametru  $\theta$ :

$$U_k(\theta) = \frac{\partial}{\partial \theta} l_k(\theta) = \sum_{i=1}^n \frac{\frac{\partial f(X_i, \theta)}{\partial \theta}}{f(X_i, \theta)} \quad (2.3)$$

Fisherovu informačnú maticu:

$$I_k(\theta) = (I_{ij}(\theta))_{i,j=1}^k \quad (2.4)$$

pričom

$$I_{ij}(\theta) = \int_M \frac{f'_i(x, \theta) f'_j(x, \theta)}{f^2(x, \theta)} f(x, \theta) d\mu(x) = E \left[ \frac{f'_i(X, \theta) f'_j(X, \theta)}{f^2(X, \theta)} \right], \quad (2.5)$$

kde sa používa značenie  $f'_i = \frac{\partial f}{\partial \theta_i}(X_i, \theta)$ .

Predpokladajme, že Fisherova informačná matica existuje a je pozitívne definitná na okolí  $\theta_o$ . Ďalej budeme pracovať s upravenou definíciou maximálne vierohodného odhadu:

Vektor  $\hat{\theta}_k \in \Omega$  sa nazýva maximálne vierohodný odhad parametru  $\theta$ , práve keď rieši vierohodnostnú rovnicu  $U_k(\hat{\theta}_k) = 0$ .

## 2.2 Testy

Uvažujme jednoduchú hypotézu  $H_o : \theta = \theta_o$  proti alternatíve  $H_1 : \theta \neq \theta_o$ . Zavedieme tri testové štatistiky pre testovanie  $H_o$  proti  $H_1$ :

1. štatistika

$$\lambda_k = 2[l(\hat{\theta}_k) - l(\theta_o)] \quad (2.6)$$

sa nazýva vierohodnostný pomer.

2. štatistika

$$W_k = n(\hat{\theta}_k - \theta_o)^T I_k(\hat{\theta}_k)(\hat{\theta}_k - \theta_o) \quad (2.7)$$

sa nazýva Waldova štatistika.

3. štatistika

$$T_k = \frac{1}{n} [U_k(\theta_o)]^T [I_k(\theta_o)]^{-1} U_k(\theta_o) \quad (2.8)$$

sa nazýva Raova (skórová) štatistika.



Matica  $I_k$ , ktorá sa vyskytuje v definícii Raovej a Waldovej štatistiky, predstavuje buď priamo Fisherovu informačnú maticu alebo jej akýkoľvek konzistentný odhad, viac v článku [2], str. 127.

Veľké hodnoty uvedených testových štatistík svedčia proti  $H_0$  v prospech  $H_1$ . Aby sme mohli posúdiť, kedy je hodnota štatistiky významne veľká, potrebujeme poznať rozdelenie štatistiky. Použijeme asymptotické rozdelenie. K tomu nám pomôže nasledujúca veta ([1], veta 7.100).

**Veta 2.1.** *Nech systém hustôt  $\{f(x, \theta), \theta \in \Omega\}$  je regulárny a má Fisherovu maticu  $I(\theta)$ . Nech platia predpoklady:*

1. *Nech  $\Omega \subset \mathbb{R}^k$  je parametrický priestor, ktorý obsahuje taký neprázdny otvorený interval  $\omega$ , že skutočná hodnota parametru  $\theta_0$  patrí do  $\omega$ .*
2. *Nech  $\mathbf{X} = (X_1, \dots, X_n)^T$ , kde  $X_i$  sú nezávislé rovnako rozdelené nahodné veličiny s hustotou  $f(x, \theta)$  vzhľadom k nejakej  $\sigma$ -konečnej miere  $\mu$ .*
3. *Nech  $M = \{x : f(x, \theta) > 0\}$  nezávisí na  $\theta$ .*
4. *Nech  $\theta_1, \theta_2 \in \Omega$ . Potom  $f(x, \theta_1) = f(x, \theta_2)$  skoro všade až na množinu miery  $\mu$  platí práve vtedy, ak je  $\theta_1 = \theta_2$ .*

*Nech sú ďalej splnené nasledujúce predpoklady:*

(a) *Derivácia  $\frac{\partial^3 f(x, \theta)}{\partial \theta_i \partial \theta_j \partial \theta_l}$  existuje pre skoro všetky  $x$ , pre všetky  $\theta \in \omega$  a pre všetky  $i, j, l = 1, \dots, k$ .*

(b) *Pre všetky  $\theta \in \omega$  platí*

$$\int_M f''_{ij}(x, \theta) d\mu(x) = 0, \quad i, j = 1, \dots, k. \quad (2.9)$$

(c) *Pre všetky  $i, j, l = 1, \dots, k$  existujú funkcie  $M_{ijl}(x) \geq 0$  tak, že*

$$E_{\theta_0} M_{ijl}(X) < \infty \quad (2.10)$$

*a*

$$\left| \frac{\partial^3 \ln f(x, \theta)}{\partial \theta_i \partial \theta_j \partial \theta_l} \right| \leq M_{ijl}(x) \quad (2.11)$$

*pre všetky  $\theta \in \omega$  a skoro všetky  $x \in M$ .*

Potom platia nasledujúce tvrdenia.

- (i) Ak  $n \rightarrow \infty$ , potom ku každému  $\varepsilon > 0$  existuje s pravdepodobnosťou blížiacou sa jednej také riešenie  $\hat{\theta}_k$  systému vierohodnostných rovníc, že  $\|\hat{\theta}_k - \theta_o\| < \varepsilon$ .
- (ii) Ak  $n \rightarrow \infty$ , potom pre skórový vektor (2.3) platí

$$\frac{1}{\sqrt{n}}U_k(\theta_o) \xrightarrow{d} N[0, I_k(\theta_o)] \quad (2.12)$$

- (iii) Ak existuje pre každé dostatočne veľké  $n$  a pre každú hodnotu  $\mathbf{X}$  taký koreň  $\hat{\theta}_k$  systému vierohodnostných rovníc, že  $\hat{\theta}_k$  je konzistentným odhadom parametru  $\theta_o$ , potom

$$\sqrt{n}(\hat{\theta}_k - \theta_o) \xrightarrow{d} N(0, [I_k(\theta_o)]^{-1}) \quad (2.13)$$

Všetky uvedené štatistiky majú asymptoticky rozdelenie  $\chi^2$  o  $k$  stupňoch voľnosti. Pre Raovu (skórovú) štatistiku toto okamžite plynie z bodu (ii) predchádzajúcej vety. Pre ostatné dve štatistiky viď [1], kapitola 8.6.1.

# Kapitola 3

## POUŽITIE NA NEYMANOVE TESTY

### 3.1 Formulácia hypotézy a alternatív

Nech  $X_1, \dots, X_n$  je výber zo spojitého rozdelenia definovaného hustotou  $f(x)$ . Predpokladajme, že existujú dve konkurujúce si hypotézy:  $H_o : f = f_o$  proti alternatíve  $A : f \neq f_o$ . V Neymanovom teste sa miesto toho testuje  $H_o : f = f_o$  proti  $k$ -rozmernej alternatívnej hypotéze  $H_k : f = f_k$ , kde

$$f_k(x, \theta) = f_o(x) \exp\{\theta^T \varphi(F_o(x)) - c_k(\theta)\}, \quad (3.1)$$

$k \geq 1$ ,  $k$  pevné,  $x \in [0, 1]$ ,  $F_o$  je distribučná funkcia odpovedajúca hustote  $f_o$  a  $\varphi = (\varphi_1, \dots, \varphi_k)^T$  je  $k$ -rozmerný vektor nejakých lineárne nezávislých funkcií v  $L^2(0, 1)$  vyjadrujúcich odlišnosť rozdelenia za alternatívy od hypotetického rozdelenia s hustotou  $f_o$ . Normujúcu konštantu  $c_k(\theta)$  môžeme vyjadriť ako

$$c_k(\theta) = \log \left[ \int_{-\infty}^{\infty} \exp\{\theta^T \varphi(F_o(x))\} f_o(x) dx \right]. \quad (3.2)$$

V Neymanovom teste dobrej zhody testovanie  $H_o : f = f_o$  proti  $H_k : f = f_k$  odpovedá testovaniu nulovosti parametru  $\theta$ , teda chceme testovať hypotézu  $H_o : \theta = 0$  proti  $H_k : \theta \neq 0$ . Môžeme použiť niektorú z troch uvedených testových štatistík (Rao, Wald, vierohodnostný pomer). Vzhľadom k tomu, že naším cieľom nie je odhadnúť parameter  $\theta$ , ale len o ňom testovať hypotézu, použijeme skórový test, ktorý nevyžaduje hľadanie maximálne vierohodného odhadu.

Poznamenajme, že test  $H_o : X \sim F_o$  je ekvivalentný testovaniu rovnomernosti na  $(0,1)$ . Transformáciou  $Y = F_o(X)$  získame náhodný výber  $Y_1, \dots, Y_n$  z rozdelenia  $R(0,1)$ . Takže platí, ak  $X \sim F_o$ , potom  $F_o(X) \sim R(0,1)$ . Testovanie hypotézy  $H_o : X \sim F_o$  je preto pre akúkoľvek  $F_o$  ekvivalentné testovaniu rovnomernosti transformovaných veličín, teda  $H_o : F_o(X) \sim R(0,1)$ .

## 3.2 Štatistika skórového testu

V našom prípade, teda ak platí  $H_o : \theta = 0$ , štatistika má tvar

$$T_k = \frac{1}{n} U_k(0)^T I_k(0)^{-1} U_k(0) \quad (3.3)$$

Vyjadruje, nakoľko je 0 dobrým odhadom  $\theta$ , teda nakoľko 0 anuluje skórový vektor.

Spočítajme jednotlivé členy. Najprv upravíme substitúciou  $t = F_o(x)$  ( $dt = f_o(x)dx$ ) normujúcu konštantu  $c_k(\theta)$  (3.2) na tvar

$$c_k(\theta) = \log \left[ \int_0^1 \exp\{\theta^T \varphi(t)\} dt \right] \quad (3.4)$$

Ďalej môžeme logaritmicke vierohodnostnú funkciu definovanú v prvej kapitole vyjadriť ako :

$$l_k(\theta) = \log L_k(\theta) = \sum_{i=1}^n \log f_k(X_i, \theta) = \sum_{i=1}^n \log f_o(X_i) + \sum_{i=1}^n \theta^T \varphi(F_o(X_i)) - n c_k(\theta) \quad (3.5)$$

Po dosadení derivácie logaritmu hustoty

$$\frac{\partial}{\partial \theta} \log f_k(x, \theta) = \varphi(F_o(x)) - \frac{\int_0^1 \varphi(t) \exp\{\theta^T \varphi(t)\} dt}{\int_0^1 \exp\{\theta^T \varphi(t)\} dt} \quad (3.6)$$

do vzťahu pre skórový vektor (2.3) dostávame:

$$U_k(\theta) = \sum_{i=1}^n \varphi(F_o(X_i)) - n \frac{\int_0^1 \varphi(t) \exp\{\theta^T \varphi(t)\} dt}{\int_0^1 \exp\{\theta^T \varphi(t)\} dt}, \quad (3.7)$$

pritom sa tento vzťah (3.7) za nulovej hypotézy ( $\theta = 0$ ) dá napísať v tvare:

$$U_k(0) = \sum_{i=1}^n \varphi(F_o(X_i)) - n \int_0^1 \varphi(t) dt \quad (3.8)$$

Vo vyjadrení Fisherovej informačnej matice (2.4) upravíme tvar *teoretickej informačnej matice*  $I_{ij}(\theta)$ . Vzorec pre  $I_{ij}(\theta)$  môžeme chápať ako integrál funkcie  $(f'/f)^2$  podľa miery, ktorá má hustotu  $f$  vzhľadom k Lebesguovej miere  $\mu$ . Potom na doplnku množiny  $M$  sa dá výraz  $(f'/f)^2$  definovať ľubovoľne a za integračný priestor vziať celý priestor  $\mathbb{R}^k$ . Tým sa hodnota  $I_{ij}(\theta)$  nezmení. Z vety o prenose integrácie ([1], str.14) teda plynie, že platí:

$$\begin{aligned} I_{ij}(\theta) &= E \left[ \frac{f'_i(X, \theta) f'_j(X, \theta)}{f^2(X, \theta)} \right] = E \left[ \frac{\partial \log f(X, \theta)}{\partial \theta_i} \frac{\partial \log f(X, \theta)}{\partial \theta_j} \right] = \\ &= E \left[ \left( \varphi_i(F_o(X)) - \frac{\int_0^1 \varphi_i(t) \exp\{\theta^T \varphi(t)\} dt}{\int_0^1 \exp\{\theta^T \varphi(t)\} dt} \right) \left( \varphi_j(F_o(X)) - \frac{\int_0^1 \varphi_j(t) \exp\{\theta^T \varphi(t)\} dt}{\int_0^1 \exp\{\theta^T \varphi(t)\} dt} \right) \right] \end{aligned} \quad (3.9)$$

Za platnosti  $H_o$  je

$$\begin{aligned} I_{ij}(0) &= E \left[ \left( \varphi_i(F_o(X)) - \int_0^1 \varphi_i(t) dt \right) \left( \varphi_j(F_o(X)) - \int_0^1 \varphi_j(t) dt \right) \right] = \\ &= \int_0^1 \varphi_i(t) \varphi_j(t) dt - \int_0^1 \varphi_i(t) dt \int_0^1 \varphi_j(t) dt \end{aligned} \quad (3.10)$$

V poslednom výraze sme znovu použili substitúciu  $t = F_o(x)$ .

Z vyjadrenia  $I_{ij}(0)$  vidíme, že matica sa zjednoduší, ak funkcie  $\varphi_1, \dots, \varphi_k$  budú ortonormálne a centrovane ( $\{1, \varphi_1, \dots, \varphi_k\}$  sú ortonormálne), teda spĺňajú

$$\int_0^1 \varphi_i(t) \varphi_j(t) dt = 0 \quad \text{pre } i \neq j \quad (3.11)$$

$$\int_0^1 \varphi_i(t)^2 dt = 1 \quad (3.12)$$

$$\int_0^1 \varphi_i(t) dt = 0 \quad (3.13)$$

potom matica  $I_k(0)$  je jednotková a testová štatistika  $T_k$  má jednoduchý tvar:

$$T_k = \frac{1}{n} \sum_{j=1}^k \left[ \sum_{i=1}^n \varphi_j(F_o(X_i)) \right]^2 \quad (3.14)$$

Ak funkcie  $\varphi_1, \dots, \varphi_k$  nie sú ortogonálne, matica  $I_k(0)$  nie je diagonálna a jej prvky musíme spočítať

$$I_k(0) = \int_0^1 \varphi(t)[\varphi(t)]^T dt \quad (3.15)$$

alebo použiť *výberovú informačnú maticu*, ktoru definujeme ako odhad asymptotickej variančnej matice skórového vektoru  $U_k(0)$ :

$$I_k(0) = \frac{1}{n} \sum_{i=1}^n \varphi(F_o(X_i))\varphi(F_o(X_i))^T. \quad (3.16)$$

### 3.3 Asymptotické rozdelenie štatistiky skórového testu

Pri aplikácii vety 2.1. na Neymanove testy sú všetky predpoklady vety splnené, predovšetkým 3. predpoklad, pretože nosič  $f_o(x)$  nezávisí na parametri  $\theta$  a exponenciálna funkcia je vždy kladná, takže množina  $M$  nezávisí na  $\theta$ . Z výsledkov uvedených v kapitole 2 teda plynie:

$$\frac{1}{\sqrt{n}}U_k(0) \xrightarrow{n \rightarrow \infty} N(0, I_k(0)). \quad (3.17)$$

Z toho odvodíme asymptotické rozdelenie  $T_k$ :

$$\frac{1}{n}U_k(0)^T I_k(0)^{-1}U_k(0) \xrightarrow{n \rightarrow \infty} \chi_k^2 \quad (3.18)$$

Toto tvrdenie zostáva v platnosti aj keď použijeme nejaký konzistentný odhad matice  $I_k(0)$ .

### 3.4 Test na hladine $\alpha$

Test hypotézy  $H_o : \theta = 0$  proti alternatíve  $H_k : \theta \neq 0$  na hladine  $\alpha$  založený na  $T_k(0)$  zamietá nulovú hypotézu v prípade, že  $T_k > b_{(1-\alpha)}$ , kde  $b_{(1-\alpha)}$  je  $(1 - \alpha)$  kvantil  $\chi_k^2$  rozdelenia.  $P$ -hodnota uvedeného testu je  $p = 1 - G_k(T_k)$ , kde  $G_k$  je distribučná funkcia rozdelenia  $\chi_k^2$ .

### 3.5 Voľba bázových funkcií $\varphi_1, \dots, \varphi_k$

Bázové funkcie  $\varphi_1, \dots, \varphi_k \in L^2(0, 1)$  predstavujú odlišnosť hustoty za alternatívy od hypotetickej hustoty  $f_o(x)$ .

Môžu to byť

1. ortonormálne funkcie (informačná matica je jednotková), tu ako bázu volíme napríklad

- cos:  $\varphi_j(t) = \sqrt{2} \cos(j\pi t) \quad t \in [0, 1]$
- Legendreove polynómy na  $[0, 1]$  definované ako

$$\varphi_j(t) = \sqrt{2j+1} \sum_{l=0}^j \binom{j}{l} \binom{j+l}{l} (-t)^l = \sqrt{2j+1} \frac{1}{j!} \frac{d^j}{dt^j} [(t^2 - t)^j] \quad (3.19)$$

Niekoľko prvých polynómov má tvar

$$\varphi_1(t) = \sqrt{3}(2t - 1) \quad (3.20)$$

$$\varphi_2(t) = \sqrt{5}(6t^2 - 6t + 1) \quad (3.21)$$

$$\varphi_3(t) = \sqrt{7}(20t^3 - 30t^2 + 12t - 1) \quad (3.22)$$

2. neortogonálne funkcie (informačná matica nie je diagonálna), sú to napríklad indikátory intervalov:  $\varphi_j(t) = I_{(t_j, t_{j+1}]}(t)$ .

# Kapitola 4

## SIMULÁCIE

### 4.1 Chovanie za platnosti nulovej hypotézy

V tejto časti nás bude zaujímať, či Neymanov test dodržiava pri nagenovaných dátach predpísanú hladinu  $\alpha = 5\%$ . Generujeme dáta s rovnomerným rozdelením a testujeme, či dáta majú toto rozdelenie. Pravdepodobnosť zamietnutia hypotézy  $H_0$  by mala byť blízka hladine  $\alpha$ . Pre porovnanie uvedieme aj test Kolmogorov–Smirnov.

Testujeme vygenerovaný náhodný výber o rozsahu buď  $n = 50$  alebo  $n = 100$  pri počte opakovaní  $M = 5000$ , zvolili sme cosinovú bázu a počet bazových funkcií  $\varphi$  je  $k$ . Smerodatná odchylka nasimulovaných čísel je  $\sqrt{\frac{0.05(1-0.05)}{M}}$ . V našom prípade sa rovná približne 0.0031. Podľa výsledkov v tabuľke 4.1. vidíme, že oba testy dobre dodržiavajú hladinu  $\alpha$ .



Tabuľka 4.1: Odhadnutie hladiny Neymanovho testu a testu Kolmogorov–Smirnov pri nominálnej hladine  $\alpha = 0,05$ .

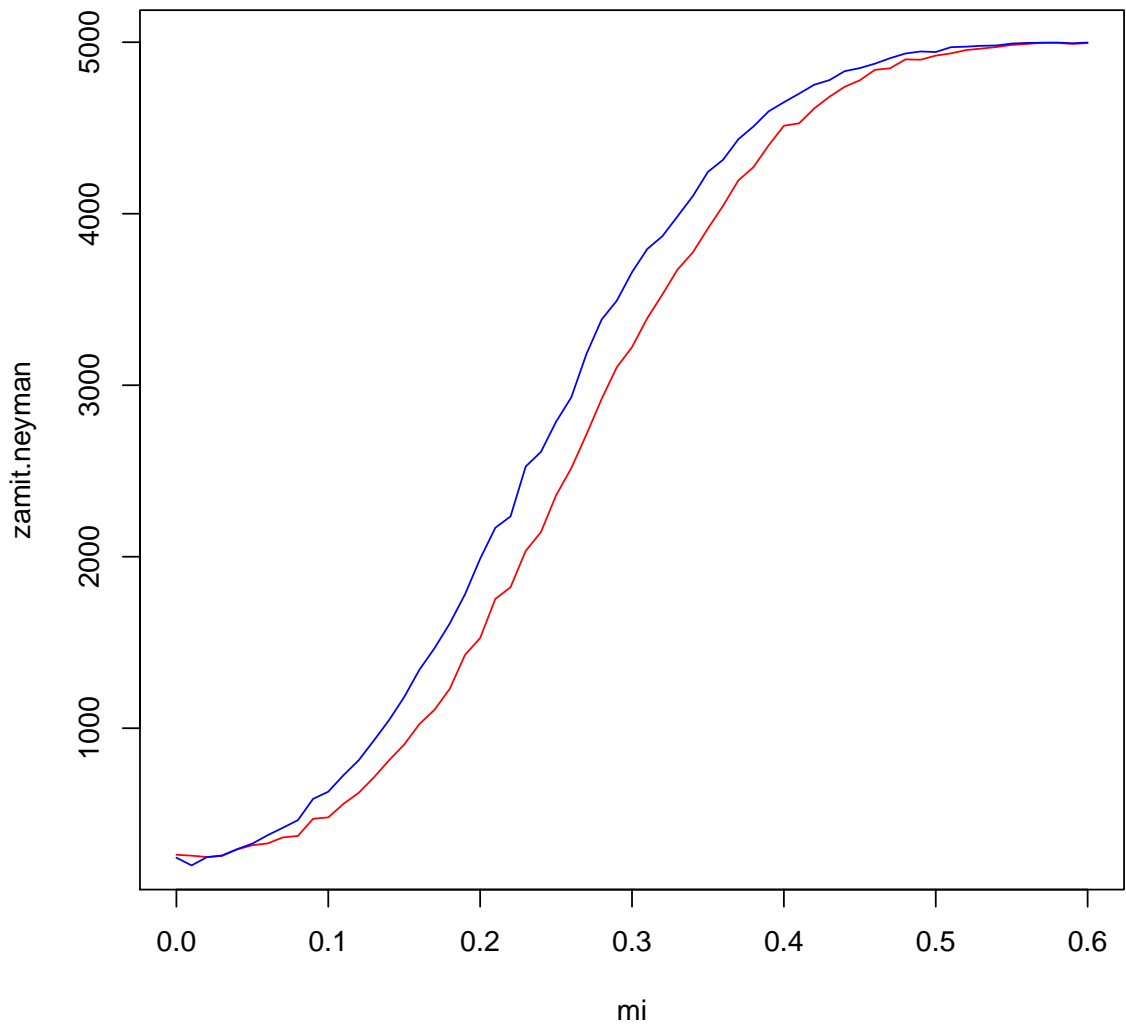
Neyman	$R(0,1)$	
	$n = 50$	$n = 100$
$k = 3$	0,0488	0,0486
$k = 4$	0,0490	0,0530
$k = 5$	0,0466	0,0500
$k = 6$	0,0546	0,0462
Kolmogorov–Smirnov	0,0380	0,0446

## 4.2 Chovanie za platnosti alternatívnej hypotézy

Budeme sledovať akú silu má Neymanov test na predpísanej hladine  $\alpha = 5\%$ . Teda aká je pravdepodobnosť zamietnutia nulovej hypotézy. Generujeme dáta z nejakého rozdelenia a testujeme, či majú tieto dáta nejaké iné rozdelenie. Takže pravdepodobnosť zamietnutia by mala byť veľká, mala by sa blížiť 1. Pre porovnanie uvedieme aj test Kolmogorov–Smirnov a sledujeme, ktorý test je v ktorých situáciách silný. Budeme generovať náhodný výber o rozsahu  $n = 100$  pri počte opakovaní  $M = 5000$  a pri testovaní použijeme cosinovú bázu, kde  $k = 4$  je počet báзовých funkcií  $\varphi$ . Smerodatná odchylka nasimulovaných čísel je najviac  $\sqrt{\frac{0,5(1-0,5)}{M}}$ . V tomto prípade je približne rovná 0,0071.

Generujeme výbery z rozdelenia  $N(\mu, 1)$  pre  $\mu$  od 0 do 0,6 a sledujeme, aká je pravdepodobnosť zamietnutia hypotézy, že tieto výbery sú z rozdelenia  $N(0, 1)$ . Z výsledkov v tabuľke 4.2 zisťujeme, že test Kolmogorov–Smirnov je trochu lepší než Neymanov. Podľa grafov na obrázku 4.1 sa dá vidieť, že ak sa rozdelenie líši strednou hodnotou, ale rozptyl zostáva rovnaký, potom testy sú zrovnateľné, ale trochu lepší je test Kolmogorov–Smirnov.

Obrázok 4.1: Silofunkcia Neymanovho testu (červenou farbou) a testu Kolmogorov–Smirnov (modrou farbou) pri výbere z  $N(\mu, 1)$ .



Tabuľka 4.2: *Odhadnutie pravdepodobnosti zamietnutia hypotézy  $N(0, 1)$  pri skutočnom rozdelení  $N(\mu, 1)$  pri hladine testu  $\alpha = 0,05$ .*

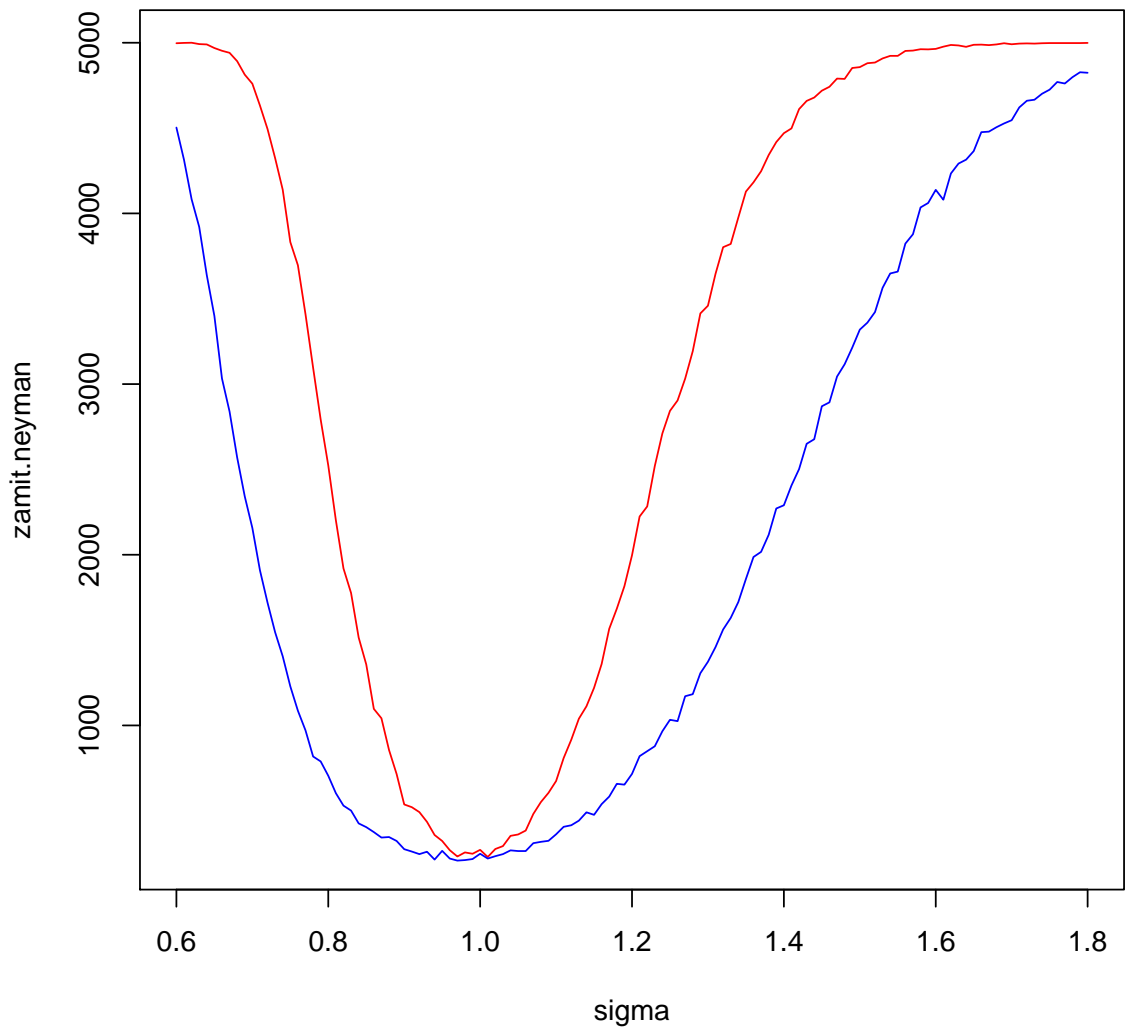
$\mu$	0	0,1	0,2	0,3	0,4	0,5	0,6
Neyman	0,0522	0,1050	0,3044	0,6452	0,8996	0,9846	0,9996
K-S	0,0490	0,1238	0,3898	0,7280	0,9300	0,9882	0,9996

Nasleduje test hypotézy, či náhodný výber z rozdelenia  $N(0, \sigma^2)$  pre rôzne hodnoty  $\sigma$  je z rozdelenia  $N(0, 1)$ . Podľa výsledkov v tabuľke 4.3 vidíme, že test Komogorov–Smirnov je slabší ako Neymanov. Z grafov na obrázku 4.2 zisťujeme, že keď sa rozdelenie líši rozptylom, ale stredná hodnota je rovnaká, potom Neymanov test je výrazne lepší než test Kolmogorov–Smirnov.

Tabuľka 4.3: *Odhadnutá pravdepodobnosť zamietnutia hypotézy  $N(0, 1)$  pri skutočnom rozdelení  $N(0, \sigma^2)$  pri nominálnej hladine  $\alpha = 0,05$ .*

$\sigma$	0,6	0,8	1	1,2	1,4	1,6	1,8	2
Neyman	0,9980	0,5114	0,0522	0,3968	0,8978	0,9948	0,9994	1,0000
K-S	0,8964	0,1368	0,0490	0,1334	0,4640	0,8166	0,9712	0,9972

Obrázok 4.2: Graf silofunkcie Neymanovho testu (červenou farbou) a testu Kolmogorov–Smirnov (modrou farbou) pri výbere z  $N(0, \sigma^2)$ .

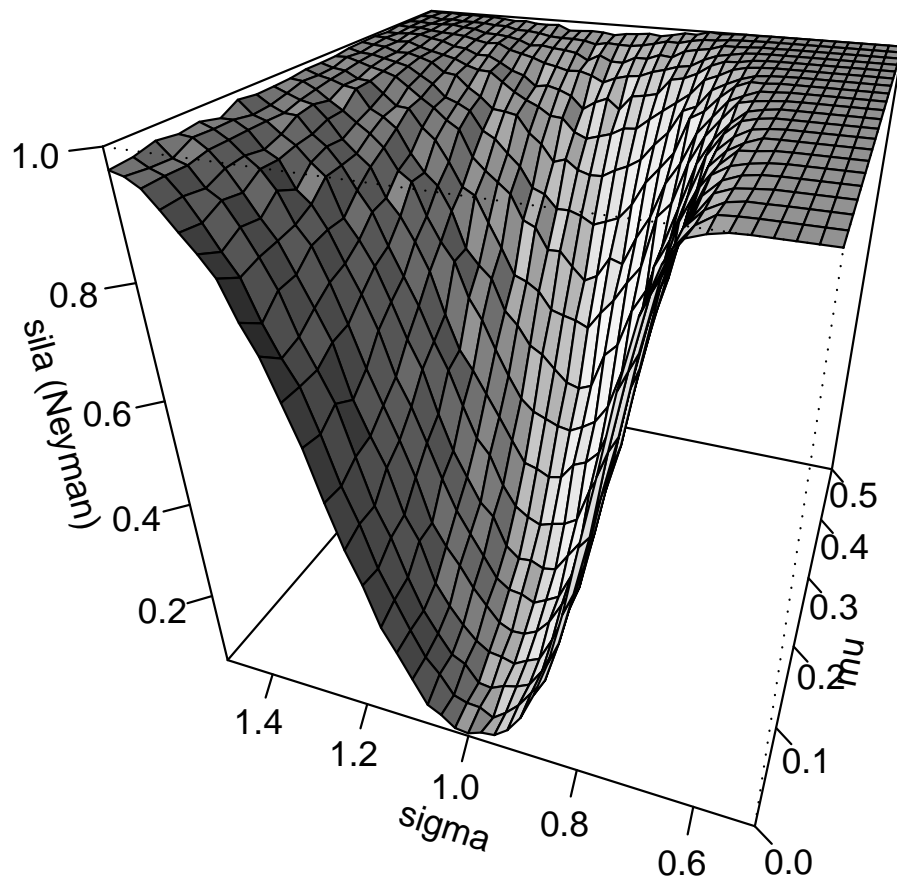


Tabuľka 4.4: Pravdepodobnosť zamietnutia hypotézy  $N(0, 1)$  Neymanovým testom pri výbere z  $N(\mu, \sigma^2)$ .

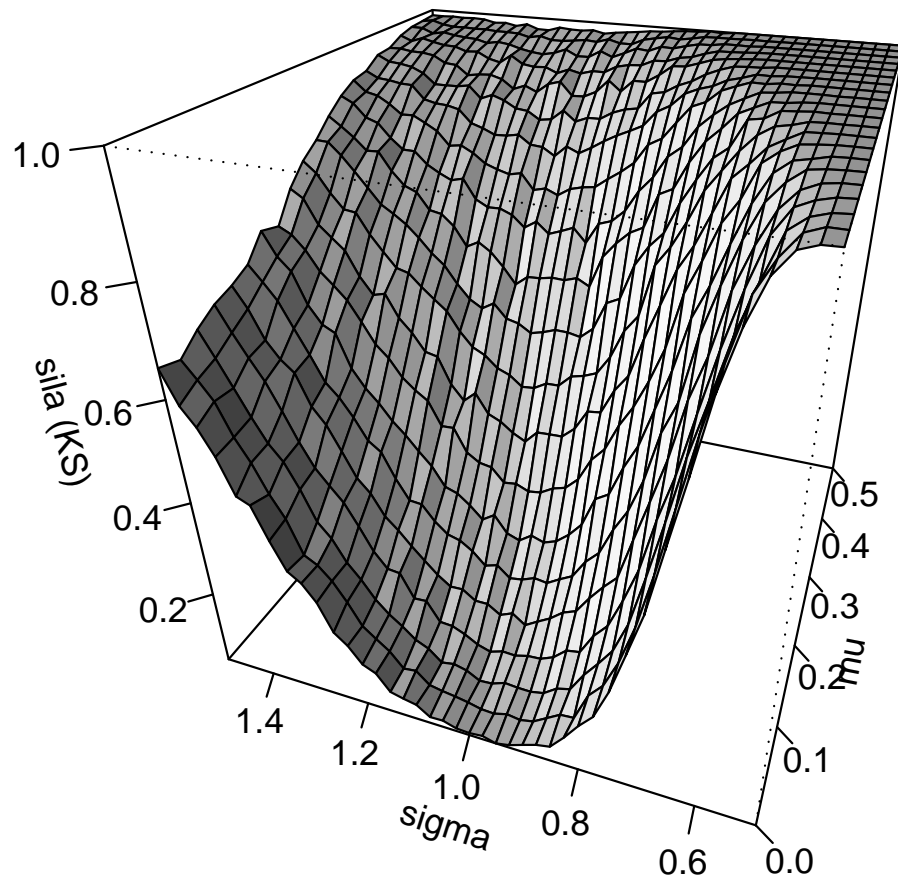
Neyman	$\sigma=0,6$	$\sigma=0,8$	$\sigma=1$	$\sigma=1,2$	$\sigma=1,4$	$\sigma=1,6$	$\sigma=1,8$
$\mu=0$	0,9998	0,5116	0,0488	0,4052	0,8934	0,9944	1,0000
$\mu=0,1$	0,9998	0,5710	0,1092	0,4520	0,9154	0,9920	0,9998
$\mu=0,2$	1,0000	0,7392	0,3056	0,6278	0,9412	0,9964	1,0000
$\mu=0,3$	1,0000	0,9166	0,6428	0,8056	0,9702	0,9974	0,9998
$\mu=0,4$	1,0000	0,9904	0,8932	0,9408	0,9910	0,9998	1,0000
$\mu=0,5$	1,0000	0,9996	0,9836	0,9860	0,9968	1,0000	1,0000
$\mu=0,6$	1,0000	1,0000	0,9996	0,9972	0,9996	0,9990	1,0000

Ďalej generujeme výbery z rozdelenia  $N(\mu, \sigma^2)$  pre rôzne hodnoty  $\mu$  a  $\sigma$  a v nasledujúcich tabuľkách 4.4 a 4.5 a obrázkoch 4.3 a 4.4 pozorujeme pravdepodobnosť zamietnutia hypotézy, že tieto výbery sú z rozdelenia  $N(0, 1)$ . Z grafov zisťujeme väčšiu citlivosť Neymanovho testu na zmenu rozptylu.

Obrázok 4.3: Silofunkcia Neymanovho testu pri testovaní hypotézy  $H_0 : X \sim N(0, 1)$ , keď skutočné rozdelenie je  $N(\mu, \sigma^2)$ .



Obrázok 4.4: Silofunkcia testu Kolmogorov–Smirnov pri testovaní hypotézy  $H_0 : X \sim N(0, 1)$ , keď skutočné rozdelenie je  $N(\mu, \sigma^2)$ .



Tabuľka 4.5: Pravdepodobnosť zamietnutia hypotézy  $N(0, 1)$  testom Kolmogorov–Smirnov pri výbere z  $N(\mu, \sigma^2)$ .

K–S	$\sigma=0,6$	$\sigma=0,8$	$\sigma=1$	$\sigma=1,2$	$\sigma=1,4$	$\sigma=1,6$	$\sigma=1,8$
$\mu=0$	0,8954	0,1422	0,0480	0,1456	0,4650	0,8152	0,9702
$\mu=0,1$	0,9596	0,3110	0,1350	0,2242	0,5344	0,8352	0,9734
$\mu=0,2$	0,9978	0,6816	0,3886	0,4798	0,6956	0,8946	0,9802
$\mu=0,3$	1,0000	0,9292	0,7250	0,7350	0,8532	0,9526	0,9876
$\mu=0,4$	1,0000	0,9960	0,9274	0,9112	0,9550	0,9832	0,9950
$\mu=0,5$	1,0000	0,9998	0,9904	0,9816	0,9862	0,9932	0,9988
$\mu=0,6$	1,0000	1,0000	1,0000	0,9962	0,9978	0,9986	0,9998

Nech  $a > 0$ ,  $b > 0$  sú dané čísla. Definujme beta rozdelenie  $B(a, b)$  s hustotou

$$f(x) = \begin{cases} \frac{1}{B(a,b)} x^{a-1} (1-x)^{b-1}, & 0 < x < 1 \\ 0, & \text{ináč} \end{cases} \quad (4.1)$$

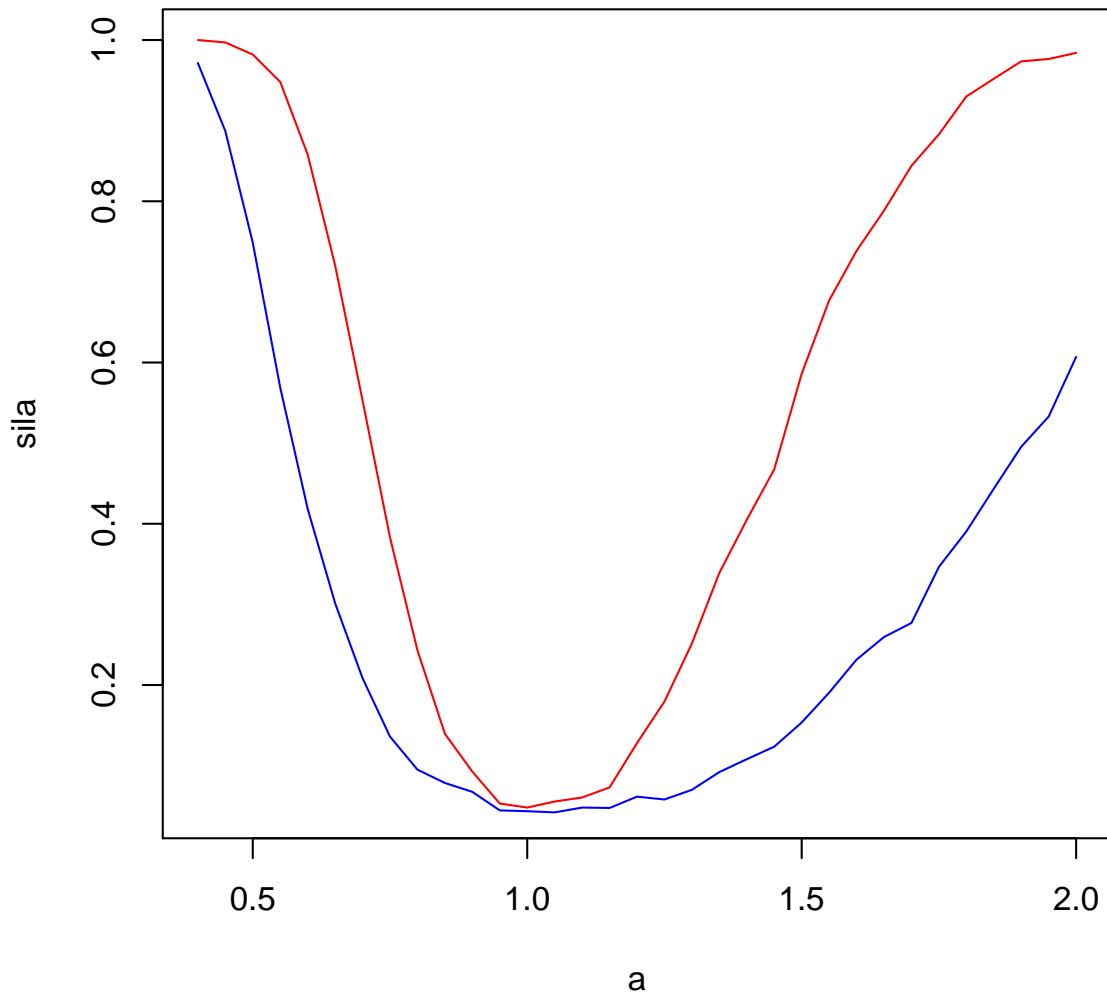
Treba poznamenať, že špeciálnym prípadom beta rozdelenia pri  $a = b = 1$  je rovnomerné rozdelenie  $R(0, 1)$ . Preto zaujímavé výsledky podáva test hypotézy, či náhodný výber z beta rozdelenia  $B(a, b)$  má rovnomerné rozdelenie  $R(0, 1)$ . Parametre  $a$  a  $b$  berieme z rozsahu 0,4 až 2. Z grafov na obrázkoch 4.5, 4.6 a 4.7 sa dá vypozerovať, že na diagonále je sila testu Kolmogorov–Smirnov ďaleko menšia než sila Neymanovho testu. Je to preto, že na diagonále je beta rozdelenie symetrické okolo  $\frac{1}{2}$  takže sa horšie rozlíši od rovnomerného rozdelenia.

### 4.3 Zhrnutie

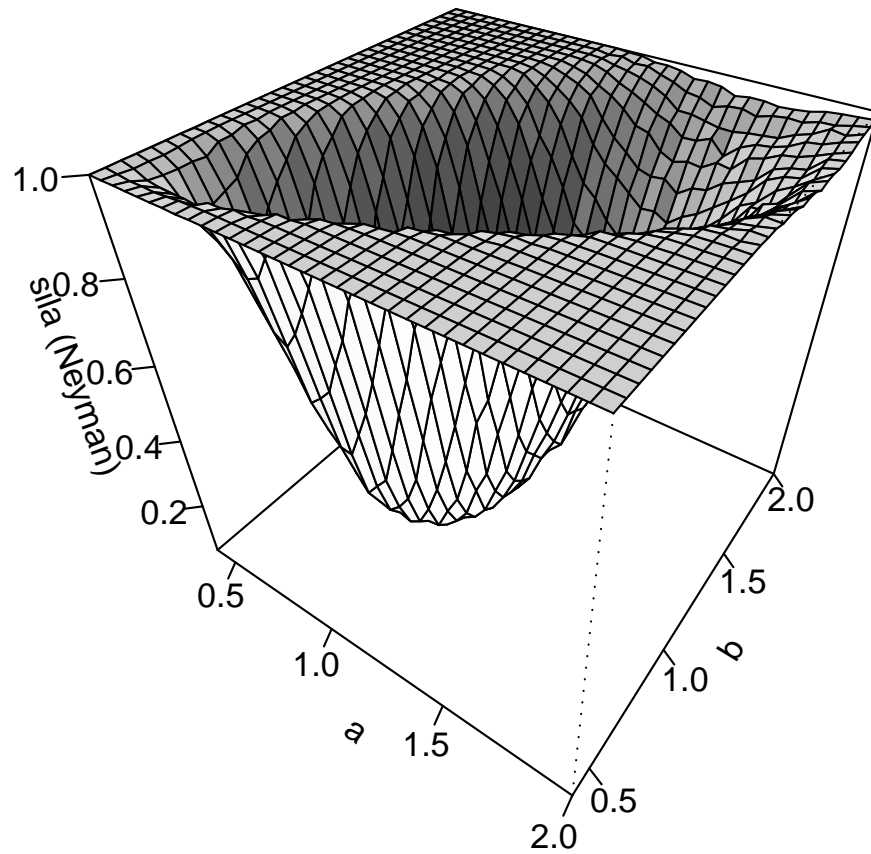
Simulačné štúdie ukázali, že Neymanove testy sú zdatnými konkurentmi testu Kolmogorov–Smirnov. Testy dodržujú predpísanú hladinu – už pri počte pozorovaní  $n = 50$  sú asymptotické kritické hodnoty použiteľné. Neymanov test je v zrovnaní s testom Kolmogorov–Smirnov lepšie schopný detekovať nesplnenie  $H_0$ , keď sa skutočné rozdelenie dáť líši od hypotetického inak než strednou hodnotou.



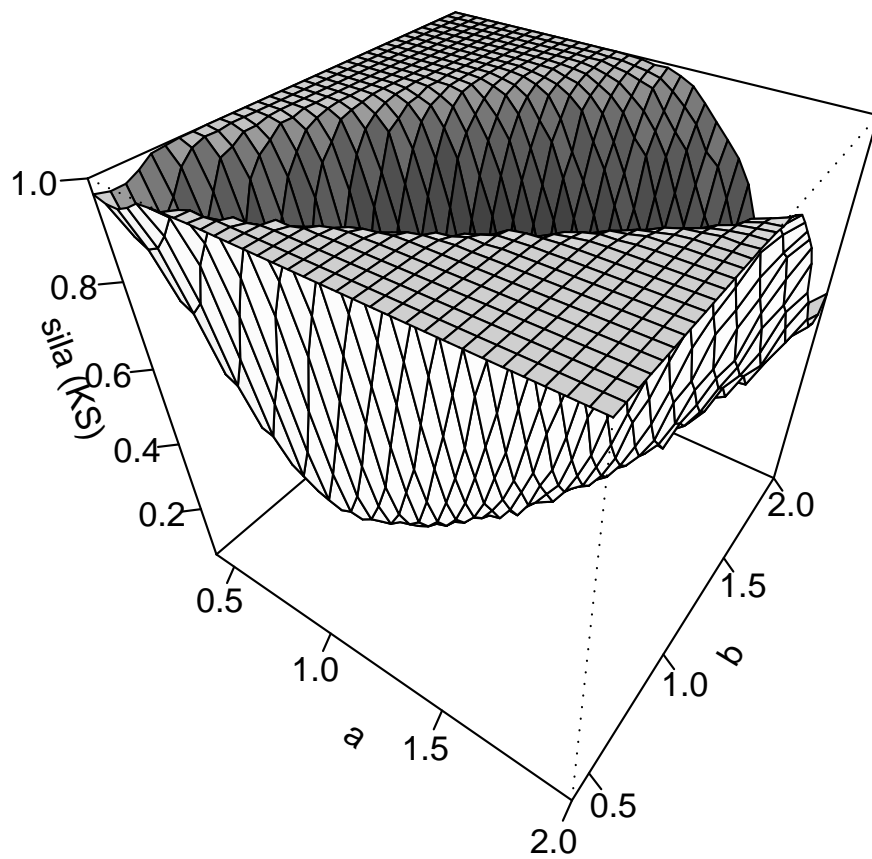
Obrázok 4.5: Graf silofunkcie Neymanovho testu (červenou farbou) a testu Kolmogorov–Smirnov (modrou farbou) na diagonále ( $a = b$ ).



Obrázok 4.6: Silofunkcia Neymanovho testu pri testovaní hypotézy  $H_0 : X \sim R(0, 1)$ , keď skutočné rozdelenie je  $B(a, b)$ .



Obrázok 4.7: Silofunkcia testu Kolmogorov–Smirnov pri testovaní hypotézy  $H_0 : X \sim R(0,1)$ , keď skutočné rozdelenie je  $B(a,b)$ .



# Literatúra

- [1] Anděl, J.: *Základy matematické statistiky*, Matfyzpress, Praha, 2001.
- [2] Kulich, M.: *Asymptotické testy hypotéz v modelech s rušivými parametry*, Robust 2000, JČMF, Praha, 2000.
- [3] Rayner, J.C.W., Best, D.J.: *Smooth tests of goodness of fit for regular distribution*, Oxford University Press, New York, 1988.
- [4] Abramowitz, M., Stegun, I.A.: *Handbook of mathematical functions*, Dover Publications, Inc., New York, 1972.