UNIVERZITA KOMENSKÉHO V BRATISLAVE
FAKULTA MATEMATIKY, FYZIKY A
INFORMATIKY

VPLYV ZAOKRÚHĽOVANIA NA FUNGOVANIE ŠTATISTICKÝCH METÓD

BAKALÁRSKA PRÁCA

2012 Michal Hojčka

Katedra aplikovanej matematiky a štatistiky
Fakulta Matematiky, Fyziky a Informatiky
Univerzita Komenského, Bratislava

Vplyv zaokrúhľovania na fungovanie štatistických metód

(Bakalárska práca)

Michal Hojčka

Vedúci: Mgr. Ján Somorčík, PhD.
Odbor: Aplikovaná matematika 1114
Program: Ekonomická a finančná matematika
Kód práce: 15bf13b3-d921-4207-bf07-adf9466c7cce Bratislava, 2012

Univerzita Komenského v Bratislave
Fakulta matematiky, fyziky a informatiky

ZADANIE ZÁVEREČNEJ PRÁCE

Meno a priezvisko študenta: Michal Hojčka
Študijný program: ekonomická a finančná matematika (Jednoodborové štúdium, bakalársky I. st., denná forma)
Študijný odbor: 9.1.9. aplikovaná matematika
Typ záverečnej práce: bakalárska
Jazyk záverečnej práce: slovenský

Názov: Vplyv zaokrúhľovania na fungovanie štatistických metód

Cieľ: Autor si z literatúry naštuduje fungovanie niekoľkých štatistických metód. Následne pomocou počítačových simulácií posúdi, ako na ich fungovanie vplýva zaokrúhľovanie vstupných dát.

Vedúci: Mgr. Ján Somorčík, PhD.

Dátum zadania: 15.10.2011

Dátum schválenia: 27.10.2011

doc. RNDr. Margaréta Halická, CSc.
garant študijného programu

.....
študent

.....
vedúci

Poďakovanie

Chcel by som poďakovať môjmu vedúcemu bakalárskej práce Mgr. Jánovi Somorčíkovi, PhD. za čas, ktorý mi venoval a za cenné rady a pripomienky. Ďalej by som rád poďakoval svojim rodičom za ich podporu a za to, že ma priviedli k matematike a Lucke za všetko, čo pre mňa urobila.

Abstrakt

Autor: Michal Hojčka
Názov bakalárskej práce: Vplyv zaokrúhľovania na fungovanie štatistických metód
Škola: Univerzita Komenského v Bratislave
Fakulta: Fakulta matematiky, fyziky a informatiky
Katedra: Katedra aplikovanej matematiky a štatistiky
Vedúci bakalárskej práce: Mgr. Ján Somorčík, PhD.
Rozsah práce: 46 strán

Bratislava, 2012

Bakalárska práca sa týka zaokrúhľovania a chýb, ktoré spôsobuje v štatistike. Odhady získané klasickými spôsobmi sú v závislosti od stupňa zaokrúhlenia a pozície rozdelenia dát vzhľadom na zaokrúhľovací interval viac či menej vychýlené, čiže sa stávajú nevhodnými na používanie v štatistike. Prvá kapitola ponúka simulácie na ukázanie vychýlenosti, vrátane grafických znázornení. V práci je možné nájsť taktiež spôsob ako zahrnúť zaokrúhľovanie do $Maximum\ \ Likelihood(ML)$ odhadovania. Fungovanie a správanie sa tohoto prístupu je v tejto práci ukázané na použití ML odhadu pri hľadaní parametrov v exponenciálnom rozdelení, normálnom rozdelení so známou aj neznámou hodnotou σ a v lineárnej regresii.

KĽÚČOVÉ SLOVÁ: zaokrúhľovanie, odhad, maximálna vierohodnosť, exponenciálne rozdelenie, normálne rozdelenie, lineárna regresia

Abstract

Author: Michal Hojčka
Name of Bachelor Thesis: Influence of rounding on the performance of statistical methods
University: Comenius University, Bratislava, Slovakia
Faculty: Faculty of Mathematics, Physics and Informatics
Department: Department of Applied Mathematics and Statistics
Thesis advisor: Mgr. Ján Somorčík, PhD
Number of Pages: 46

Bratislava, 2012

This bachelor thesis relates to rounding and errors it causes in statistics. Estimates obtained by conventional methods are more or less biased, depending on the degree of rounding and the position of data distribution in regard of rounding interval, so they become inappropriate for use in statistics. The first chapter provides simulations to present bias including graphical representations. The way how to involve rounding in Maximum Likelihood(ML) estimate can be also found in this thesis. Operation and performance of this approach is shown in this thesis on using ML estimate to look for parametres of the exponential distribution, normal distribution with both known and unknown value of σ and also in the linear regression.

KEYWORDS: rounding, estimate, maximum likelihood, exponential distribution, normal distribution, linear regression

Obsah

Úvod **1**

1 Vplyv zaokrúhľovania na klasické odhady **2**

2 Odhad parametrov pomocou ML **21**

2.1 Odhad parametra λ exponenciálneho rozdelenia 23

2.2 Odhad parametra μ normálneho rozdelenia pri známej hodnote σ 27

2.3 Odhad parametra μ a σ normálneho rozdelenia pri neznámej hodnote σ 33

2.4 Odhad parametrov α, β a σ v lineárnej regresii 38

Záver **46**

Úvod

Ak sledujeme náhodnú premennú, ktorá má spojité rozdelenie, je nevyhnutné, že napozorované hodnoty budú iba zaokrúhlené odhady ich skutočných hodnôt. Veľkosť zaokrúhlenia a tým pádom aj veľkosť odchýliek od skutočných hodnôt je určená presnosťou meracích prístrojov, z ktorých sme dané pozorovania získali. Z náhodnej premennej so spojitým rozdelením sa stane náhodná premenná s diskrétnym rozdelením. Sú teda veľmi ľahko pochopiteľné obavy, že z takto zdeformovaných dát sa nebudú dať získať presné odhady a štatistické metódy, používajúce zaokrúhlené dáta budú generovať nesprávne závery. Priekopníkom v tejto oblasti štatistiky bol Sheppard(1898) a jeho článok [6]. Tomuto problému sa taktiež venoval aj Hejtan, napríklad aj v [3]. Kvôli nárastu počtu štatistických výskumov a pozorovaní s nástupom počítačov je táto oblasť stále veľmi dôležitá a aj v súčasnosti sa jej venuje veľká pozornosť, čo dokumentuje niekoľko rokov starý článok [1].

V tejto práci sa budeme snažiť ukázať, akým spôsobom vplýva zaokrúhľovanie na odhadovanie parametrov náhodného výberu z normálneho rozdelenia, naše výsledky porovnáme s už existujúcimi simuláciami, popísanými v [1]. V ďalšej časti opíšeme spôsob odhadovania parametrov, prezentovaný v [1], ktorý zahŕňa aj zaokrúhlenie hodnôt náhodného výberu. Následne ho použijeme na odhad parametrov niektorých známych rozdelení, ako aj v lineárnej regresii. Pomocou ďalších simulácii overíme, či tento spôsob naozaj poskytne lepšie odhady ako keď nezohľadňujeme zaokrúhlenie.

Kapitola 1

Vplyv zaokrúhľovania na klasické odhady

V tejto kapitole pomocou niekoľkých simulácii demonštrujeme, ako sa bude zaokrúhľovanie prejavovať na náhodnom výbere z normálneho rozdelenia. Základné vedomosti o štatistických metódach sme čerpali z [2]. Náhodný výber z $N(\mu, \sigma^2)$ si označme ako $\{x_1, x_2, \cdots, x_n\}$. Naše pozorované dáta sú celé čísla $\{y_1, y_2, \cdots, y_n\}$, čo sú vlastne zaokrúhlené hodnoty $\{x_1, x_2, \cdots, x_n\}$. V našom prípade budeme zaokrúhľovať na celé čísla. Interval $[y-0.5, y+0.5)$, resp. $(y-0.5, y+0.5]$ si nazvime zaokrúhľovací interval. Pre všetky x_i, ktoré budú ležať v danom intervale bude platiť $y_i = y$. Jeho šírku budeme označovať ako h. Keďže ten náš má šírku 1, tak v našom prípade platí $h = 1$. Môžeme si označiť, že pre zaokrúhlený výber platí: $E(Y) = \tilde{\mu}$, $D(Y) = \tilde{\sigma}^2$. Ak by sme zanedbali fakt, že dáta boli zaokrúhlené, tak strednú hodnotu $\tilde{\mu}$ by sme mali odhadovať pomocou výberového priemeru ako $\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$ a disperziu $\tilde{\sigma}^2$ pomocou výberovej disperzie ako $\tilde{s}^2 = \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})^2$. Pokúsime sa overiť, či $\bar{y}$ je dobrý odhad aj pre skutočné μ, čiže budeme testovať hypotézu $H_0 : \tilde{\mu} = \mu$. Ak platí H_0, tak testovacia štatistika $\frac{\sqrt{n}\cdot(\bar{y}-\mu)}{\tilde{s}}$ musí mať

Studentovo rozdelenie s $(n-1)$ stupňami voľnosti. Potom by chyba I. druhu pre obojstranný Studentov t-test na hladine významnosti 5% mala byť rovná 5%, malo by teda platiť nasledovné:

$$P\left(\frac{\sqrt{n}\cdot|\bar{y}-\mu|}{\tilde{s}} \geq t_{n-1,2.5\%}\right) = 5\%,$$

kde $t_{n-1,2.5\%}$ je 2.5%-ná kritická hodnota pre Studentovo rozdelenie s $(n-1)$ stupňami voľnosti. V [1] sa na podobnej simulácii pre σ rovné 0.25 a pre rôzne hodnoty μ ukázalo, že v danom prípade náhodné premenné $\bar{y}$ a $\tilde{s}^2$ nie sú vhodné na používanie ako odhady pre μ a σ^2 a chyba I.druhu pre veľké hodnoty n dosahovala 100%. V [3] sa tvrdí, že pre určité, nie príliš extrémne hodnoty n a h platí, že $\bar{y}$ je dobrý odhad pre μ a že $\tilde{s}^2$ nie je dobrý odhad pre σ^2, ale dá sa napraviť odčítaním hodnoty $\frac{h^2}{12}$, kde h je dĺžka zaokrúhľovacieho intervalu, v našom prípade používame $h = 1$, čiže dá sa povedať, že $\tilde{s}^2$ je dobrý odhad pre $\sigma^2 + \frac{1}{12}$. Táto úprava sa zvykne nazývať Sheppardova korekcia. Čiže po odrátaní $\frac{1}{12}$ dostaneme z vyrátanej disperzie skutočnú disperziu. Platnosť tejto korekcie sa pokúsime overiť pomocou chí-kvadrát testu. Vieme, že ak sa dve disperzie rovnajú, tak testovacia štatistika pre hypotézu $H_0 : \sigma_1^2 = \sigma^2$ je rovná $\frac{(n-1)\cdot s_1^2}{\sigma^2}$ a má chí-kvadrát rozdelenie s $(n-1)$ stupňami voľnosti. V našom pozmenenom prípade sa bude jednať o hypotézu $H_0 : \tilde{\sigma}^2 = \sigma^2 + \frac{1}{12}$ a teda platí

$$\frac{(n-1)\cdot\tilde{s}^2}{\sigma^2+\frac{1}{12}} \sim \chi^2_{n-1}.$$

Chyba I. druhu na hladine významnosti 5% je rovná 5%, čiže ak hypotéza H_0 platí, tak

$$P\left(\frac{(n-1)\cdot\tilde{s}^2}{\sigma^2+\frac{1}{12}} \leq \chi^2_{n-1,97.5\%} \vee \frac{(n-1)\cdot\tilde{s}^2}{\sigma^2+\frac{1}{12}} \geq \chi^2_{n-1,2.5\%}\right) = 5\%.$$

V tejto kapitole ukážeme výsledky niekoľkých našich vlastných simulácii k tejto téme pre rôzne hodnoty n, μ a σ^2 , kde sa pokúsime ukázať správnosť

tvrdení z [1] a [3] a pokúsime sa určiť hranicu, odkedy sú odhady, pri ktorých zanedbávame fakt, že dáta boli zaokrúhlené, vhodné na používanie.

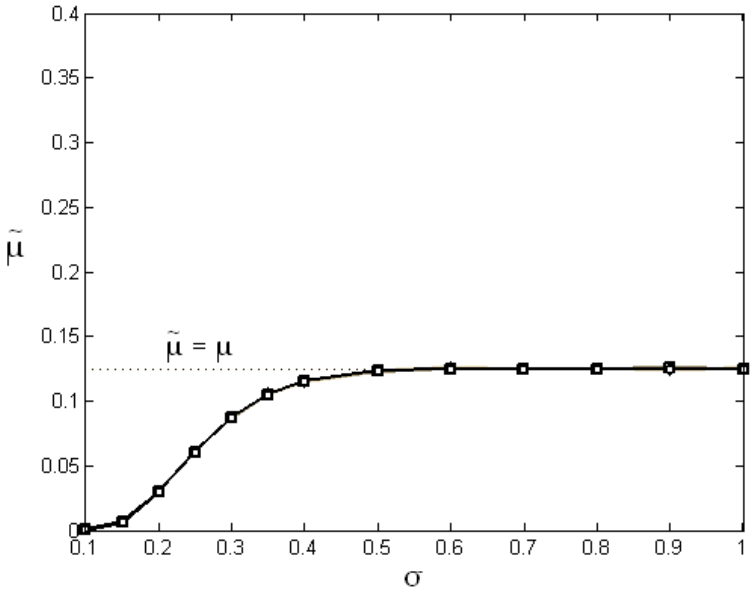
Simulácia 1.1 (sledovanie správania sa odhadu $\bar{y}$)

V tejto simulácii sa zameriame na sledovanie $\bar{y}$ a jeho vychýlenosti. Budeme teda hľadať $E(\bar{y}) = \tilde{\mu}$. Pre nevychýlený odhad by sa stredná hodnota mala rovnať μ. Budeme sledovať hodnotu $\tilde{\mu}$ vzhľadom na zmenu n, σ a μ. Ako hodnoty n budeme používať hodnoty 10, 100 *a* 1000. *Chceme zistiť, či vychýlenosť $\bar{y}$ závisí od veľkosti náhodného výberu. Ako hodnoty σ budeme používať hodnoty* 0.1, 0.15, 0.2, 0.25, 0.3, 0.35, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9 *a* 1. *Očakávame, že pre malé hodnoty σ bude vychýlenosť značná ale s rastúcou sigmou dosiahneme nevychýlený odhad, ako sa tvrdí v [3]. Ako hodnoty μ budeme používať hodnoty* 0, 0.125, 0.25, 0.375 *a* 0.5. *Chceme zistiť, pre akú polohu μ vzhľadom na zaokrúhľovací interval má zaokrúhľovanie najväčšie účinky. Keďže zaokrúhľovanie je symetrické okolo hodnoty* 0.5, *je zbytočné sledovať výsledky aj pre μ medzi* 0.5 *a* 1, *pretože sú to len ekvivalenty hodnôt μ medzi* 0 *a* 0.5, *napríklad výsledky pre $\mu = 0.75$ budú symetrické k výsledkom pre $\mu = 0.25$ podľa čísla* 0.5. *Simuláciu opakujeme* 100000*-krát, čo je dostatočne významná vzorka, takže naše výsledky môžeme považovať za štatisticky významné.*

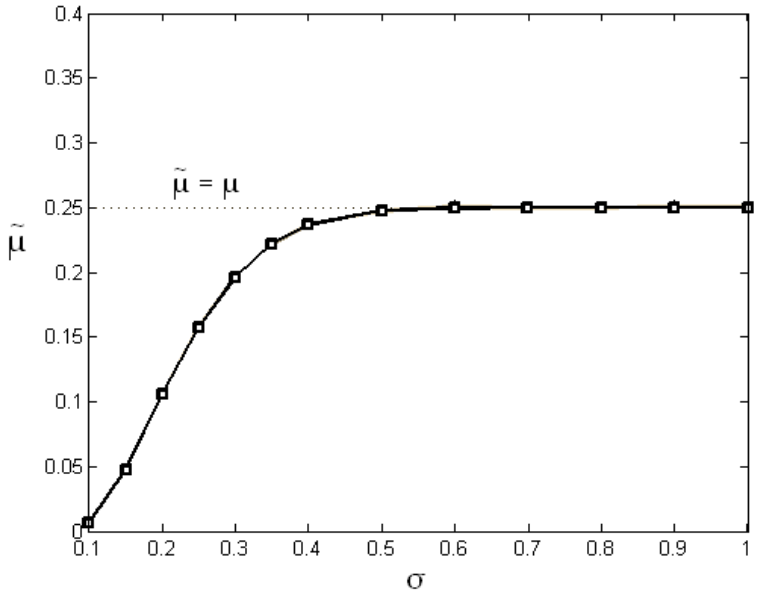
Pre hodnotu $\mu = 0$ a $\mu = 0.5$ nám vyšlo $E(\bar{y}) = \tilde{\mu}$ veľmi blízke μ pre všetky kombinácie σ a n. To nie je až také prekvapujúce, keďže zaokrúhľovanie je symetrické vzhľadom na čísla 0 *a* 0.5, *jedno leží na okraji medzi dvoma intervalmi a druhé v strede zaokrúhľovacieho intervalu. Pre $\mu = 0$ je vždy bez ohľadu na σ a n rovnaká pravdepodobnosť, že $y_i = 1$ alebo $y_i = -1$, takisto to platí pre ľubovoľné iné celé čísla k a $-k$, preto je jasné, že zaokrúhľovanie strednú hodnotu neovplyvní. Pre $\mu = 0.5$ je obdobne vždy rovnaká pravdepodobnosť, že $y_i = 0$ alebo $y_i = 1$, takisto to platí pre ľubovoľné iné celé čísla k a $1 - k$, preto zaokrúhľovanie strednú hodnotu ani v tomto prípade neovplyvní. Preto nemá význam prikladať obrázky k týmto prípadom.*

Pre hodnoty $\mu = 0.125$, 0.25 a 0.375 nám vyšli ale celkom zaujímavé výsledky, čo dokumentujú nasledujúce obrázky. Pre každú hodnotu μ máme samostatný obrázok, kde je znázornená závislosť hodnoty $\tilde{\mu}$ od hodnoty σ(os x) a veľkosti výberu n(3 rôznofarebné grafy na každom obrázku - na obrázkoch v tejto simulácii sa ale nedajú rozoznať) . V obrázkoch je vyznačená čiara $\tilde{\mu} = \mu$, kde by sa $\tilde{\mu}$ malo nachádzať, ak by $\bar{y}$ bol nevychýlený odhad pre μ.

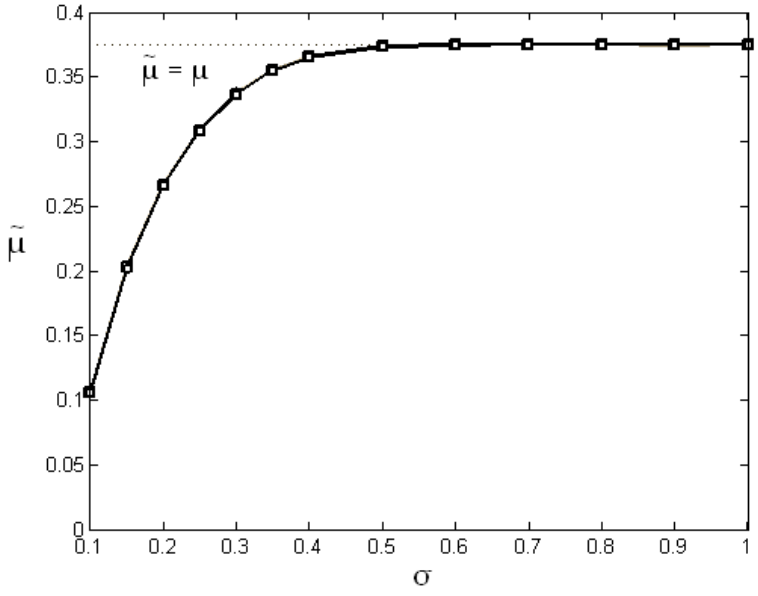
Obr. 1.1: Správanie sa $\tilde{\mu}$ ak $\mu = 0.125$



Obr. 1.2: Správanie sa $\tilde{\mu}$ ak $\mu = 0.25$



Obr. 1.3: Správanie sa $\tilde{\mu}$ ak $\mu = 0.375$



Prvá vec, ktorú si môžeme všimnúť je, že v každom obrázku grafy pre $n = 10$, $n = 100$ *a* $n = 1000$ *takmer úplne splývajú, takže sa zdá, že na obrázku je iba jeden graf. Takisto sa môžeme presvedčiť, že pre* $\sigma = 0.25$ *nám vyšli veľmi podobné výsledky ako v* [1] *. Napríklad pre* $\mu = 0.25$ *nám vyšiel odhad* $\tilde{\mu} = 0.1571$, *v* [1] *bola použitá hodnotu* $\mu = 2.25$ *a v ich simulácii vyšla hodnota* $\tilde{\mu} = 2.157$, *v oboch prípadoch je odchýlka rovná* 0.093. *Presvedčili sme sa, že odhady môžu byť veľmi nepresné pre malé hodnoty* σ, *ale so zväčšujúcou sa hodnotou* σ *sa hodnota* $\tilde{\mu}$ *blíži k skutočnej hodnote* μ. *Natíska sa otázka, od akej hodnoty* σ *sa dá* $\tilde{\mu}$ *považovať za rovné* μ.

Výsledky simulácie môžeme zhrnúť do nasledujúcich záverov.

1. *Skutočná hodnota* $E(\bar{y}) = \tilde{\mu}$ *nezávisí od veľkosti náhodného výberu.*

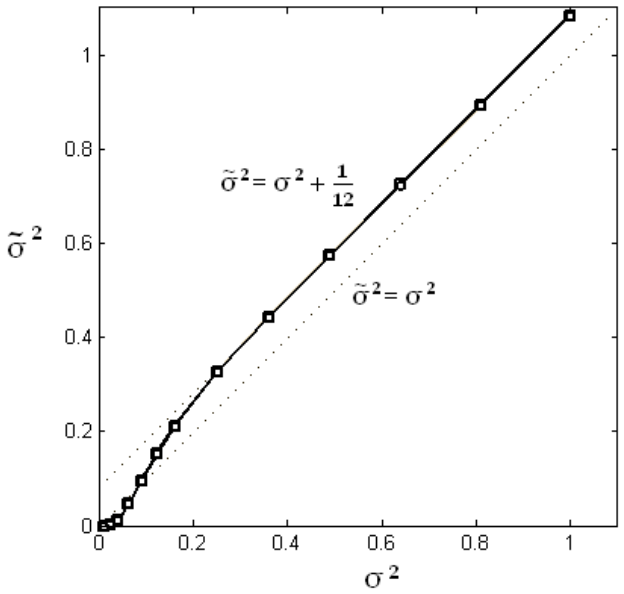
2. *Pre* σ *menšie ako určitá hranica je odhad* $\bar{y}$ *vychýleným odhadom pre* μ, *pre veľké* σ *je* $\bar{y}$ *dobrý odhad pre* μ.

Simulácia 1.2 (sledovanie správania sa odhadu $\tilde{s}^2$)

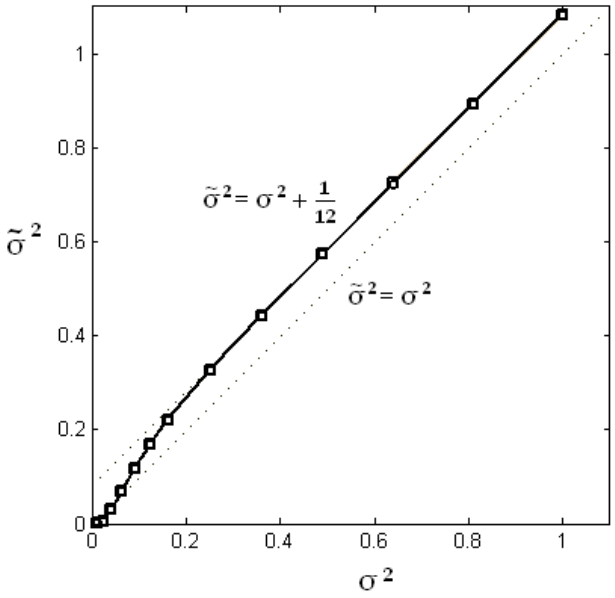
V tejto simulácii sa zameriame na sledovanie $\tilde{s}^2$ *a jeho vychýlenosti. Budeme teda hľadať* $E(\tilde{s}^2)$. *Pre nevychýlený odhad by sa mala rovnať* σ^2. *Podľa* [3] *by sa mala rovnať* $\sigma^2 + \frac{1}{12}$. *Budeme sledovať zmeny vzhľadom na zmenu* n, σ *a* μ. *Ako hodnoty* n *budeme používať hodnoty* 10, 100 *a* 1000. *Chceme zistiť, či vychýlenosť* $\tilde{s}^2$ *závisí od veľkosti náhodného výberu. Ako hodnoty* σ *budeme používať hodnoty* 0.1, 0.15, 0.2, 0.25, 0.3, 0.35, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9 *a* 1. *Ako hodnoty* μ *budeme opäť používať hodnoty* 0, 0.125, 0.25, 0.375 *a* 0.5, *pričom nakreslíme obrázok pre každú z týchto stredných hodnôt. Opäť použijeme* 100000 *opakovaní, aby sme zistili dostatočne presnú hodnotu* $E(\tilde{s}^2) = \tilde{\sigma}^2$. *Výsledky dokumentujú nasledujúce obrázky. Na osi x je vždy skutočná hodnota* σ^2, *na osi y je vypočítaná hodnota* $\tilde{\sigma}^2$. *Na obrázkoch sú vyznačené* 2 *dôležité priamky, prvou je priamka* $\tilde{\sigma}^2 = \sigma^2$ *na ktorej sa nachádzajú nevychýlené odhady pre* σ^2. *Druhou je priamka* $\tilde{\sigma}^2 = \sigma^2 + \frac{1}{12}$, *kde sa nachádzajú nevychýlené odhady pre* $\sigma^2 + \frac{1}{12}$ *a ku ktorej by sa podľa* [3]

mali blížiť aj naše hodnoty $\tilde{\sigma}^2$. Na každom obrázku sú opäť 3 grafy pre rôzne hodnoty n, ale prekrývajú sa, takže sa opäť zdá, že na každom obrázku je iba jeden graf.

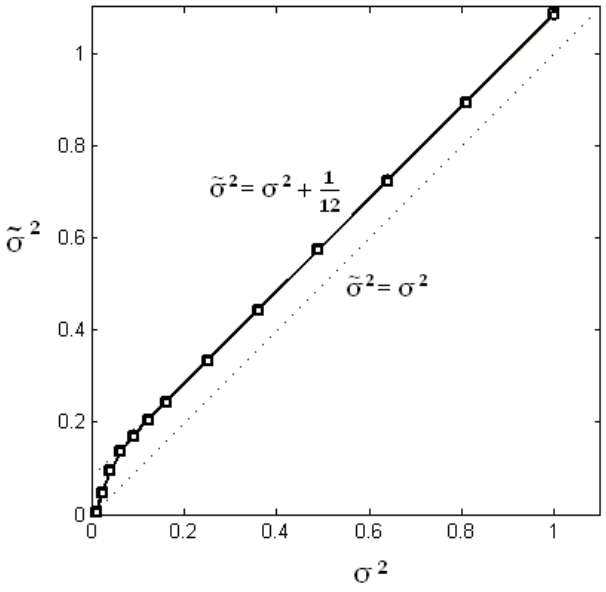
Obr. 1.4: Správanie sa $\tilde{\sigma}^2$ ak $\mu = 0$



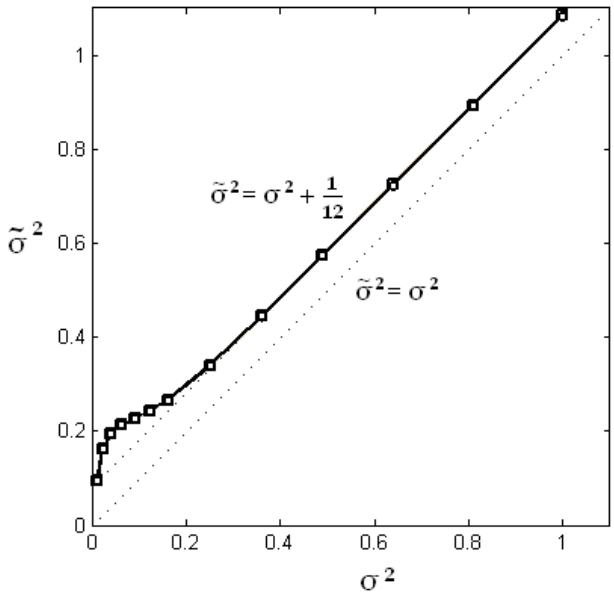
Obr. 1.5: Správanie sa $\tilde{\sigma}^2$ ak $\mu = 0.125$



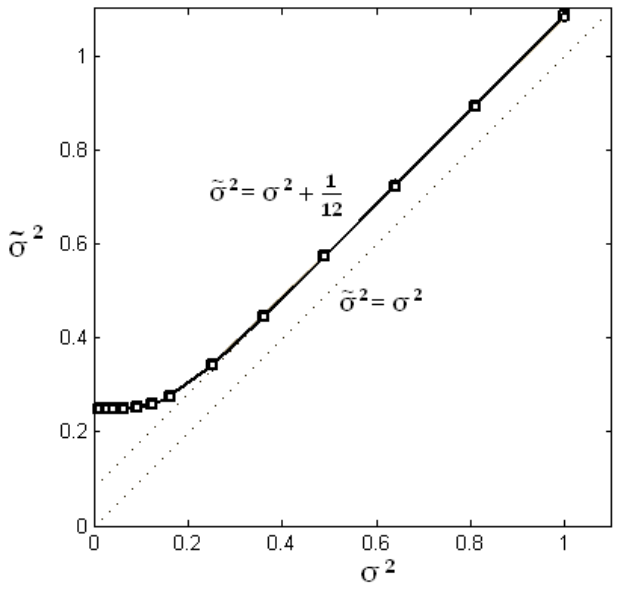
Obr. 1.6: Správanie sa $\tilde{\sigma}^2$ ak $\mu = 0.25$



Obr. 1.7: Správanie sa $\tilde{\sigma}^2$ ak $\mu = 0.375$



Obr. 1.8: Správanie sa $\tilde{\sigma}^2$ ak $\mu = 0.5$



Opäť prvá vec, čo si môžeme všimnúť je fakt, že aj keď sú na každom obrázku 3 *grafy, takmer úplne sa prekrývajú, takže viditeľný je iba jeden, čo znamená, že veľkosť náhodného výberu* n *nevplýva na veľkosť hodnoty* $\tilde{\sigma}^2$*. Znovu môžeme naše výsledky porovnať so simuláciou v [1], kde pre* $\mu = 2.25$*,* $\sigma^2 = 0.25^2$ *vyšiel výsledok* $\tilde{\sigma}^2 = 0.135$*. Nám pre* $\mu = 0.25$ *a* $\sigma^2 = 0.25^2$ *vyšiel výsledok* $\tilde{\sigma}^2 = 0.1352$*, čo opäť potvrdilo správnosť našej simulácie. Vidíme, že hodnota* $\tilde{\sigma}^2$ *sa pre veľké hodnoty* σ^2 *správa rovnako pre ľubovoľnú kombináciu* n *aj* μ *a totiž je veľmi blízka hodnote* $\sigma^2 + \frac{1}{12}$*, čo súhlasí s tvrdeniami v [3]. Tiež vidíme, že pre malé hodnoty* σ^2 *odhad pre* $\tilde{\sigma}^2$ *naopak výrazne závisí od hodnoty* μ *a jeho polohy v zaokrúhľovacom intervale. To nie je až také neočakávané. Vezmime si napríklad prípad* $\mu = 0$*, čo je presne stred zaokrúhľovacieho intervalu* $(-0.5, 0.5)$*. Na to aby* $y_i \neq 0$*, musí byť* $|x_i| \geq 0.5$*, čo je ale pri veľkosti štandardnej odchýlky menšej ako* 0.1 *takmer nemožné. No a keď* $\forall i : y_i = 0$*, tak potom aj výberová disperzia bude rovná* 0*. Opačným extrémom je prípad* $\mu = 0.5$*, ktorý leží na hranici dvoch zaokrúhľovacích intervalov* $(-0.5, 0.5)$ *a* $[0.5, 1.5)$*. Nech zvolím ľubovoľne malú nenulovú* σ^2*, tak platí, že* $y_i = 0 \vee y_i = 1$*. A keďže vieme, že v spojitom rozdelení pravdepodobnosti je nulová šanca, že hodnota* x_i *bude presne* 0.5*, je rovnaká pravdepodobnosť, že* $y_i = 0$ *ako, že* $y_i = 1$*. Keďže uvažujeme len o malých* σ^2*, je takmer nemožné, aby* $|x_i - 0.5| \geq 1$ *a teda takmer určite* $\forall i : y_i \in \{0, 1\}$*, čo je obyčajné alternatívne rozdelenie s pravdepodobnosťou* 0.5 *a teda výberová disperzia bude mať strednú hodnotu rovnú* 0.25*. Ostatné prípady budú ležať niekde medzi týmito dvomi extrémami.*

Výsledky môžeme zhrnúť do nasledujúcich záverov.

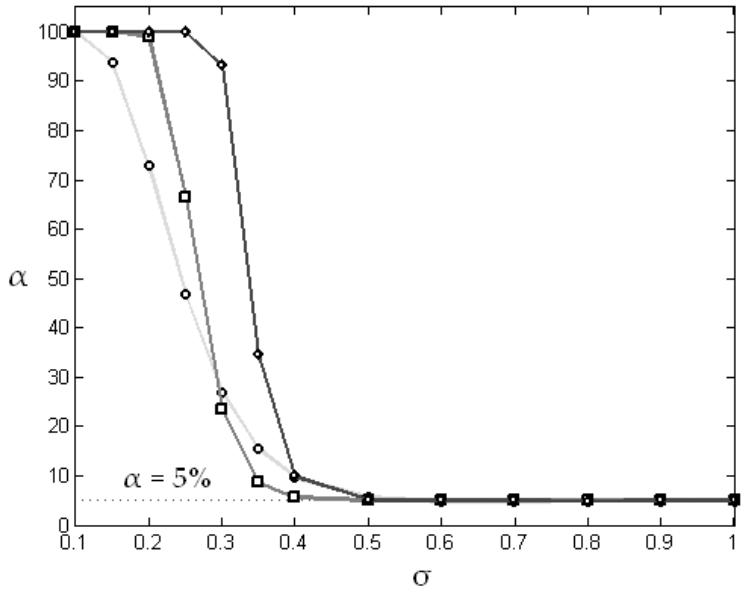
1. *Vychýlenosť odhadu* $\tilde{s}^2$ *nezávisí od veľkosti náhodného výberu.*
2. $\tilde{s}^2$ *je vychýlený odhad pre* σ^2*.*
3. *Pre malé hodnoty* σ^2 *veľkosť* $\tilde{\sigma}^2$ *výrazne závisí od polohy* μ *vzhľadom na zaokrúhľovací interval.*

4. *Pre σ väčšie ako určitá hranica je odhad $\tilde{s}^2$ nevychýleným odhadom pre $\sigma^2 + \frac{1}{12}$ pre ľubovoľnú polohu μ.*

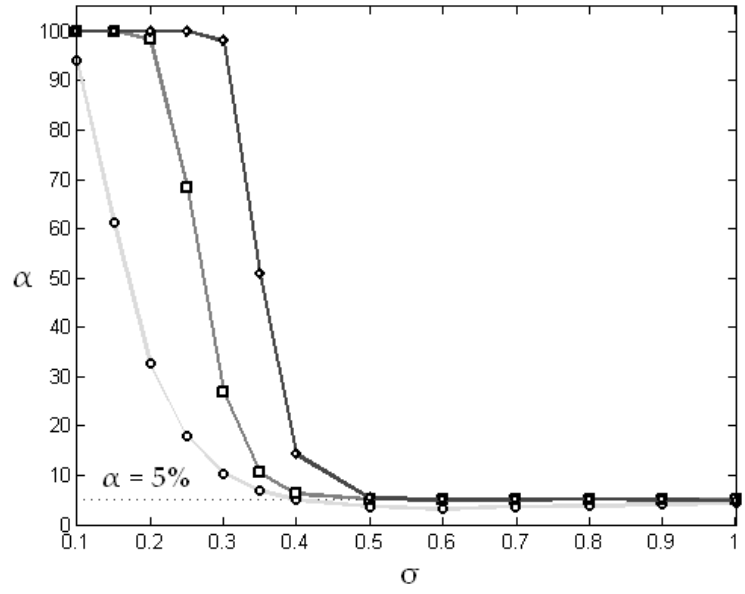
Simulácia 1.3 (sledovanie správania sa t-testu)

V tejto simulácii budeme overovať použiteľnosť $\bar{y}$ a $\tilde{s}^2$ v t-teste o μ. Budeme to robiť pomocou sledovania hodnoty chyby I. druhu Studentovho t-testu na hladine významnosti 5%. Vychádzať budeme z výsledkov dosiahnutých v prvých dvoch simuláciach, po každom zo 100000 opakovaní pomocou Studentovho t-testu overíme hypotézu $H_0 : \tilde{\mu} = \mu$. Chyba I. druhu je určená ako $\alpha = P\left(\frac{\sqrt{n}\cdot|\bar{y}-\mu|}{\tilde{s}} \geq t_{n-1,2.5\%}\right)$. Pre nezaokrúhlené hodnoty x sa hodnota α rovná 5%. Keďže nám konečne vyšli rôzne grafy pre rôzne hodnoty n, musíme ich vedieť rozlíšiť. Najtmavšia čiara s kosoštvorcami ako pozorovanými bodmi vyjadruje závislosť pre $n = 1000$, stredne tmavá čiara so štvorčekmi zastupuje $n = 100$ a najsvetlejšia čiara s krúžkami symbolizuje $n = 10$. Pre $\mu = 0$ a malú hodnotu σ bohužiaľ počítač nebol schopný zrátať hodnotu $\frac{\sqrt{n}\cdot|\bar{y}-\mu|}{\tilde{s}}$, keďže sa pri výpočte vyskytne podiel $\frac{0}{0}$. Preto obrázky a výsledky prikladáme len pre hodnoty $\mu = 0.125,\ 0.25,\ 0.375,\ a\ 0.5$.

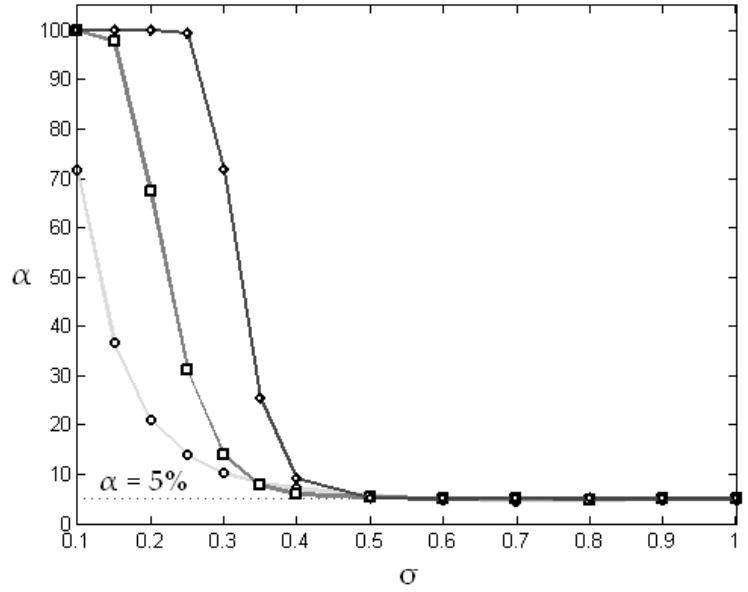
Obr. 1.9: Sledovanie chyby I. druhu v t-teste ak $\mu = 0.125$



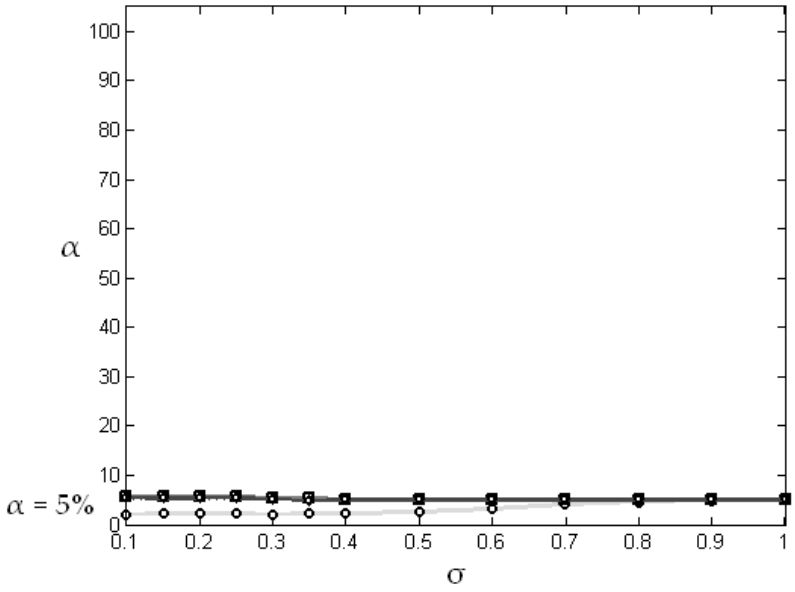
Obr. 1.10: Sledovanie chyby I. druhu v t-teste ak $\mu = 0.25$



Obr. 1.11: Sledovanie chyby I. druhu v t-teste ak $\mu = 0.375$



Obr. 1.12: Sledovanie chyby I. druhu v t-teste ak $\mu = 0.5$



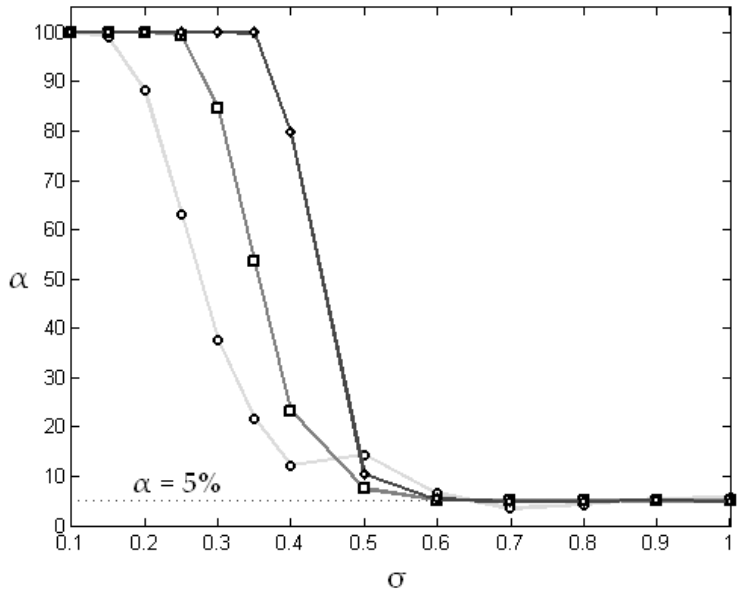
Z teórie pravdepodobnosti vieme, že ak p je skutočná pravdepodobnosť nejakého javu a $\hat{p}$ je vypozorovaná pravdepodobnosť tohoto javu pri n opakovaniach, tak z centrálnej limitnej vety vyplýva, že $\frac{\hat{p}-p}{\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}} \sim N(0,1)$. Vybavení touto vedomosťou vieme určiť 95%-ný interval spoľahlivosti okolo odhadu chyby I. druhu $\hat{p}$ ako $\left(\hat{p} - u_{2.5\%} \cdot \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}, \hat{p} + u_{2.5\%} \cdot \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\right)$. Skutočná hodnota pravdepodobnosti chyby I. druhu p sa nachádza v tomto intervale s pravdepodobnosťou 95%. V našom prípade overujeme, či $p = \alpha = 5\% = 0.05$. Ak sa 0.05 nebude nachádzať v danom intervali spoľahlivosti, môžeme zamietnuť hypotézu, že skutočná hodnota pravdepodobnosti I. druhu je 5%. Je zjavné, že hodnota 0.05, čiže 5% bude ležať v tomto intervale práve vtedy, keď $0.05 \geq \hat{p} - u_{2.5\%} \cdot \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \wedge 0.05 \leq \hat{p} + u_{2.5\%} \cdot \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$. Keď z daných nerovníc spravíme rovnice tak získame hraničné prípady, kedy sa 0.05 bude ešte nachádzať v intervale spoľahlivosti. Z rovnice $0.05 = \hat{p} - u_{2.5\%} \cdot \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$ získame riešenie $\hat{p} = 0.051368$. Z rovnice $0.05 = \hat{p} + u_{2.5\%} \cdot \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$ získame riešenie $\hat{p} = 0.048664$. Môžeme teda tvrdiť, že 0.05 bude ležať v intervale spoľahlivosti práve vtedy, keď $\hat{p}$ bude ležať v intervale s približnými hranicami $(0.048664,\ 0.051368)$. Teraz už môžeme určiť, pre ktoré μ, σ a n nemá zaokrúhľovanie vplyv na t-test. Pre $n = 10$ sa z našich pozorovaní nedá vypozorovať žiadna hranica, za ktorou by skutočná hodnota chyby I.druhu už vždy mohla byť rovná 5%. Mohli by sme teda tvrdiť, že pre $n = 10$ a $\sigma < 1$ zaokrúhľovanie prílíš ovplyvňuje veľkosť odhadov na to, aby sa dali používať v t-teste. Ale pre hodnoty $n = 100$ a $n = 1000$ sa celkom presne dá určiť hranica, odkedy už 5% leží v intervale spoľahlivosti. V nasledujúcej tabuľke uvedieme, pri akej hodnote σ sa chyba I. druhu už dá považovať za 5%, teda kde zhruba je hranica odkedy sú už dáta príliš zničené. Riešenie uvádzame ako interval, pretože nepoznáme presnú hranicu, odkedy sa 0.05 už nachádza v intervale spoľahlivosti, poznáme iba bod, kde odhad ešte nevyhovuje a bod, kde odhad už vyhovuje a vieme, že skutočná hranica bude niekde medzi nimi.

	n	*100*	*1000*
μ			
0.125		*(0.4,0.5)*	*(0.4,0.5)*
0.25		*(0.4,0.5)*	*(0.5,0.6)*
0.375		*(0.5,0.6)*	*(0.4,0.5)*
0.5		*(0.35,0.4)*	*(0.25,0.3)*

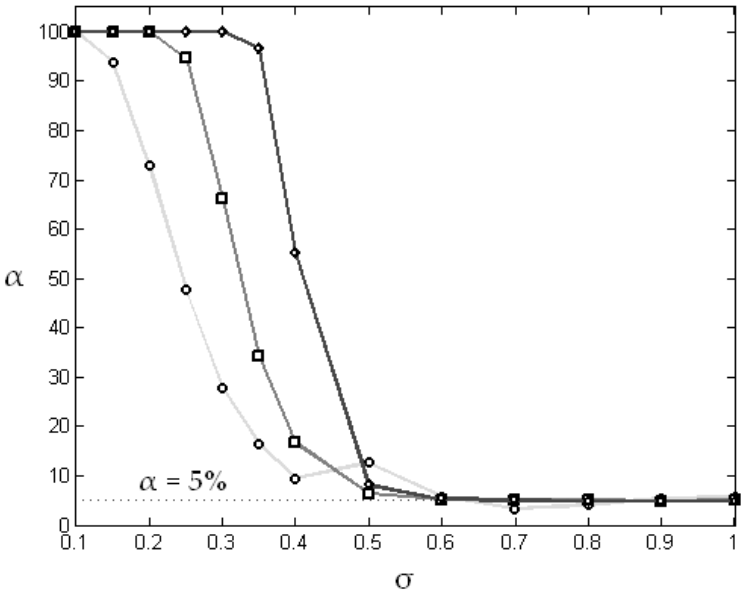
Keď si odmyslíme prípad, kde $\mu = 0.5$*, ktorý ako vieme je prípad degenerovaný, keďže* 0.5 *leží presne na hranici medzi dvomi zaokrúhľúhlovacími intervalmi a ako sa môžeme presvedčiť na obrázku, chyba I. druhu sa preň správa inak ako v ostatných prípadoch, tak vieme celkom vierohodne prehlásiť, že pre nie príliš malé hodnoty* n *a hodnoty* μ *z vnútra zaokrúhľovacieho intervalu sa t-test dá považovať za dôveryhodný, ak hodnota* σ *je nad nejakou hranicou. Podľa našich výsledkov sa táto hranica nachádza niekde v intervale* $(0.4, 0.6)$ *pre všetky takéto prípady.*

Simulácia 1.4 (sledovanie správania sa odhadu $\tilde{s}^2$ v chí-kvadrát teste) *V tejto simulácii budeme testovať vhodnosť korekcie* $\left(-\frac{h^2}{12}\right)$*, v našom prípade* $\left(-\frac{1}{12}\right)$*. Našou testovanou hypotézou bude hypotézu* $H_0 : \tilde{\sigma}^2 = \sigma^2 + \frac{1}{12}$*. Budeme to robiť pomocou sledovania hodnoty chyby I. druhu chí-kvadrát testu na overenie hodnoty disperzie na hladine významnosti* 5%*. Tá sa ráta ako* $\alpha = P\left(\frac{(n-1)\cdot\tilde{s}^2}{\sigma^2+\frac{1}{12}} \leq \chi^2_{n-1,97.5\%} \vee \frac{(n-1)\cdot\tilde{s}^2}{\sigma^2+\frac{1}{12}} \geq \chi^2_{n-1,2.5\%}\right)$*. Vychádzať budeme z výsledkov dosiahnutých v prvých dvoch simuláciach, po každom zo* 100000 *opakovaní pomocou chí-kvadrát testu zamietneme alebo nezamietneme* H_0*. Pre odhady získané z nezaokrúhlených hodnôt* x *sa* α *rovná* 5%*. Opäť vyšli rôzne výsledky pre rôzne hodnoty* n*, čiže najtmavšia čiara s kosoštvorcami ako pozorovanými bodmi vyjadruje závislosť pre* $n = 1000$*, stredne tmavá čiara so štvorčekmi zastupuje* $n = 100$ *a najsvetlejšia čiara s krúžkami symbolizuje* $n = 10$*. Obrázky a výsledky prikladáme pre všetky hodnoty* $\mu = 0,\ 0.125,\ 0.25,\ 0.375$ *a* 0.5*.*

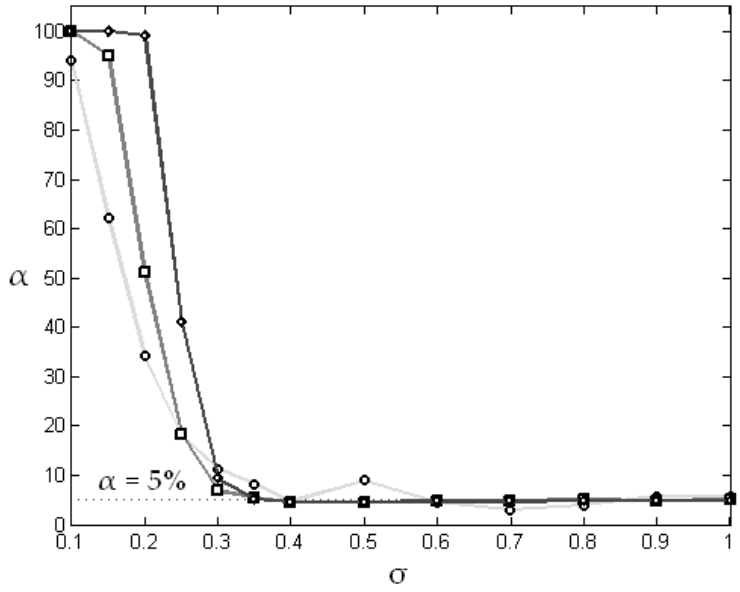
Obr. 1.13: Sledovanie chyby I. druhu v χ^2-teste ak $\mu = 0$



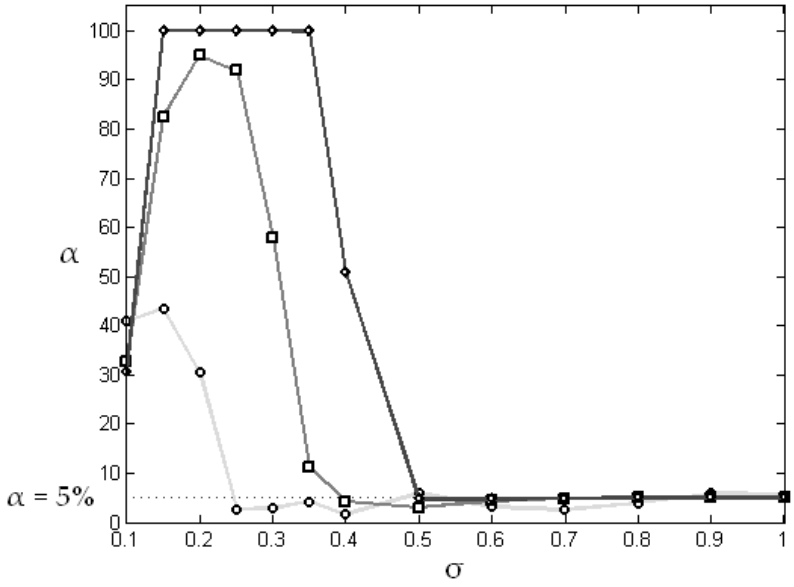
Obr. 1.14: Sledovanie chyby I. druhu v χ^2-teste ak $\mu = 0.125$



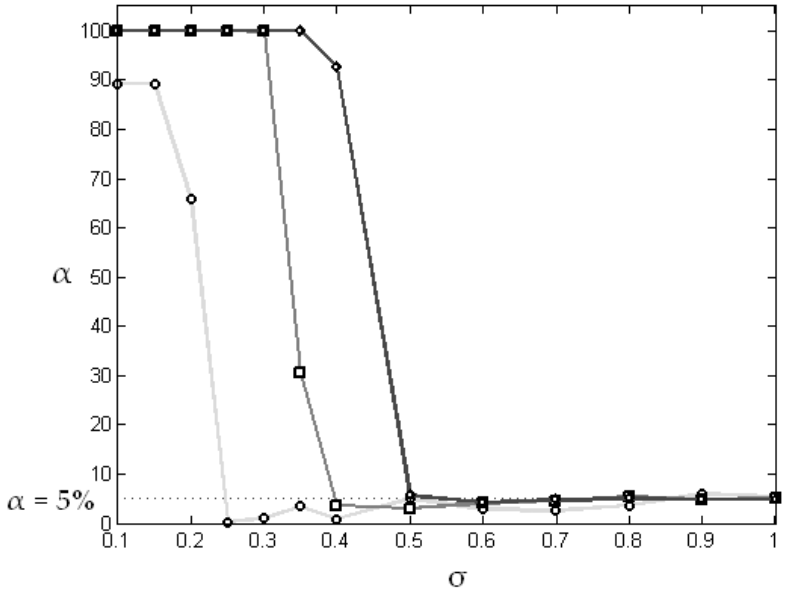
Obr. 1.15: Sledovanie chyby I. druhu v χ^2-teste ak $\mu = 0.25$



Obr. 1.16: Sledovanie chyby I. druhu v χ^2-teste ak $\mu = 0.375$



Obr. 1.17: Sledovanie chyby I. druhu v χ^2-teste ak $\mu = 0.5$



Rovnakými úvahami ako v predchádzajúcej simulácii dostaneme, že chybu I. druhu môžeme považovať za rovnú 5% *vtedy, ak vypočítaná pravdepodobnosť je z intervalu* (0.048664, 0.051368). *Opäť sme zistili, chyba I. druhu v prípade, keď* $n = 10$ *sa pre takto malé hodnoty* σ *správa stále veľmi chaoticky a nepresne a teda odhad* $\tilde{s}^2$ *pre* $n = 10$ *a* $\sigma < 1$ *nie je vhodný na používanie. Ale pre prípady, kde* $n = 100$ *a* $n = 1000$ *sa opäť môžeme pokúsiť určiť hranicu, odkedy je odhad dobrý.*

	n	*100*	*1000*
μ			
0		*(0.7,0.8)*	*(0.7,0.8)*
0.125		*(0.5,0.6)*	*(0.9,1)*
0.25		*(0.7,0.8)*	*(0.9,1)*
0.375		*(0.7,0.8)*	*(0.7,0.8)*
0.5		*(0.8,0.9)*	*(0.7,0.8)*

Nevyšli nám také pekné výsledky ako v predchádzajúcej simulácii, ale vidíme, že pre väčšie σ *už* 5% *vždy leží v intervale spoľahlivosti, môžeme teda prehlásiť, že pre nami skúmané* μ *a* $n = 100$ *resp.* $n = 1000$ *a* $\sigma > 1$ *platí, že* $\tilde{\sigma}^2 = \sigma^2 + \frac{1}{12}$.

Kapitola 2

Odhad parametrov pomocou ML

V tejto kapitole sa zameriame na spôsob, ako zo zaokrúhleného výberu vieme získať čo najlepšie odhady pre parametre pôvodného rozdelenia z ktorého bol výber získaný. Na odhad použijeme ML, teda "maximum-likelihood", tento termín sa do slovenčiny prekladá ako odhad pomocou metódy maximálnej vierohodnosti, kde sa snažíme nastavením parametra maximalizovať pravdepodobnosť, že odhadované dáta sú také, ako sme napozorovali. Tento spôsob bol prezentovaný v článku [1]. V ďalšej časti vypočítame ML odhad pre rôzne rozdelenia, konkrétne exponenciálne rozdelenie, normálne rozdelenie a lineárnu regresiu a pokúsime sa ukázať, že dosahuje lepšie výsledky ak zahrnieme do odhadu zaokrúhlenie ako tváriť sa, že náš výber je nezaokrúhlený.

Predpokladajme, že $X = (X_1, \cdots, X_n)$ je iid výber zo spojitého rozdelenia $F(\theta|x)$ s hustotou $f(\theta|x)$. Potom vierohodnostnou funkciou nazveme funkciu $L(\theta|x) = \prod_{i=1}^{n} f(\theta|x_i)$. Odhadom parametra θ pomocou ML sa označuje $\hat{\theta}$ a je daná ako riešenie problému $\hat{\theta} = \arg\max_{\theta \in \Theta} L(\theta|x)$. Často je výhodnejšie

pracovať s funkciou $\ell(\theta|x) = \ln L(\theta|x)$. ML odhad $\hat{\theta}$ sa spravidla stanovuje ako riešenie rovnice

$$\frac{\partial \ell(\theta|x)}{\partial \theta} = 0$$

Vieme, že ak sú splnené niektoré podmienky regulárnosti, tak ML odhad $\hat{\theta}$ je jediný a asymptotický rovný skutočnej hodnote parametra θ.

Podobne sa dá ML použiť aj na zaokrúhlené dáta. Bez ujmy na všeobecnosti nech opäť zaokrúhľujeme na celé čísla. Keďže zaokrúhľovanie nám zo spojitého rozdelenia spravilo diskrétne, nemôžeme do funkcie $L(\theta|x)$ dosadiť hustoty, miesto nich tam dosadíme pravdepodobnosti, že naše pozorované dáta sú s konkrétneho rozdelenia.

$$\begin{aligned} L(\theta|x) &= \prod_{i=1}^{n} P(X_i = x_i) \\ &= \prod_{i=1}^{n} [F(\theta|x_i + 0.5) - F(\theta|x_i - 0.5)] \\ &= \prod_{j=-\infty}^{\infty} [F(\theta|j + 0.5) - F(\theta|j - 0.5)]^{n_j}, \end{aligned}$$

kde n_j je počet výskytov čísla j vo výbere $x_1, \cdots, x_n$. Najvýhodnejšie je zrejme používať tretí zápis, keďže rôznych pozorovaných hodnôt (j) je zrejme oveľa menej ako všetkých pozorovaní (x_j), ale v niektorých prípadoch je nutné používať druhý zápis, napríklad v lineárnej regresii (podkapitola 2.4).

Často je výhodnejšie pracovať s logaritmickou funkciou vierohodnosti

$$\ell(\theta|x) = \ln L(\theta|x) = \sum_{i=-\infty}^{\infty} n_i \cdot \ln(p_i),$$

kde $p_i = p_i(\theta) = F(\theta|i+0.5) - F(\theta|i-0.5)$. Úloha sa nám teda zredukovala na hľadanie $\hat{\theta} = \underset{\theta \in \Theta}{\arg\max}\ \ell(\theta|x)$. Z už vyššie spomenutého vieme, že ak existuje parciálna derivácia $\ell(\theta|x)$, tak $\hat{\theta}$ je dané ako riešenie rovnice

$$\frac{\partial \ell(\theta|x)}{\partial \theta} = \sum_{i=-\infty}^{\infty} \frac{n_i}{p_i} \cdot \frac{\partial p_i}{\partial \theta} = 0.$$

2.1 Odhad parametra λ exponenciálneho rozdelenia

Exponenciálne rozdelenie je príkladom spojitého rozdelenia pravdepodobnosti. Hovoríme, že spojitá náhodná premenná X má exponenciálne rozdelenie s parametrom $\lambda > 0$, ak jej hustota pravdepodobnosti je

$$f(x) = \begin{cases} \lambda e^{-\lambda x} & ; x \geq 0 \\ 0 & ; x < 0. \end{cases}$$

Jej distribučná funkcia je daná ako:

$$F(x) = \begin{cases} 1 - e^{-\lambda x} & ; x \geq 0 \\ 0 & ; x < 0. \end{cases}$$

Označujeme:

$$X \sim Exp(\lambda).$$

Urobíme ML odhad pre nezaokrúhlený výber $x_1, x_2, \cdots, x_n$ z exponenciálneho rozdelenia.

$$L(\lambda|x) = \lambda^n \cdot e^{-\lambda \cdot \sum_{i=1}^{n} x_i} = \lambda^n \cdot e^{-\lambda \cdot n \cdot \bar{x}}$$

$$\ell(\lambda|x) = \ln L(\lambda|x) = n \cdot \ln \lambda - n \cdot \lambda \cdot \bar{x}$$

λ_{ML} spĺňa rovnicu $\frac{\partial \ell(\lambda|x)}{\partial \lambda} = 0$, teda platí:

$$\frac{n}{\lambda_{ML}} - n \cdot \bar{x} = 0 \Rightarrow \lambda_{ML} = \frac{1}{\bar{x}}.$$

Označme si náhodný výber $x_1, x_2 \cdots, x_n$ zaokrúhlený na celé čísla ako $y_1, y_2, \cdots, y_n$. Ak nezohľadníme ich zaokrúhlenie, ML odhad pre λ bude rovný $\frac{1}{\bar{y}}$, označme si tento odhad ako $\tilde{\lambda}$.

Keď zoberieme do úvahy zaokrúhľovanie, tak odhad sa zmení, ako je prezentované v [1]. Označme si ML odhad zohľadňujúci zaokrúhľovanie pre exponenciálne rozdelenie ako $\hat{\lambda}$. Keď upravíme vzťah spomínaný v predchádzajúcej časti pre potreby exponenciálneho rozdelenia, dostaneme

$$(\star)\ \ \frac{\partial \ell(\lambda|y)}{\partial \lambda} = \sum_{i=0}^{\infty} \frac{n_i}{p_i} \cdot \frac{\partial p_i}{\partial \lambda} = 0,$$

kde

$$p_i = p_i(\lambda) = F\left(\lambda|i + \frac{h}{2}\right) - F\left(\lambda|i - \frac{h}{2}\right) = \begin{cases} e^{-\lambda i} \cdot \left(e^{\frac{h\cdot\lambda}{2}} - e^{-\frac{h\cdot\lambda}{2}}\right) & ; i \neq 0 \\ 1 - e^{-\frac{h\cdot\lambda}{2}} & ; i = 0. \end{cases}$$

Teraz už len potrebujeme z $(\star)$ vyjadriť λ:

$$\frac{\partial p_i}{\partial \lambda} = \begin{cases} e^{-\lambda i} \cdot \left[-i \cdot \left(e^{\frac{h\cdot\lambda}{2}} - e^{-\frac{h\cdot\lambda}{2}}\right) + \frac{h}{2} \cdot \left(e^{\frac{h\cdot\lambda}{2}} + e^{-\frac{h\cdot\lambda}{2}}\right)\right] & ; i \neq 0 \\ \frac{h}{2} \cdot e^{-\frac{h\cdot\lambda}{2}} & ; i = 0 \end{cases}$$

$$\sum_{i=1}^{\infty} \frac{n_i \cdot e^{-\lambda i} \cdot \left[-i \cdot \left(e^{\frac{h\cdot\lambda}{2}} - e^{-\frac{h\cdot\lambda}{2}}\right) + \frac{h}{2} \cdot \left(e^{\frac{h\cdot\lambda}{2}} + e^{-\frac{h\cdot\lambda}{2}}\right)\right]}{e^{-\lambda i} \cdot \left(e^{\frac{h\cdot\lambda}{2}} - e^{-\frac{h\cdot\lambda}{2}}\right)} + \frac{h \cdot n_0}{2} \cdot \left(\frac{e^{-\frac{h\cdot\lambda}{2}}}{1 - e^{-\frac{h\cdot\lambda}{2}}}\right) = 0$$

$$\sum_{i=1}^{\infty} n_i \cdot \left(-i + \frac{h}{2} \cdot \frac{e^{\frac{h\cdot\lambda}{2}} + e^{-\frac{h\cdot\lambda}{2}}}{e^{\frac{h\cdot\lambda}{2}} - e^{-\frac{h\cdot\lambda}{2}}}\right) + \frac{h \cdot n_0}{2} \cdot \left(\frac{e^{-\frac{h\cdot\lambda}{2}} - 1 + 1}{1 - e^{-\frac{h\cdot\lambda}{2}}}\right) = 0$$

$$-\sum_{i=1}^{n} x_i + \frac{h \cdot (n - n_0)}{2} \cdot \frac{e^{h\cdot\lambda} + 1}{e^{h\cdot\lambda} - 1} - \frac{h \cdot n_0}{2} + \frac{h \cdot n_0}{2} \cdot \frac{1}{1 - e^{-\frac{h\cdot\lambda}{2}}} = 0.$$

Pre ľahšie počítanie použijeme substitúciu $e^{\frac{h\cdot\lambda}{2}} = k$.

$$-2 \cdot n \cdot \bar{y} - h \cdot n_0 + h \cdot (n - n_0) \cdot \left(\frac{k^2 + 1}{k^2 - 1}\right) + h \cdot n_0 \cdot \frac{1}{1 - \frac{1}{k}} = 0$$

$$\frac{2 \cdot h \cdot (n - n_0)}{k^2 - 1} + h \cdot n_0 \cdot \frac{k}{k - 1} - 2 \cdot n \cdot \bar{y} - h \cdot 2 \cdot n_0 + h \cdot n = 0$$

$$2 \cdot h \cdot (n - n_0) + h \cdot n_0 \cdot k \cdot (k + 1) - (k^2 - 1) \cdot (2 \cdot n \cdot \bar{y} + 2 \cdot h \cdot n_0 - h \cdot n) = 0$$

$$k^2 \cdot (-2 \cdot n \cdot \bar{y} - h \cdot n_0 + h \cdot n) + k \cdot h \cdot n_0 + h \cdot n + 2 \cdot n \cdot \bar{y} = 0$$

Dostali sme sa k obyčajnej kvadratickej rovnici, ktorú vieme ľahko vyriešiť.

$$D = h^2 \cdot n_0^2 + 4 \cdot (h \cdot n + 2 \cdot n \cdot \bar{y}) \cdot (2 \cdot n \cdot \bar{y} + h \cdot n_0 - h \cdot n)$$

Keďže k musí byť kladné, do úvahy pripadá iba riešenie

$$k = \frac{h \cdot n_0 + \sqrt{D}}{2 \cdot (2 \cdot n \cdot \bar{y} + h \cdot n_0 - h \cdot n)} = e^{\frac{h \cdot \lambda}{2}}.$$

A teda sme získali odhad

$$\tilde{\lambda} = \frac{2}{h} \cdot \ln \left(\frac{h \cdot n_0 + \sqrt{h^2 \cdot n_0^2 + 4 \cdot (h \cdot n + 2 \cdot n \cdot \bar{y}) \cdot (2 \cdot n \cdot \bar{y} + h \cdot n_0 - h \cdot n)}}{2 \cdot (2 \cdot n \cdot \bar{y} + h \cdot n_0 - h \cdot n)} \right).$$

Ak položíme $h = 0$, dostaneme prípad bez zaokrúhľovania, čiže spojité rozdelenie, pre ktoré platí $\hat{\lambda} = \frac{1}{\bar{y}}$. Ak je náš odhad, zohľadňujúci zaokrúhľovanie dobrý, požadujeme aby platilo

$$\lim_{h \to 0} \tilde{\lambda} = \hat{\lambda} = \frac{1}{\bar{y}}.$$

Jednoduchým dosadením 0 za h sa dostaneme k podielu $\frac{0}{0}$, čiže musíme použiť L´Hospitalovo pravidlo a zderivovať čitateľa aj menovateľa podľa h

$$\lim_{h \to 0} \tilde{\lambda} = \lim_{h \to 0} \frac{4 \cdot (2 \cdot n \cdot \bar{y} + h \cdot n_0 - h \cdot n)}{h \cdot n_0 + \sqrt{D}} \cdot$$

$$\cdot \left(\frac{\left(2 \cdot n_0 + \frac{1}{\sqrt{D}} \cdot (2 \cdot n_0^2 \cdot h + 8 \cdot (n \cdot n_0 \cdot \bar{y} + n \cdot n_0 \cdot h - n^2 \cdot h))\right) \cdot (2 \cdot n \cdot \bar{y} + h \cdot n_0 - h \cdot n)}{4 \cdot (2 \cdot n \cdot \bar{y} + h \cdot n_0 - h \cdot n)^2} - \right.$$

$$\left. - \frac{2 \cdot (n_0 - n) \cdot (h \cdot n_0 + \sqrt{D})}{4 \cdot (2 \cdot n \cdot \bar{y} + h \cdot n_0 - h \cdot n)^2} \right)$$

Ak $h \to 0$ potom aj $n_0 \to 0$, keďže pre spojité rozdelenie je nulová pravdepodobnosť, že y_i bude presne 0. Po dosadení dostaneme

$$\lim_{h \to 0} \tilde{\lambda} = \frac{8 \cdot n \cdot \bar{y}}{4 \cdot n \cdot \bar{y}} \cdot \frac{8 \cdot n^2 \cdot \bar{y}}{16 \cdot n^2 \cdot \bar{y}^2} = \frac{1}{\bar{y}}.$$

Čo sme presne chceli ukázať.

Simulácia 2.1 (Porovnanie správania sa odhadov pre λ)

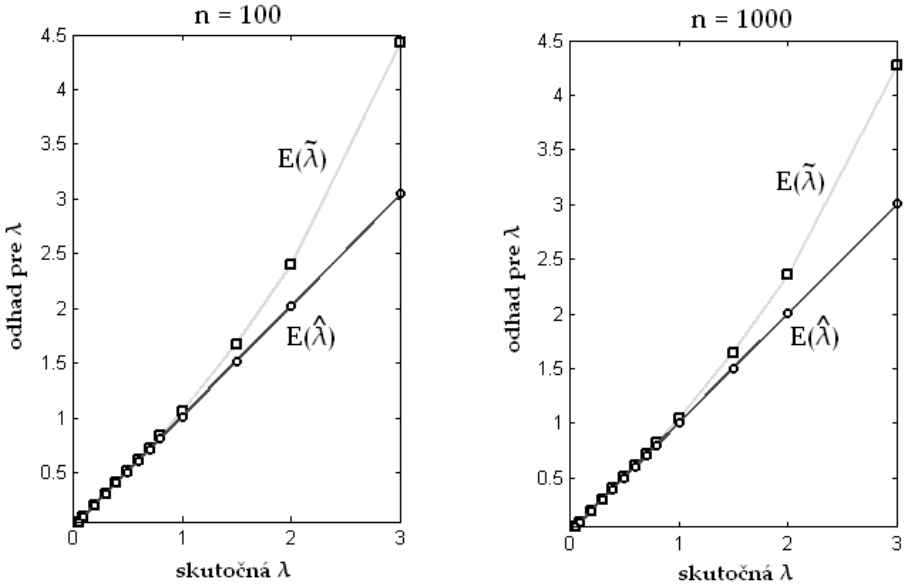
V nasledovnej simulácii sa pokúsime ukázať, že $\tilde{\lambda}$ je lepší odhad pre λ ako $\hat{\lambda}$. Pre rôzne hodnoty λ sme vygenerovali výber $x_1, \cdots, x_n \sim Exp(\lambda)$ a zaokrúhlili ho na celé čísla, čím sme dostali výber $y_1, \cdots, y_n$ zo zaokrúhleného exponenciálneho rozdelenia. Pre vypočítanie odhadu pre λ sme použili dva spôsoby uvedené v tejto kapitole, $\tilde{\lambda}$ je ML odhad, kde sa tvárime akoby sme mali ozajstný výber z exponenciálneho rozdelenia a $\hat{\lambda}$ je ML odhad zohľadňujúci zaokrúhlenie, vypočítaný podľa postupu v článku [1]. Tieto odhady sme vypočítali ako

$$\hat{\lambda} = \frac{1}{\bar{y}}$$

$$\tilde{\lambda} = \frac{2}{h} \cdot \ln\left(\frac{h \cdot n_0 + \sqrt{h^2 \cdot n_0^2 + 4 \cdot (h \cdot n + 2 \cdot n \cdot \bar{y}) \cdot (2 \cdot n \cdot \bar{y} + h \cdot n_0 - h \cdot n)}}{2 \cdot (2 \cdot n \cdot \bar{y} + h \cdot n_0 - h \cdot n)}\right)$$

V našej simulácii sme zaokrúhľovali na celé čísla, teda $h = 1$ a používali sme veľkosti náhodného výberu $n = 100$ a $n = 1000$. Za λ sme postupne volili hodnoty 0.05, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 1, 1.5, 2, 3. Danú simuláciu sme opakovali 100000-krát, aby sme zabezpečili jej významnosť. Keďže odhady $\hat{\lambda}$ a $\tilde{\lambda}$ sú vypočítané zo zaokrúhleného výberu z náhodnej premennej X a tá závisí od hodnoty skutočného λ, dá sa povedať, že $E(\hat{\lambda})$ a $E(\tilde{\lambda})$, ktoré chceme určiť, sú funkcie závislé od λ. Výsledky sú ilustrované na nasledujúcich obrázkoch, tmavšou farbou s krúžkami ako pozorovanými bodmi je znázornený ML odhad, zohľadňujúci zaokrúhľovanie, bledšou farbou so štvorcami odhad nezohľadňujúci zaokrúhľovanie.

Obr. 2.1: Odhady pre λ v exponenciálnom rozdelení



Ako môžeme vidieť na obrázkoch, ML odhad zohľadňujúci zaokrúhľovanie dáva lepšie výsledky ako klasický ML odhad pre všetky hodnoty λ. So zväčšujúcimi sa hodnotami λ, a teda so zmenšujúcimi sa hodnotami $\bar{y}$ sa odhad $\tilde{\lambda}$ čoraz viac odchyľuje od skutočnej hodnoty λ. Naopak zaokrúhlený ML odhad $\hat{\lambda}$ dáva výborné odhady pre skoro ľubovoľnú hodnotu λ. Priamka $\lambda = \lambda$ je úplne prekrytá grafom získaným spojením odhadov, vypočítaných pomocou ML zohľadňujúceho zaokrúhľovanie. Je teda celkom zjavné, že tento spôsob nám zlepší odhadovanie parametra λ v exponenciálnom rozdelení.

2.2 Odhad parametra μ normálneho rozdelenia pri známej hodnote σ

V tejto časti uvedieme spôsob pre odhad parametra μ normálneho rozdelenia so zohľadnením zaokrúhľovania podľa návodu v [1] a porovnáme ho so

známym odhadom pre parameter μ nezaokrúhleného normálneho rozdelenia, čiže s $\bar{x}$. Pre jednoduchosť budeme predpokladať, že parameter σ je známy.

Normálne rozdelenie je príkladom spojitého rozdelenia pravdepodobnosti. Hovoríme, že spojitá náhodná premenná X má normálne rozdenie s parametrami μ, σ^2, ak jej hustota pravdepodobnosti je

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \cdot e^{\frac{-(x-\mu)^2}{2\sigma^2}} = \frac{1}{\sigma} \cdot \phi\left(\frac{x-\mu}{\sigma}\right).$$

Jej distribučná funkcia je daná ako:

$$F(x) = \Phi\left(\frac{x-\mu}{\sigma}\right) = \int_{-\infty}^{\frac{x-\mu}{\sigma}} e^{\frac{-t^2}{2}} dt.$$

Označujeme:

$$X \sim \mathcal{N}(\mu, \sigma^2).$$

Uvažujme z neho náhodný výber $x_1, x_2 \cdots, x_n$ a ich zaokrúhlené hodnoty $y_1, y_2 \cdots, y_n$, zaokrúhlené na celé čísla. ML odhad parametra μ nezohľadňujúci zaokrúhľovanie pre normálne rozdelenie sme už prezentovali v kapitole 1, dá sa ukázať, že je rovný $\bar{y} = \frac{1}{n} \cdot \sum y_i$ a nezávisí od hodnoty σ. ML odhad pre μ zohľadňujúci zaokrúhľovanie maximalizuje logaritmickú funkciu vierohodnosti, ktorou je v tomto prípade

$$\ell(\mu|y) = \sum_{i=-\infty}^{\infty} n_i \cdot \ln(p_i) = \sum_{i=-\infty}^{\infty} n_i \cdot \ln\left(F\left(i+\frac{h}{2}\right) - F\left(i-\frac{h}{2}\right)\right) =$$

$$= \sum_{i=-\infty}^{\infty} n_i \cdot \ln\left(\Phi\left(\frac{i+\frac{h}{2}-\mu}{\sigma}\right) - \Phi\left(\frac{i-\frac{h}{2}-\mu}{\sigma}\right)\right).$$

Keďže jediná neznáma tu je μ, tak tento ML odhad je daný riešením rovnice

$$\frac{\partial \ell(\mu|y)}{\partial \mu} = \sum_{i=-\infty}^{\infty} \frac{n_i}{p_i} \cdot \frac{\partial p_i}{\partial \mu} = 0,$$

kde

$$p_i = p_i(\mu) = F\left(\mu, \sigma | i + \frac{h}{2}\right) - F\left(\mu, \sigma | i - \frac{h}{2}\right) = \Phi\left(\frac{i + \frac{h}{2} - \mu}{\sigma}\right) - \Phi\left(\frac{i - \frac{h}{2} - \mu}{\sigma}\right).$$

Na výpočet $\frac{\partial p_i}{\partial \mu}$ použijeme znalosti z matematickej analýzy, konkrétne vzorec

$$(1) \quad \frac{d}{dx} \int_{\varphi(x)}^{\psi(x)} f(t)dt = \frac{d\psi(x)}{dx} \cdot f(\psi(x)) - \frac{d\varphi(x)}{dx} \cdot f(\varphi(x)),$$

uvedený napríklad v [4], pomocou ktorého vieme vypočítať, že

$$\frac{\partial p_i}{\partial \mu} = -\frac{\sqrt{2\pi}}{\sigma} \cdot \left(\phi\left(\frac{i + \frac{h}{2} - \mu}{\sigma}\right) - \phi\left(\frac{i - \frac{h}{2} - \mu}{\sigma}\right)\right)$$

Násobenie konštantou $-\frac{\sqrt{2\pi}}{\sigma}$ je nepodstatné, pretože ho môžeme vyňať pred sumu a zároveň vieme, že sa nerovná 0 a preto týmto členom vieme vydeliť celú rovnicu. Dosadením vypočítaných členov do pôvodnej rovnice dostávame

$$\sum_{i=-\infty}^{\infty} n_i \cdot \left(\frac{\phi\left(\frac{i+\frac{h}{2}-\mu}{\sigma}\right) - \phi\left(\frac{i-\frac{h}{2}-\mu}{\sigma}\right)}{\Phi\left(\frac{1+\frac{h}{2}-\mu}{\sigma}\right) - \Phi\left(\frac{1-\frac{h}{2}-\mu}{\sigma}\right)}\right) = 0$$

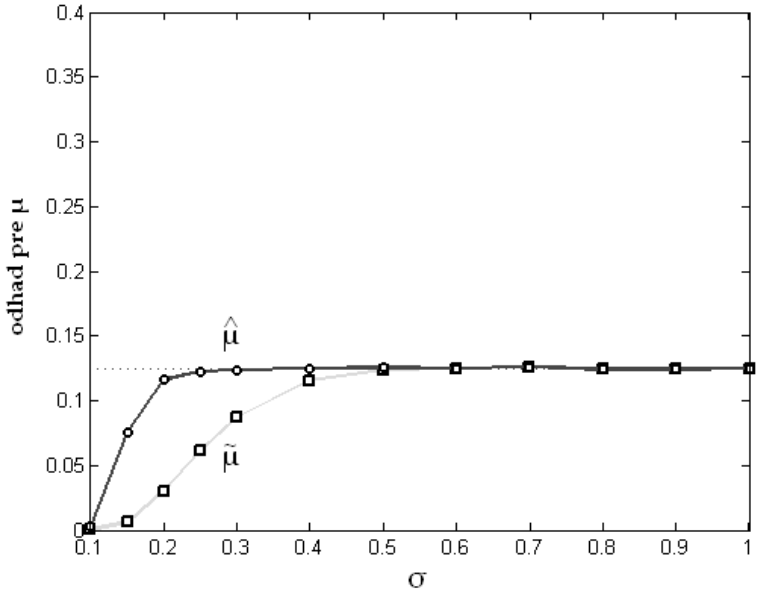
V tejto rovnici ale vystupujú aj iné ako elementárne funkcie, preto bude veľký problém vyjadriť z tejto rovnice μ. Ale nejakými numerickými metódami by sa dala odhadnúť jeho hodnota s ľubovoľnou presnosťou. Jednou z najznámejších a najjednoduchších je Newtonova metóda dotyčníc, ktorá hľadá riešenie rovnice $f(x) = 0$ pomocou iteračného predpisu $x_{n+1} = x_n - \frac{f(x)}{f'(x)}$ o ktorom vieme, že veľmi rýchlo konverguje, ak začneme dostatočne blízko riešenia. Pri viacrozmerných úlohách sa používa miesto $f'(x)$ matica parciálnych derivácii funkcie f, čiže Jakobián. Iteračná schéma sa zmení na $x_{n+1} = x_n - J_n^{-1} \cdot f(x)$. Táto metóda je síce najjednoduchšia, ale môže nastať problém, ak začiatočný bod je príliš ďaleko od skutočného riešenia alebo $f'(x)$ je blízke 0 resp.

pri viacrozmerných prípadoch je Jakobián veľmi blízky singulárnej matici. Preto sa väčšinou používajú zložitejšie algoritmy, ktoré tieto problémy dokážu obísť. Najjednoduchšie je použiť funkcie na výpočet sústav nelineárnych rovníc v nejakom počítačovom programe, my sme používali MATLAB a jeho funkciu *fsolve*. Tá má na riešenie sústavy nelineárnych rovníc zabudovaných niekoľko algoritmov, pričom základným je tzv. " Trust-Region Dogleg " algoritmus. Bližšie informácie o tomto a ďalších podobných algoritmoch sa dajú nájsť napríklad v [5]. Výhodou je taktiež fakt, že MATLAB si vie aj sám zrátať derivácie v Jakobiáne potrebné v tomto algoritme numericky z definície derivácie. Tým nám odbudne veľa práce, keďže derivácie vyjdu dosť zložité. V nasledujúcej simulácii sa prezentovaným postupom pokúsime zlepšiť odhad pre μ.

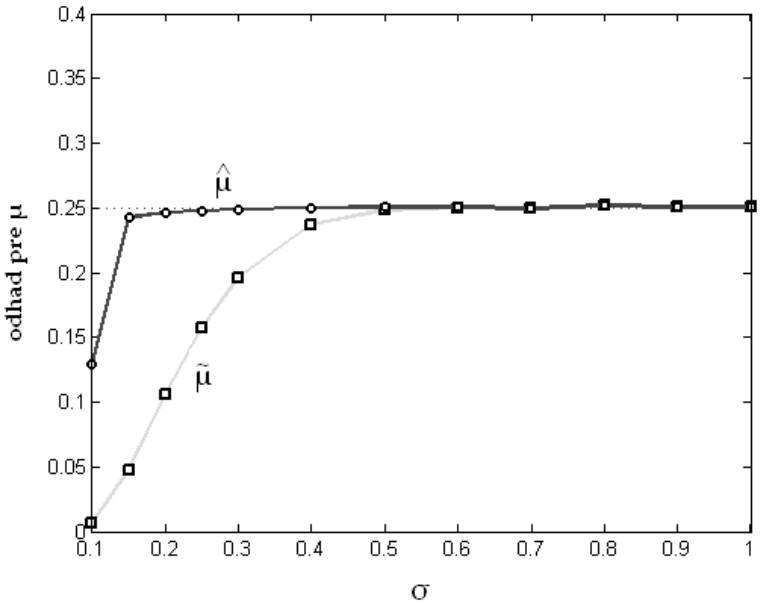
Simulácia 2.2 (Správanie sa odhadov pre μ pri známej hodnote σ) *V tejto simulácii sa pokúsime zistiť, či nám ML odhad μ zohľadňujúci zaokrúhľovanie, jeho strednú hodnotu (odhadnutú ako priemer z* 10000 *pozorovaní) si označme ako $\hat{\mu}$ dá lepší výsledok ako odhad μ pre nezaokrúhlené dáta (výberový priemer), jeho strednú hodnotu, odhadnutú z* 10000 *pozorovaní si označme ako $\tilde{\mu}$. Táto simulácia bude podobná simulácii* 1.1*, pretože sa budeme zaujímať o odhad hodnoty μ. Rozdiel bude v tom, že popri obyčajnom výberovom priemere odhadneme μ aj pomocou vyššie popísaného postupu na získanie ML odhadu. Pre zjednodušenie už nesledujeme správanie sa odhadov pre μ vzhľadom na hodnotu n, ktorá nemá skoro žiadny vplyv na hodnotu odhadov pre μ, ale používame iba hodnotu $n = 100$. Ako hodnoty σ, ktoré sme brali ako dopredu známe sme používali postupne hodnoty* $0.1, 0.15, 0.2, 0.25, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9$ *a* 1*. A keďže sa ukázalo, že zaokrúhľovanie nemá veľký vplyv na odhady pre μ ak $\mu = 0$ alebo $\mu = 0.5$, tak ako hodnoty μ používame hodnoty* $0.125, 0.25$ *a* 0.375*, ku každej z nich prikladáme obrázok. Na každom obrázku možno vidieť* 2 *grafy, tmavší z nich s krúžkami znázorňujúcimi pozorované body je graf pre odhad $\hat{\mu}$, získaný pomo-*

cou ML odhadu zohľadňujúceho zaokrúhľovanie a bledší graf so štvorčekami je graf pre odhad pomocou výberového priemeru, čiže nezahŕňa zaokrúhľovanie.

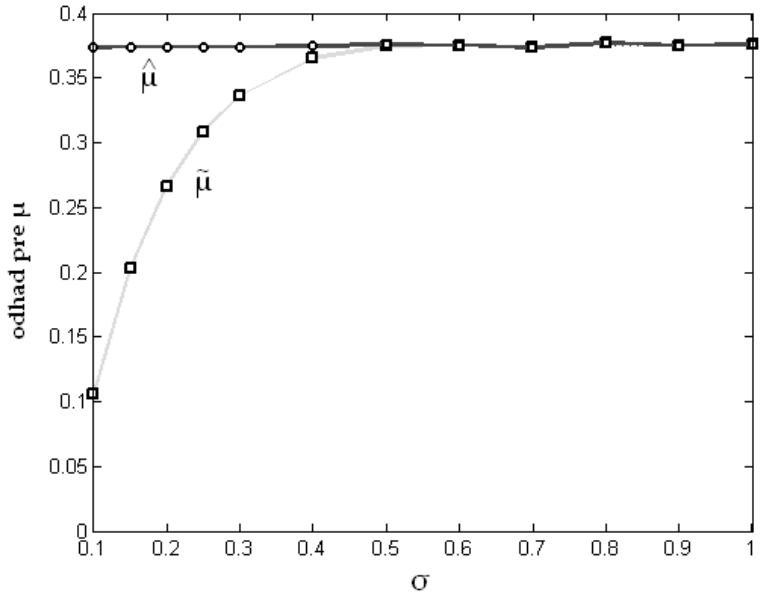
Obr. 2.2: Porovnanie odhadov pre μ ak skutočné $\mu = 0.125$



Obr. 2.3: Porovnanie odhadov pre μ ak skutočné $\mu = 0.25$



Obr. 2.4: Porovnanie odhadov pre μ ak skutočné $\mu = 0.375$



Môžeme si všimnúť že $\hat{\mu}$ je oproti $\tilde{\mu}$ bližšie ku skutočnej hodnote μ vo všetkých prípadoch. V tomto prípade predstavuje ML odhad výrazné zlepšenie oproti výberovému priemeru. Vidíme, že pre $\mu = 0.125$ a $\sigma < 0.2$ nedáva ML odhad až také presné výsledky, ale to je pochopiteľné, keďže sú to prípady, kde pozorované hodnoty sú takmer samé 0. Z toho istého dôvodu môžeme pri $\mu = 0.25$ pozorovať niečo podobné, avšak iba pri hodnotách $\sigma < 0.15$. Pri $\mu = 0.375$ je už hodnota $\sigma = 0.1$ dostatočne veľká na to, aby ML odhad dával celkom presné výsledky.

2.3 Odhad parametra μ a σ normálneho rozdelenia pri neznámej hodnote σ

V tejto časti budeme vychádzať z rovnakých rovníc ako v podkapitole 2.2, iba σ nebudeme brať ako známu konštantu, ale ako ďalšiu neznámu premennú, budeme sa teda snažiť odhadovať 2 parametre naraz, μ aj σ. Snažíme sa maximalizovať tú istú logaritmickú funkciu vierohodnosti

$$\ell(\mu, \sigma|y) = \sum_{i=-\infty}^{\infty} n_i \cdot \ln(p_i) = \sum_{i=-\infty}^{\infty} n_i \cdot \ln\left(F\left(i + \frac{h}{2}\right) - F\left(i - \frac{h}{2}\right)\right) =$$

$$= \sum_{i=-\infty}^{\infty} n_i \cdot \ln\left(\Phi\left(\frac{i + \frac{h}{2} - \mu}{\sigma}\right) - \Phi\left(\frac{i - \frac{h}{2} - \mu}{\sigma}\right)\right).$$

Keďže máme už 2 neznáme, riešenie maximalizačného problému musí spĺňať podmienky

$$\frac{\partial \ell(\mu, \sigma|y)}{\partial \mu} = \sum_{i=-\infty}^{\infty} \frac{n_i}{p_i} \cdot \frac{\partial p_i}{\partial \mu} = 0$$

$$\frac{\partial \ell(\mu, \sigma|y)}{\partial \sigma} = \sum_{i=-\infty}^{\infty} \frac{n_i}{p_i} \cdot \frac{\partial p_i}{\partial \sigma} = 0.$$

p_i aj $\frac{\partial p_i}{\partial \mu}$ už poznáme z podkapitoly 2.2, $\frac{\partial p_i}{\partial \sigma}$ vypočítame pomocou vzorca (1) ako

$$\frac{\partial p_i}{\partial \sigma} = \frac{\sqrt{2\pi}}{\sigma^2} \cdot \left(\left(i + \frac{h}{2} - \mu \right) \cdot \phi \left(\frac{i + \frac{h}{2} - \mu}{\sigma} \right) - \left(i - \frac{h}{2} - \mu \right) \cdot \phi \left(\frac{i - \frac{h}{2} - \mu}{\sigma} \right) \right).$$

Dosadením týchto výrazov do maximalizačných rovníc dostávame sústavu 2 rovníc o 2 neznámych:

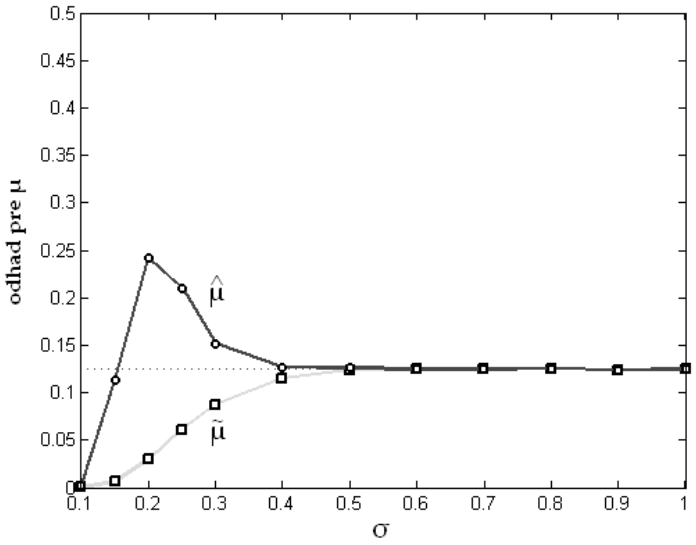
$$\sum_{i=-\infty}^{\infty} n_i \cdot \left(\frac{\phi\left(\frac{i+\frac{h}{2}-\mu}{\sigma}\right) - \phi\left(\frac{i-\frac{h}{2}-\mu}{\sigma}\right)}{\Phi\left(\frac{i+\frac{h}{2}-\mu}{\sigma}\right) - \Phi\left(\frac{i-\frac{h}{2}-\mu}{\sigma}\right)} \right) = 0$$

$$\sum_{i=-\infty}^{\infty} n_i \cdot \left(\frac{(i+\frac{h}{2}-\mu) \cdot \phi\left(\frac{i+\frac{h}{2}-\mu}{\sigma}\right) - (i-\frac{h}{2}-\mu) \cdot \phi\left(\frac{i-\frac{h}{2}-\mu}{\sigma}\right)}{\Phi\left(\frac{i+\frac{h}{2}-\mu}{\sigma}\right) - \Phi\left(\frac{1-\frac{h}{2}-\mu}{\sigma}\right)} \right) = 0$$

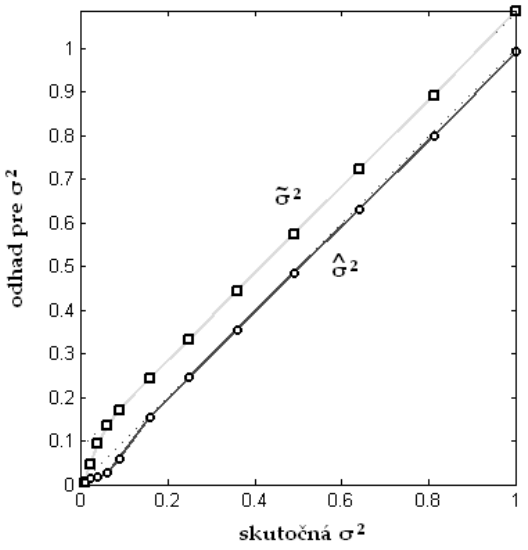
Túto sústavu vyriešime rovakým spôsobom ako v podkapitole 2.2, akurát zvýšime jej rozmer na 2.

Simulácia 2.3 (Správanie sa odhadov pre μ a σ pri neznámej hodnote σ) *V tejto simulácii sme postupovali presne tak isto ako v simulacii* 2.2*, akurát sme odhadovali aj parameter σ spoločne s parametrom μ pomocou riešenia sústavy rovníc z podkapitoly* 2.3*. Strednú hodnotu ML odhadov, odhadnutú ako priemer z* 10000 *pozorovaní pre μ a σ^2 si označme ako $\hat{\mu}$ resp. $\hat{\sigma}^2$. Strednú hodnotu, odhadnutú ako priemer z* 10000 *pozorovaní výberového priemeru ($\frac{1}{n} \cdot \sum y_i$) a výberovej disperzie ($\frac{1}{n-1} \cdot \sum (y_i - \bar{y})^2$) si označme ako $\tilde{\mu}$ resp. $\tilde{\sigma}^2$. Rovnako ako v simulácii* 2.2 *používame hodnoty $n = 100$, $\sigma = 0.1, 0.15, 0.2, 0.25, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9$ a 1 a $\mu = 0.125, 0.25$ a* 0.375*. Ku každej hodnote μ prikladáme* 2 *obrázky, jeden ukazuje odhady pre μ a druhý odhady pre σ^2.*

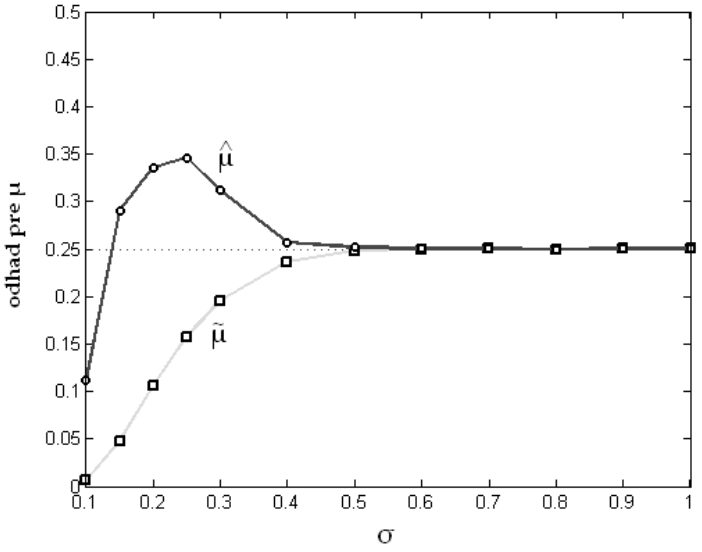
Obr. 2.5: Porovnanie odhadov pre μ ak skutočné $\mu = 0.125$



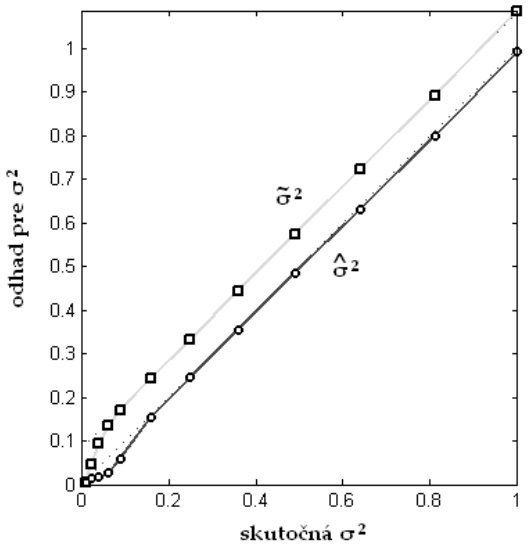
Obr. 2.6: Porovnanie odhadov pre σ^2 ak skutočné $\mu = 0.125$



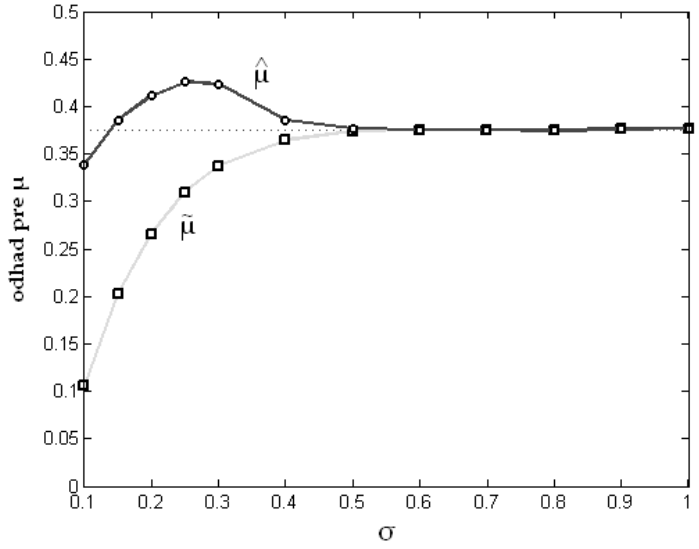
Obr. 2.7: Porovnanie odhadov pre μ ak skutočné $\mu = 0.25$



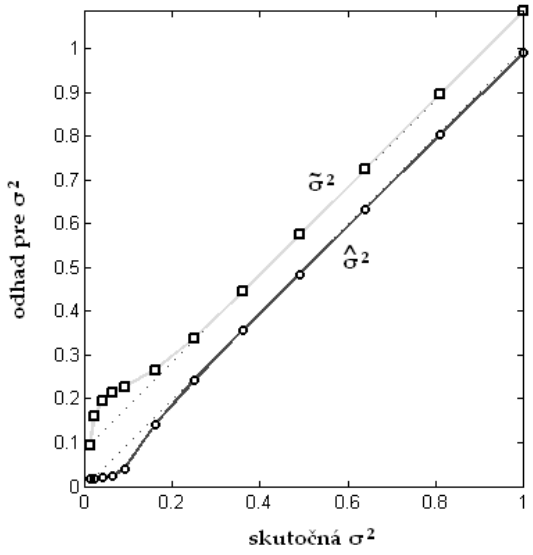
Obr. 2.8: Porovnanie odhadov pre σ^2 ak skutočné $\mu = 0.25$



Obr. 2.9: Porovnanie odhadov pre μ ak skutočné $\mu = 0.375$



Obr. 2.10: Porovnanie odhadov pre σ^2 ak skutočné $\mu = 0.375$



Ako si môžeme všimnúť, vyskytol sa tu jeden problém. Pre menšie hodnoty σ, zhruba okolo 0.2 a 0.3 dochádza k zaujímavému fenoménu, ML odhad $\hat{\mu}$ pre μ je výrazne nadhodnotený a naopak ML odhad $\hat{\sigma}^2$ pre σ^2 je výrazne podhodnotený. Dôvod je však celkom jednoduchý. Pre malé hodnoty σ vyzerá zaokrúhlený náhodný výber v drvivej väčšine prípadov ako veľa núl a semtam nejaká jednotka. Tento výber mohol vzniknúť z normálneho rozdelenia so strednou hodnotou o kúsok menšou ako 0.5 a minimálnou disperziou tak isto ako aj z našich hodnôt μ a σ, pričom najpravdepodobnejšie sú hodnoty, ktoré môžeme vidieť na obrázkoch a sú niekde medzi spomínanými dvoma prípadmi. Pre väčšie hodnoty σ je už ML odhad $\hat{\mu}$ pre μ dobrý, lenže v tých prípadoch už aj odhad nezohľadňujúci zaokrúhľovanie dáva pomerne dobré výsledky, čiže nedá sa úplne jednoznačne povedať, či sme si v prípade odhadovania μ pomohli. Výrazne sa ale zlepšil ML odhad $\hat{\sigma}^2$ pre σ^2 pri väčších hodnotách σ^2, ktorý odstraňuje potrebu korekcie $-\frac{1}{12}$ a dáva podstatne bližšie odhady k skutočnej hodnote σ^2 ako $\tilde{\sigma}^2$.

2.4 Odhad parametrov α, β a σ v lineárnej regresii

V tejto časti sa zameriame na lineárny regresný model, čiže model $X = \alpha + \beta \cdot s + \varepsilon,\ \varepsilon \sim N(0, \sigma^2)$. Označme si zaokrúhlené hodnoty X ako Y. V tomto modeli máme až 3 neznáme parametre: α, β a σ. Pre jednoduchšie počítanie si ako vektor s zvoľme aritmetickú postupnosť $0, 1, 2, \cdots, n-1$, kde n je počet pozorovaní, resp. rozmer X. Premenná x_i sa správa ako náhodná premenná so strednou hodnotou $\alpha + (i-1) \cdot \beta$ a disperziou σ^2 a y_i je jej hodnota zaokrúhlená na celé čísla. Je problém vypočítať funkciu vierohodnosti pomocou prechodu cez všetky možné hodnoty k (vzorec $\sum_{k=-\infty}^{\infty} n_k \cdot \ln[F(\theta|k+0.5) - F(\theta|k-0.5)]$), pretože z toho, že pre nejaké neznáme

i platí, že $y_i = k$ nevieme určiť, z akého s_i bola vypočítaná daná hodnota, čiže nevieme aká je stredná hodnota rozdelenia z ktorého pochádza dané pozorovanie a tým pádom nevieme vyčísliť hodnotu p_i. Preto budeme musieť prechádzať cez všetky pozorovania (vzorec $\sum_{i=1}^{n} \ln[F(\theta|y_i+0.5)-F(\theta|y_i-0.5)]$), ak poznáme i, vieme, že y_i je zaokrúhlená hodnota merania x_i, ktoré má normálne rozdelenie so strednou hodnotou $\alpha + (i-1)\cdot\beta$. To znamená, že logaritmická funkcia vierohodnosti bude vyzerať nasledovne

$$\ell(\alpha,\beta,\sigma|y) = \sum_{i=1}^{n} \ln(p_{y_i}) = \sum_{i=1}^{n} \ln\left(F\left(y_i + \frac{h}{2}\right) - F\left(y_i - \frac{h}{2}\right)\right) =$$

$$= \sum_{i=1}^{n} \ln\left(\Phi\left(\frac{y_i + \frac{h}{2} - \alpha - (i-1)\cdot\beta}{\sigma}\right) - \Phi\left(\frac{y_i - \frac{h}{2} - \alpha - (i-1)\cdot\beta}{\sigma}\right)\right).$$

Tento výraz maximalizujú také hodnoty parametrov α, β, σ, pre ktoré platí

$$\frac{\partial \ell(\alpha,\beta,\sigma|y)}{\partial \alpha} = \sum_{i=1}^{n} \frac{1}{p_{y_i}} \cdot \frac{\partial p_{y_i}}{\partial \alpha} = 0$$

$$\frac{\partial \ell(\alpha,\beta,\sigma|y)}{\partial \beta} = \sum_{i=1}^{n} \frac{1}{p_{y_i}} \cdot \frac{\partial p_{y_i}}{\partial \beta} = 0$$

$$\frac{\partial \ell(\alpha,\beta,\sigma|y)}{\partial \sigma} = \sum_{i=1}^{n} \frac{1}{p_{y_i}} \cdot \frac{\partial p_{y_i}}{\partial \sigma} = 0.$$

Po dosadení za p_{y_i} a za derivácie dostávame systém 3 rovníc o 3 neznámych α, β, σ

$$\sum_{i=1}^{n} \left(\frac{\phi\left(\frac{y_i+\frac{h}{2}-\alpha-(i-1)\cdot\beta}{\sigma}\right) - \phi\left(\frac{y_i-\frac{h}{2}-\alpha-(i-1)\cdot\beta}{\sigma}\right)}{\Phi\left(\frac{y_i+\frac{h}{2}-\alpha-(i-1)\cdot\beta}{\sigma}\right) - \Phi\left(\frac{y_i-\frac{h}{2}-\alpha-(i-1)\cdot\beta}{\sigma}\right)} \right) = 0$$

$$\sum_{i=1}^{n} (i-1) \cdot \left(\frac{\phi\left(\frac{y_i+\frac{h}{2}-\alpha-(i-1)\cdot\beta}{\sigma}\right) - \phi\left(\frac{y_i-\frac{h}{2}-\alpha-(i-1)\cdot\beta}{\sigma}\right)}{\Phi\left(\frac{y_i+\frac{h}{2}-\alpha-(i-1)\cdot\beta}{\sigma}\right) - \Phi\left(\frac{y_i-\frac{h}{2}-\alpha-(i-1)\cdot\beta}{\sigma}\right)} \right) = 0$$

$$\sum_{i=1}^{n} \left(\frac{H^+ \cdot \phi\left(\frac{y_i+\frac{h}{2}-\alpha-(i-1)\cdot\beta}{\sigma}\right) - H^- \cdot \phi\left(\frac{y_i-\frac{h}{2}-\alpha-(i-1)\cdot\beta}{\sigma}\right)}{\Phi\left(\frac{y_i+\frac{h}{2}-\alpha-(i-1)\cdot\beta}{\sigma}\right) - \Phi\left(\frac{y_i-\frac{h}{2}-\alpha-(i-1)\cdot\beta}{\sigma}\right)} \right) = 0,$$

kde $H^+ = y_i + \frac{h}{2} - \alpha - (i-1)\cdot\beta$ a $H^- = y_i - \frac{h}{2} - \alpha - (i-1)\cdot\beta$. Na vypočítanie riešenia $(\hat{\alpha}, \hat{\beta}, \hat{\sigma})$ opäť využijeme MATLAB a jeho funkciu *fsolve*.

Ak by sme nezohľadňovali zaokrúhľovanie, na odhad parametrov α a β by sme mohli použiť napríklad metódu najmenších štvorcov, ktorá je určená nasledovne. Nech S je matica, ktorá má v prvom stĺpci samé jednotky a v druhom stĺpci hodnoty s_i. Potom je naša regresia určená ako $y = S \cdot \binom{\alpha}{\beta} + \varepsilon$ a odhad pomocou metódy najmenších štvorcov je určený ako

$$\begin{pmatrix} \tilde{\alpha} \\ \tilde{\beta} \end{pmatrix} = (S^T S)^{-1} S^T y$$

a $\tilde{s}^2 = \frac{RSS}{n-k}$, kde RSS je súčet štvorcov reziduí a k je počet parametrov v regresii, v našom prípade 2.

Simulácia 2.4 (Správanie sa odhadov v lineárnej regresii)

V tejto simulácii sa už snažíme odhadovať 3 *parametre, čiže MATLAB v každom kroku musí numericky počítať* 9 *derivácii (Jakobián je matica* 3×3*), výpočty sa stávajú čoraz zložitejšími. Preto sme túto simuláciu robili pre menej rôznych hodnôt parametrov ako predchádzajúce. Ako veľkosť náhodného výberu n sme stále požívali hodnotu* 100*, keďže sme sa už presvedčili, že na veľkosti odhadov parametrov veľkosť výberu veľmi nevplýva. Ako hodnotu* α *sme stále používali hodnotu* 0.25*. Ako hodnotu* β *sme používali hodnoty* $0.1, 0.4$ *a* 1*, aby sme simulovali rôzne situácie v regresii. Napríklad ak* $\beta = 1$*,* $\alpha + \beta \cdot s$ *bude stále mať hodnotu* $a + 0.25$*,* $a \in \mathbb{Z}$*, čiže bude stále na rovnakej pozícii vzhľadom na zaokrúhľovací interval a zaokrúhľovanie by malo mať najväčší účinok, pretože pôsobí primárne jedným smerom (čísla blízke* $a + 0.25$ *zaokrúhli vždy na a). Ak* $\beta = 0.4$ *alebo* 0.1*, účinky by nemali byť až také výrazné, pretože* $\alpha + \beta \cdot s$ *sa bude striedavo nachádzať na rôznych miestach v zaokrúhľovacom intervale, čím by sa účinky zaokrúhľovania mali vzájomne trochu rušiť, čiže očakávame, že zaokrúhľovanie nebude mať až také veľké účinky. Ako hodnotu* σ *sme používali hodnoty* $0.1, 0.2, 0.4, 0.6$ *a* 1*.*

Ďalej chceme zistiť, či aj v regresii bude odhad $\tilde{s}^2 = \frac{RSS}{n-2}$ *nadhodnocovať skutočnú disperziu o už spomínanú hodnotu* $\frac{1}{12}$.

Keďže ide už o príliš veľa údajov, výsledky prikladáme vo forme tabuľky. Odhadnuté stredné hodnoty odhadov získaných pomocou metódy najmenších štvorcov budeme označovať $\tilde{\alpha}, \tilde{\beta}$ *a* $\tilde{\sigma}^2$ *a odhadnuté stredné hodnoty odhadov pomocou ML budeme označovať* $\hat{\alpha}, \hat{\beta}$ *a* $\hat{\sigma}^2$. *Otázkou je s akou presnosťou máme uvádzať dosiahnuté výsledky. Táto otázka úzko súvisí s veľkosťou intervalu spoľahlivosti pre danú hodnotu, my budeme používať* 95%*-ný interval spoľahlivosti. Ak je napríklad veľkosť intervalu spoľahlivosti* 0.02*, vieme, že skutočná stredná hodnota odhadu môže byť s celkom rozumnou pravdepodobnosťou o* 0.01 *väčšia, ale aj o* 0.01 *menšia ako nami vypočítaná hodnota. Preto by v tomto prípade bol nezmysel vyčíslovať vypočítanú hodnotu na viac ako* 2 *desatinné miesta. Aby sa výsledkom dalo veriť, zaokrúhlime ich na toľko desatinných miest, koľko je núl v šírke intervalu spoľahlivosti. Vypočítať intervaly spoľahlivosti pre ML odhady by bolo veľmi zložité, ale pre obyčajné odhady pomocou najmenších štvorcov to nebude ťažké, pretože poznáme stredné hodnoty a disperzie odhadov jednotlivých parametrov ako aj ich rozdelenie. Potom aj ML odhady budeme uvádzať na rovnaký počet desatinných miest. Keďže vektor* $(\tilde{\alpha}, \tilde{\beta})$ *obsahuje odhady, vypočítané pomocou metódy najmenších štvorcov, vieme, že má dvojrozmerné normálne rozdelenie. Disperzia* $\tilde{\alpha}$ *bude určená prvkom s variančno-kovariančnej matice* $\sigma^2 \cdot (S^T S)^{-1}$ *so súradnicami* $[1,1]$*, disperzia* $\tilde{\beta}$ *bude určená prvkom so súradnicami* $[2,2]$. *Ostatné prvky matice sú kovariancie parametrov medzi sebou, tieto nás nebudú zaujímať. Prvok so súradnicami* $[1,1]$ *je určený ako* $\sigma^2 \cdot \left(\sum_{i=1}^{n} i\right)^{-1}$*, čiže platí* $D(\tilde{\alpha}) = \frac{\sigma^2}{n}$ *. Prvok so súradnicami* $[2,2]$ *je určený ako* $\sigma^2 \cdot \left(\sum_{i=1}^{n} (i-1)^2\right)^{-1}$. *Použijeme známy vzorec na súčet prvých k štvorcov prirodzených čísel a dostaneme výsledok* $D(\tilde{\beta}) = \frac{6 \cdot \sigma^2}{n \cdot (n-1) \cdot (2n-1)}$ *. V našej simulácii robíme* 10000 *opakovaní a potom hodnotu* $E(\tilde{\alpha})$ *aproximujeme pomocou* $\bar{\tilde{\alpha}}$*, podobne je to s*

betou. Priemer $\tilde{\alpha}$ *má normálne rozdelenie so strednou hodnotou* $E(\tilde{\alpha})$ *a disperziou* $\frac{D(\tilde{\alpha})}{p}$*, kde* p *je počet opakovaní simulácie, v našom prípade* 10000*. Z toho nám vyplýva, že*

$$\frac{(\bar{\tilde{\alpha}} - E(\tilde{\alpha})) \cdot \sqrt{p} \cdot \sqrt{n}}{\sigma} \sim \mathcal{N}(0,1).$$

Z tohto určíme šírku 95%*-ného intervalu spoľahlivosti ako*

$$2 \cdot u_{2.5\%} \cdot \frac{\sigma}{\sqrt{n} \cdot \sqrt{p}},$$

kde $u_{2.5\%}$ *je* 2.5%*-ná kritická hodnota normovaného normálneho rozdelenia. Dosadením čísel dostaneme pre šírku intervalu spoľahlivosti pre* $\bar{\tilde{\alpha}}$ *vyjadrenie približne* 0.004σ*. Budeme teda zaokrúhľovať na* 2 *resp.* 3 *desatinné miesta, podľa hodnoty* σ*.*

Pre priemer $\tilde{\beta}$ *platí*

$$\frac{\left(\bar{\tilde{\beta}} - E(\tilde{\beta})\right) \cdot \sqrt{p} \cdot \sqrt{n \cdot (n-1) \cdot (2n-1)}}{6 \cdot \sigma} \sim \mathcal{N}(0,1),$$

z čoho dostaneme šírku intervalu ako

$$2 \cdot u_{2.5\%} \cdot \frac{6 \cdot \sigma}{\sqrt{n \cdot (n-1) \cdot (2n-1)} \cdot \sqrt{p}},$$

čo je približne rovné 0.00007σ*. Keďže MATLAB automaticky zaokrúhlil výsledok na* 4 *desatinné miesta, ďalšie zaokrúhlenie výsledku už v tomto prípade nebude potrebné.*

Najzložitejšie je odhadnúť disperziu $\tilde{s}^2$*. Vychádzať budeme zo vzťahu* $\frac{(n-2)\cdot\tilde{s}^2}{\sigma^2} \sim \chi^2_{n-2}$*. Vieme, že pre* $X \sim \chi^2_k$ *a veľké* k *bude platiť približne* $\frac{X-k}{\sqrt{2k}} \sim \mathcal{N}(0,1)$*. Pri* $k > 50$ *je rozdiel už takmer zanedbateľný a my používame* $k = 98$*. Čiže platí*

$$\frac{(n-2) \cdot \tilde{s}^2}{\sigma^2 \cdot \sqrt{2 \cdot (n-2)}} - \frac{n-2}{\sqrt{2 \cdot (n-2)}} \sim \mathcal{N}(0,1)$$

Z toho vyplýva

$$\frac{\tilde{s}^2 - \sigma^2}{\frac{\sqrt{2}\cdot\sigma^2}{\sqrt{n-2}}} \sim \mathcal{N}(0,1) \Rightarrow \tilde{s}^2 \sim \mathcal{N}\left(\sigma^2, \frac{2\cdot\sigma^4}{n-2}\right) \Rightarrow D(\tilde{s}^2) = \frac{2\cdot\sigma^4}{n-2}.$$

Z toho už vieme, že šírka intervalu spoľahlivosti bude

$$2 \cdot u_{2.5\%} \cdot \sqrt{\frac{D(\tilde{s}^2)}{p}} = 2 \cdot u_{2.5\%} \cdot \sqrt{\frac{2\cdot\sigma^4}{(n-2)\cdot p}} \doteq 0.0056\sigma^2,$$

čiže budeme zaokrúhľovať na 2 *až* 4 *desatinné miesta, v závislosti od hodnoty* σ.

V celej simulácii sme si pevne zvolili $n = 100$, $\alpha = 0.25$

	σ	*0.1*	*0.2*	*0.4*	*0.6*	*1*
	σ^2	*0.01*	*0.04*	*0.16*	*0.36*	*1*
β						
	$\hat{\alpha} =$	*0.250*	*0.250*	*0.25*	*0.25*	*0.25*
	$\tilde{\alpha} =$	*0.240*	*0.246*	*0.25*	*0.25*	*0.25*
0.1	$\hat{\beta} =$	*0.1000*	*0.1000*	*0.1000*	*0.1000*	*0.1000*
	$\tilde{\beta} =$	*0.1002*	*0.1001*	*0.1000*	*0.1000*	*0.1000*
	$\hat{\sigma}^2 =$	*0.0087*	*0.037*	*0.155*	*0.35*	*0.98*
	$\tilde{\sigma}^2 =$	*0.0939*	*0.124*	*0.244*	*0.44*	*1.08*
	$\hat{\alpha} =$	*0.253*	*0.250*	*0.25*	*0.25*	*0.25*
	$\tilde{\alpha} =$	*0.245*	*0.246*	*0.25*	*0.25*	*0.25*
0.4	$\hat{\beta} =$	*0.4000*	*0.4000*	*0.4000*	*0.4000*	*0.3999*
	$\tilde{\beta} =$	*0.4001*	*0.4001*	*0.4000*	*0.4000*	*0.3999*
	$\hat{\sigma}^2 =$	*0.0082*	*0.037*	*0.154*	*0.35*	*0.98*
	$\tilde{\sigma}^2 =$	*0.0941*	*0.123*	*0.243*	*0.44*	*1.08*
	$\hat{\alpha} =$	*0.067*	*0.332*	*0.26*	*0.25*	*0.25*
	$\tilde{\alpha} =$	*0.006*	*0.105*	*0.24*	*0.25*	*0.25*
1	$\hat{\beta} =$	*1.0001*	*1.000*	*1.000*	*0.9999*	*1.000*
	$\tilde{\beta} =$	*1.000*	*1.000*	*1.000*	*0.9999*	*1.000*
	$\hat{\sigma}^2 =$	*0.0091*	*0.017*	*0.150*	*0.35*	*0.98*
	$\tilde{\sigma}^2 =$	*0.0062*	*0.095*	*0.243*	*0.44*	*1.08*

Prvá vec, ktorú si môžeme všimnúť je, že odhady parametra β *sú v podstate nevychýlené pre ľubovoľnú kombináciu ostatných hodnôt a to pre odhady zohľadňujúce ako aj odhady nezohľadňujúce zaokrúhlenie. Čiže môžeme tvrdiť, že zaokrúhľovanie neovplyvňuje odhad sklonu regresnej priamky. Potvrdili sa taktiež naše očakávania, že pre* $\beta = 1$ *bude mať zaokrúhľovanie ďaleko najväčší vplyv, ukázalo sa však, že iba na parametre* α *a* σ*. Výsledky pre tento*

prípad sú podobné tým, ktoré vyšli v simulácii 2.3, *akurát miesto* μ *v tejto simulácii vystupuje* α. *Dokonca aj ML odhad bol vychýlený podobne ako v tejto simulácii. Pre* $\beta = 0.1$ *resp.* 0.4 *boli výchylky odhadov pre* α *podstatne menšie a ML odhad skoro úplne presný. Odhady pre* σ^2 *nezohľadňujúce zaokrúhľovanie opäť nadhodnocovali skutočnú disperziu o približne* $\frac{1}{12}$, *ML odhady to dokázali celkom úspešne napraviť.*

Záver

V našej práci sme sa venovali vplyvu zaokrúhľovania v štatistike. Tému sme poňali skôr z praktického hľadiska, v prvej kapitole sme pomocou simulácii ukázali, že keď nezahrnieme zaokrúhľovanie do našich odhadov, výsledky budú vychýlené. V druhej kapitole sme prezentovali spôsob opísaný v [1], kde zahŕňame zaokrúhľovanie do ML odhadu parametrov. Následne sme vypočítali ML odhady pre parametre v exponenciálnom rozdelení, v normálnom rozdelení so známou aj neznámou disperziou a v lineárnej regresii. Pre exponenciálne rozdelenie sa nám podarilo nájsť explicitné vyjadrenie pre ML odhad parametra λ, pre zvyšné prípady sme výsledky odhadli pomocou numerických metód. Na simuláciach sme následne ukázali, že vo väčšine prípadov sa nám podarilo zlepšiť odhad oproti odhadu, kde nezohľadňujeme, že pozorovania boli zaokrúhlené. Takisto sme objavili príklad, kde zaokrúhlenie nemá žiaden účinok, totiž odhad sklonu regresnej priamky je nevychýlený pri zahrnutí aj nezahrnutí zaokrúhlenia do odhadu. Myslíme si, že k danej téme by sa dalo pristupovať aj teoretickejšie, napríklad uviesť dôkaz korekcie "$-\frac{1}{12}$"v odhadovaní disperzie normálneho rozdelenia, ktorý ale obtiažnosťou presahuje túto prácu. Vzhľadom na množstvo rôznych štatistických výskumov v súčasnosti si myslíme, že danej téme bude venovaná ešte veľká pozornosť.

Literatúra

[1] Bai, Z., Zheng, S., Zhang, B., Hu, G.: *Statistical analysis for rounded data*, Journal of Statistical Planning and Inference - J STATIST PLAN INFER , Volume 139, Number 8(2009), 2526-2542

[2] Casella, G., Berger, R. L.: *Statistical Inference: Second Edition*, Thomson Learning, Ann Arbor, 2002

[3] Heitjan, D.F.: *Inference from Grouped Continuous Data: A Review*, Statistical Science, Volume 4, Number 2 (1989), 164-179.

[4] Kubáček, Z., Valášek, J., *Cvičenia z matematickej analýzy II*, vysokoškolské skriptá, Univerzita Komenského, Bratislava, 1996

[5] Powell, M. J. D., *A Fortran Subroutine for Solving Systems of Nonlinear Algebraic Equations, Numerical Methods for Nonlinear Algebraic Equations*, P. Rabinowitz, ed., Ch.7, 1970.

[6] Sheppard, W. F., *On the calculation of the most probable values of frequency constants for data arranged according to equidistant divisions of a scale*, Proc. London Math. Soc.,29 (1898), 353–380.