# Západočeská univerzita v Plzni

## Fakulta aplikovaných věd

## Katedra matematiky


# Diplomová práce

# Dvouvýběrový Kolmogorovův-Smirnovův test a zaokrouhlená data


**Plzeň, 2017** **Bc. Zuzana Vlasáková**

# Prohlášení

Prohlašuji, že jsem diplomovou práci vypracovala samostatně a výhradně s použitím citovaných pramenů.

V Plzni dne ……

Zuzana Vlasáková

# Poděkování

V první řadě bych chtěla velmi poděkovat svému vedoucímu diplomové práce, Mgr. Michalu Frieslovi, Ph.D., za odborné vedení, trpělivost, ochotný přístup a přínosné rady během zpracování této práce. Poděkování patří i těm, kteří mě během mého dosavadního studia podporovali.

# Abstrakt

Diplomová práce se zabývá dvouvýběrovým Kolomogorovovým-Smirnovovým testem. Hlavním cílem je vyšetřit, jak zaokrouhlení vstupních dat ovlivní výsledky testu o shodě rozdělení. V první části práce jsou popsány výsledky simulací, v druhé části jsou uvedeny výsledky pro konkrétní data. Námětem pro vznik této práce byla bakalářská práce Martiny Kocandové Srovnání vlivu relativního věku ve sportu.

**Klíčová slova:** dvouvýběrový Kolmogorovův-Smirnovův test, simulace, zaokrouhlení dat, kritická hodnota, chyba 1. druhu, silofunkce

# Abstract

This thesis focuses on Two-sample Kolmogorov-Smirnov test. The main objective of the thesis is to find out how the rounding of the input data affects the results of the hypothesis of the same distribution. The first part of the thesis describes the results of the simulations. The second part gives the results for specific data. The reason of this thesis was the bachelor thesis of Martina Kocandová Comparison of the influence of the relative age in sport.

**Keywords:** Two-sample Kolmogorov-Smirnov test, simulation, data rounding, critical value, type 1 error, power of a test

# Obsah

# Seznam obrázků

# Seznam tabulek

# 1 Úvod

Cílem této diplomové práce bylo vyšetřit vliv zaokrouhlení vstupních dat na výsledek dvouvýběrového Kolmogorovova-Smirnovova testu o shodě rozdělení. Hlavním námětem pro vznik práce byla bakalářská práce [1]. Studentka zvolila pro testování shody výběrů data narození sportovců v hokeji, fotbale a šachu. V našem případě se omezíme pouze na šachisty. Všechna data však byla zaokrouhlena na celé měsíce, tím byl porušen předpoklad spojitosti výběrů. V práci jsme si proto položili otázku, zda zaokrouhlení vstupních dat ovlivní výsledky dvouvýběrového Kolmogorovova-Smirnovova testu. Nejprve je popsána situace pomocí simulací a poté byly poznatky ověřeny na konkrétních datech o šachistech. Text je členěn do sedmi kapitol.

Druhá kapitola je zaměřena na popsání dvouvýběrového Kolmogorovova-Smirnovova testu. Tento test byl použit při všech výpočtech. V kapitole je vysvětlen hlavní princip testu. Pro ukázku je uveden i důkaz Smirnovovy věty, která popisuje rozdělení testovací statistiky.

Postup jednotlivých simulací je uveden v třetí kapitole. Obsahuje volbu a způsob zaokrouhlování výběrů, které vstupují do testování. Zmíněny jsou sledované výstupní parametry, které se mohou lišit vlivem zaokrouhlení. Je ukázán postup odhadu kritických hodnot pro zvolený test, které se mohou měnit právě v závislosti na míře zaokrouhlení.

Další dvě kapitoly obsahují rozbor případů, pokud oba výběry pocházejí z rovnoměrného nebo normálního rozdělení. Simulace jsou provedeny pro různé rozsahy výběrů a pro změny parametrů jednotlivých rozdělení. Vždy jsou porovnávány dva přístupy. První je, zanedbává-li se zaokrouhlení vstupních dat. A v druhém případě je zahrnut vliv zaokrouhlení do testování.

V závěrečné části jsou shrnuty výsledky pro konkrétní data o šachistech. K dispozici byla data narození šachistů zaokrouhlena na měsíce. Všechna uvedená data byla čerpána z textu [1]. V práci jsou uvedeny výsledky testů pro tři vybrané kategorie šachistů. Je uváděn rozdíl mezi výsledky testování se zanedbáním zaokrouhlení vstupních pozorování, a pokud míru zaokrouhlení nezanedbáváme. Výsledky jsou uvedeny v jednotlivých podkapitolách.

Všechny výpočty a grafické výstupy byly provedeny v software Matlab a v MS Excel. Všechny zdrojové kódy a výpočty jsou dostupné na přiloženém CD.

# 2 Dvouvýběrový Kolmogorovův-Smirnovův test

Kolmogorovův-Smirnovův test pro dva výběry je použit při všech testech uvedených v diplomové práci. V [1] byl použit pro porovnání dat narození šachistů a české populace. Kapitola uvádí formulaci testu, jehož autory jsou Andrej Nikolajevič Kolmogorov a Vladimir Ivanovič Smirnov.

Kolmogorovův-Smirnovův test patří do třídy neparametrických metod porovnávajících shodu rozdělení dvou výběrů. Jako první zavedeme empirické distribuční funkce. [2]

Nechť $X_1, \dots, X_m$ je náhodný výběr z rozdělení s distribuční funkcí $F$. Pro $i = 1, \dots, m$

$$\xi_i(x) = \begin{cases} 1, pro\ X_i < x, \\ 0, pro\ X_i \geq x, \end{cases}$$

jsou náhodné veličiny. Náhodný proces $F_m(x) = \frac{1}{m}\sum_{i=1}^{m} \xi_i(x)$ se nazývá empirická distribuční funkce. Analogicky zavedeme empirickou distribuční funkci $G_n(x)$ pro náhodný výběr $Y_1, \dots, Y_n$ s distribuční funkcí $G$. Ukážeme, že pro takto zavedené empirické distribuční funkce platí následující tvrzení. [2]

**Věta 2.1** Pro každé $x$ platí skoro jistě

$$F_m(x) \xrightarrow{m\to\infty} F(x)$$

Obdobně $G_n(x) \to G(x)$ skoro jistě pro $n \to \infty$.

*Důkaz:*Víme, že pro pevně zvolená $x$ jsou $\xi_i(x)$ nezávislé stejně rozdělené veličiny a platí pro ně

$$P[\xi_i(x) = 1] = F(x), \;\; E\xi_i(x) = F(x).$$

Dokazování tvrzení spočívá na základě silného zákonu velkých čísel [3]. Označme $S_m = \sum_{i=1}^{m} \xi_i(x)$ jako součet náhodných veličin a $E(\xi_1) = \mu$ je konečná střední hodnota. Zákon velkých čísel nám říká, že s pravděpodobností jedna podíl $\frac{S_m}{m}$ konverguje pro $m \to \infty$ ke střední hodnotě $\mu$. Vidíme, že $F_m(x) = \frac{1}{m}\sum_{i=1}^{m} \xi_i(x) = \frac{S_m}{m}$ konverguje pro

$m \to \infty$ s pravděpodobností 1 ke své střední hodnotě, což je právě $E\xi_i(x) = F(x)$. Opět analogicky platí pro $G_n(x)$.

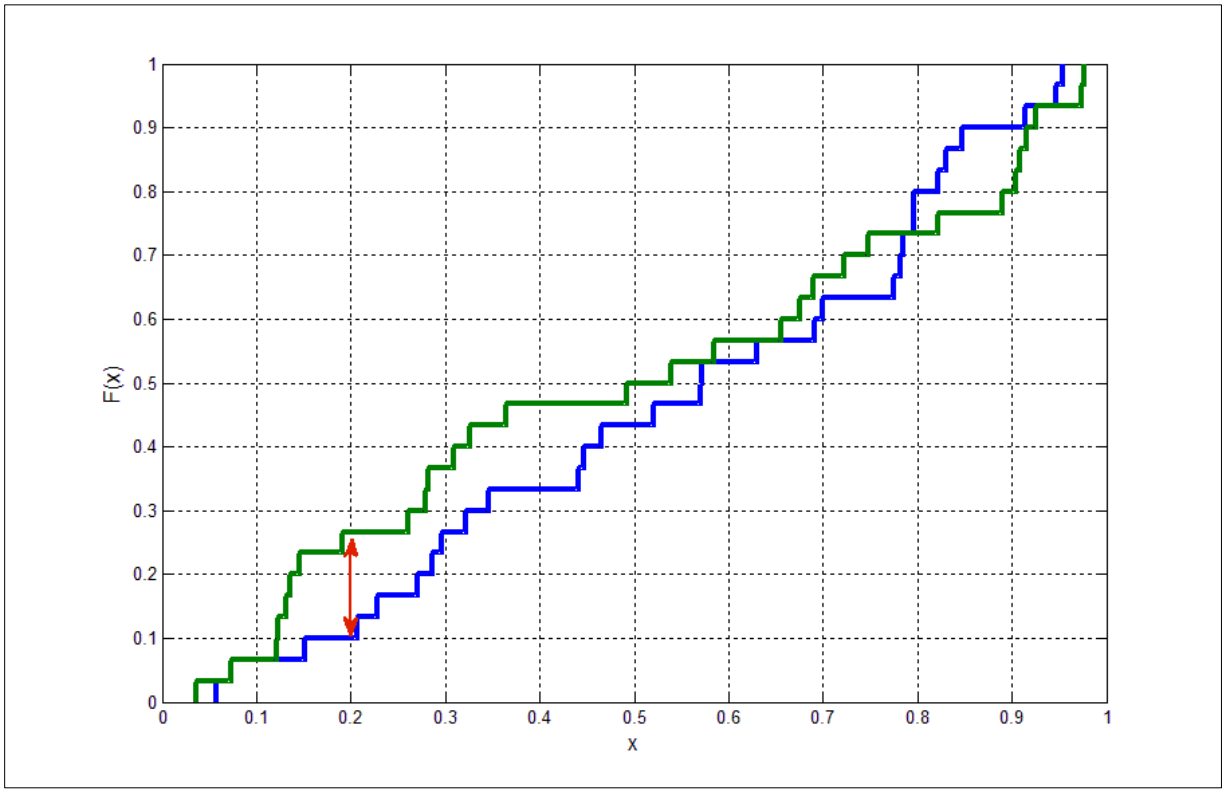
Ještě silnější tvrzení vyplývá z Glivenkovy věty, které navíc říká, že s pravděpodobností 1 empirická distribuční funkce $F_m(x)$ konverguje k distribuční funkci $F(x)$, roste-li počet prvků náhodného výběru ($m \to \infty$) stejnoměrně. Neboli z dostatečně velkého statistického souboru můžeme s pravděpodobností 1 získat libovolně podrobnou informaci o distribuční funkci $F(x)$.

**Věta 2.2 Glivenkova** Označíme si $D_m = sup_x|F_m(x) - F(x)|$. Potom platí

$$P\left(\lim_{m\to\infty} D_m = 0\right) = 1.$$

Důkaz Glivenkovy věty lze najít například v [4] na straně 340.

Pro porovnání dvou výběrů potřebujeme rozhodnout, zda pocházejí ze stejného rozdělení, tedy zda platí $F \equiv G$, rozdělení $F$ a $G$ může být libovolné. Při rozhodování o shodě rozdělení se pracuje s odchylkou $|F_m(x) - G_n(x)|$, konkrétně s maximální odchylkou $sup_x|F_m(x) - G_n(x)|$. Znázornění odchylky můžeme vidět červeně na Obrázku 2.1.




**Obrázek 2.1: Odchylka empirických distribučních funkcí**

Nechť $X_1, \dots, X_m$ je náhodný výběr pocházející ze spojitého rozdělení s distribuční funkcí $F, F_m$ je empirická distribuční funkce výběru a $Y_1, \dots, Y_n$ je náhodný výběr pocházející ze spojitého rozdělení s distribuční funkcí $G, G_n$ je empirická distribuční funkce výběru. Nechť oba výběry jsou navzájem nezávislé. Hypotézy o shodě rozdělení formulujeme ve tvaru (oboustrannou alternativu)

$$H_0: F = G$$
$$H_1: F \neq G.$$

Z Věty 2.1 už víme, že se empirické distribuční funkce $F_m$, $G_n$ pro $m, n \to \infty$ blíží k distribučním funkcím $F\ a\ G$.

Testovací statistika pro dvouvýběrový Kolmogorovův-Smirnovův test je ve tvaru

$$D_{m,n} = sup_x |F_m(x) - \ G_n(x)|,$$

varianta pro stejné rozsahy výběrů je pak ve tvaru

$$D_{n,n} = sup_x |F_n(x) - \ G_n(x)|.$$

Nulovou hypotézu $H_0$ na hladině významnosti $\alpha$ nezamítáme, pokud

$$D^*_{m,n}(\alpha) > D_{m,n},$$

naopak nulovou hypotézu o shodě rozdělení zamítáme, pokud platí

$$D^*_{m,n}(\alpha) \leq D_{m,n},$$

kde $D^*_{m,n}(\alpha)$ je kritická hodnota, určená jako $100 * (1 - \alpha)\%$ kvantil rozdělení veličiny $D_{m,n}$ [2].

Pokud jsou veličiny $X$ a $Y$ spojité, rozdělení veličiny $D_{m,n}$ je vždy stejné. Místo přesné hodnoty kvantilu se někdy používá aproximativní hodnota, která vychází z limitního rozdělení $D_{m,n}$. Aproximace má v tomto případě tvar

$$D^*_{m,n}(\alpha) \approx \sqrt{\frac{1}{2\frac{mn}{m+n}} ln \frac{2}{\alpha}},$$

varianta pro stejné rozsahy

$$D^*_{n,n}(\alpha) \approx \sqrt{\frac{1}{n} ln \frac{2}{\alpha}}.$$

Aproximativní kritická hodnota je odvozena z limitní Věty 2.8, která bude uvedena později. Kritická hodnota nezávisí na rozdělení veličin $X$ a $Y$.

Nyní uvedeme jednostranné alternativy o shodě rozdělení.

$$H_0: F = G$$
$$H_1: F > G.$$

Alternativa popisuje, že výběr $X_1, \dots, X_m$ pochází z rozdělení, jehož distribuční funkce nabývá ve všech bodech větších hodnot než druhá distribuční funkce pro náhodný výběr $Y_1, \dots, Y_n$. Testovací statika je ve tvaru

$$D^+{}_{m,n} = sup_x(F_m(x) - \ G_n(x)).$$

Nulovou hypotézu $H_0$ na hladině významnosti $\alpha$ nezamítáme, pokud

$$D^{+*}{}_{m,n}(\alpha) > D^+{}_{m,n},$$

naopak nulovou hypotézu o shodě rozdělení zamítáme, pokud platí

$$D^{+*}{}_{m,n}(\alpha) \leq D^+{}_{m,n},$$

kde $D^{+*}{}_{m,n}(\alpha)$ je kritická hodnota.

V případě, že alternativou budou záporné hodnoty rozdílu mezi $F_m(x)$ a $G_n(x)$, formulujeme hypotézy ve tvaru

$$H_0: F = G$$
$$H_1: F < G.$$

Alternativa popisuje, že výběr $X_1, \dots, X_m$ pochází z rozdělení, jehož distribuční funkce nabývá ve všech bodech menších hodnot než druhá distribuční funkce pro náhodný výběr $Y_1, \dots, Y_n$. Testovací statistika je ve tvaru

$$D^-{}_{m,n} = sup_x(G_n(x) - F_m(x)).$$

Nulovou hypotézu $H_0$ na hladině významnosti $\alpha$ nezamítáme, pokud

$$D^{-*}{}_{m,n}(\alpha) > D^{-}{}_{m,n},$$

naopak nulovou hypotézu o shodě rozdělení zamítáme, pokud platí

$$D^{-*}{}_{m,n}(\alpha) \leq D^{-}{}_{m,n},$$

kde $D^{-*}{}_{m,n}(\alpha)$ je kritická hodnota.

K rozhodnutí, kdy odchylka $|F_m(x) - G_n(x)|$, resp. $\left(F_m(x) - G_n(x)\right)$ a $(G_n(x) - F_m(x))$, dvou rozdělení už je významná, lze použít tzv. Smirnovovy věty, které hovoří o přesném rozdělení veličiny $sup_x|F_m(x) - G_n(x)|$. V diplomové práci při simulačních pokusech bylo počítáno se shodnými rozsahy výběrů. Docházelo tedy k porovnání empirických distribučních funkcí $F_n(x)$ a $G_n(x)$. Dokazovat v práci proto budeme Smirnovovu větu formulovanou právě pro shodné rozsahy výběrů. Různé rozsahy byly použity v práci při srovnávání výsledků z [1]. Věta pro různé rozsahy bude uvedena na konci kapitoly.

**Věta 2.4** (Pro jednostranný test) Pokud $F(x) \equiv G(x)$, potom

$$\lim_{n\to\infty} P\left(\sqrt{\frac{n}{2}} sup_{-\infty<x<\infty}\left(F_n(x) - G_n(x)\right) < y\right) = \begin{cases} 1 - e^{-2y^2} & pro\ y > 0, \\ 0 & jinak. \end{cases}$$

**Věta 2.5** (Pro oboustranný test) Pokud $F(x) \equiv G(x)$, potom

$$\lim_{n\to\infty} P\left(\sqrt{\frac{n}{2}} sup_{-\infty<x<\infty}|F_n(x) - G_n(x)| < y\right) = \begin{cases} K(y) & pro\ y > 0, \\ 0 & jinak, \end{cases}$$

kde $K(y) = \sum_{k=-\infty}^{\infty}(-1)^k e^{-2k^2y^2}$.

Pro dokázání Smirnovových vět jdou využít Věty 2.6 a Věty 2.7 Koroljuka a Gněděnka. V následujícím textu bude uveden důkaz Věty 2.4 limitním přechodem Věty 2.6 ([4] na straně 426 nebo [5] na straně 171) a bude uvedena myšlenka důkazu Věty 2.5.

**Věta 2.6** Pokud $\{x\}$ je nejmenší celé číslo, které není menší než $x$, pokud $c = \{z\sqrt{2n}\}$ a zároveň $F(x) \equiv G(x)$, potom

$$P\left(\sqrt{\frac{n}{2}} sup_{-\infty<x<\infty}\left(F_n(x) - \ G_n(x)\right) < z\right) = \begin{cases} 0 & pro\ z \le 0, \\ 1 - \frac{\binom{2n}{n-c}}{\binom{2n}{n}}\ pro\ 0 < z \le \sqrt{\frac{n}{2}}, \\ 1 & jinak. \end{cases}$$

**Věta 2.7** Pokud $\{x\}$ je nejmenší celé číslo, které není menší než $x$, pokud $c = \{z\sqrt{2n}\}$ a zároveň $F(x) \equiv G(x)$, potom

$$P\left(\sqrt{\frac{n}{2}} sup_{-\infty<x<\infty}|F_n(x) - \ G_n(x)| < z\right)$$

$$= \begin{cases} 0 & pro\ z \le \frac{1}{\sqrt{2n}}, \\ \frac{1}{\binom{2n}{n}} \sum_{k=-\left(\frac{n}{c}\right)}^{\left(\frac{n}{c}\right)} (-1)^k \binom{2n}{n-kc}\ \ pro \frac{1}{\sqrt{2n}} < z \le \sqrt{\frac{n}{2}}, \\ 1 & jinak. \end{cases}$$

*Důkaz Věty 2.6:* Jádro důkazu spočívá v řešení kombinatorické úlohy. Nejprve si vytvoříme seřazenou posloupnost $XY$ o rozsahu $2n$. Posloupnost vznikne seřazením veličin $X_1, \dots, X_n$ a $Y_1, \dots, Y_n$ podle velikosti. Nyní prvky posloupnosti $XY$ nahradíme číslem 1, pokud prvek pochází z výběru $X_1, \dots, X_n$ a číslem $-1$, pokud pochází z $Y_1, \dots, Y_n$. Takto vzniklou posloupnost označíme $U$ o rozsahu $2n$, k-tý prvek posloupnosti $U$ označíme $U_k$.

Př. $X = \{1, 4, 12\}, Y = \{2, 5, 10\}, XY = \{1, 2, 4, 5, 10, 12\}, U = \{1, -1, 1, -1, -1, 1\}, U_3 = 1$

Nyní si zavedeme součet prvních k-členů posloupnosti $U$ jako $S_k = \ U_1 + U_2 + \cdots + U_k$. Před samotným dokazováním uvedeme ještě pomocné tvrzení

$$sup_{-\infty<x<\infty}\left(F_n(x) - \ G_n(x)\right) = \frac{1}{n}\max_{1\le k\le 2n}(S_k),$$

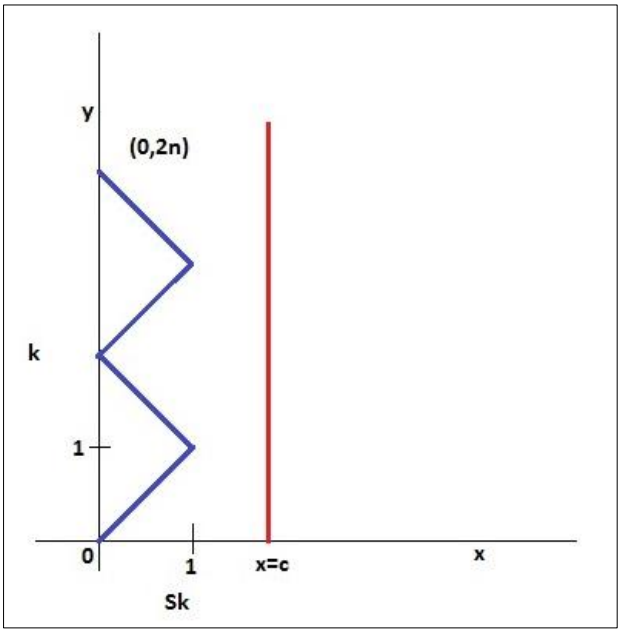
$$sup_{-\infty<x<\infty}|F_n(x) - \ G_n(x)| = \frac{1}{n}\max_{1\le k\le 2n}|S_k|,$$

kde číslo $n\big(F_n(x) - \ G_n(x)\big)$ je rozdíl mezi počtem prvků posloupnosti $X_1, \dots, X_n$ menších než $x$ a počtem prvků $Y_1, \dots, Y_n$ menších než $x$. Hodnota výrazu se mění jen tehdy, pokud $x$ přesáhne hodnotu $XY_k$ (k-tý prvek posloupnosti $XY$). Potom pomocné tvrzení dokážeme následujícím

$$sup_{-\infty<x<\infty} n\big(F_n(x) - \ G_n(x)\big) = \max_{1\le k\le 2n} n\big(F_n(XY_k\ + 0) - \ G_n(XY_k\ + 0)\big) = \max_{1\le k\le 2n} S_k,$$

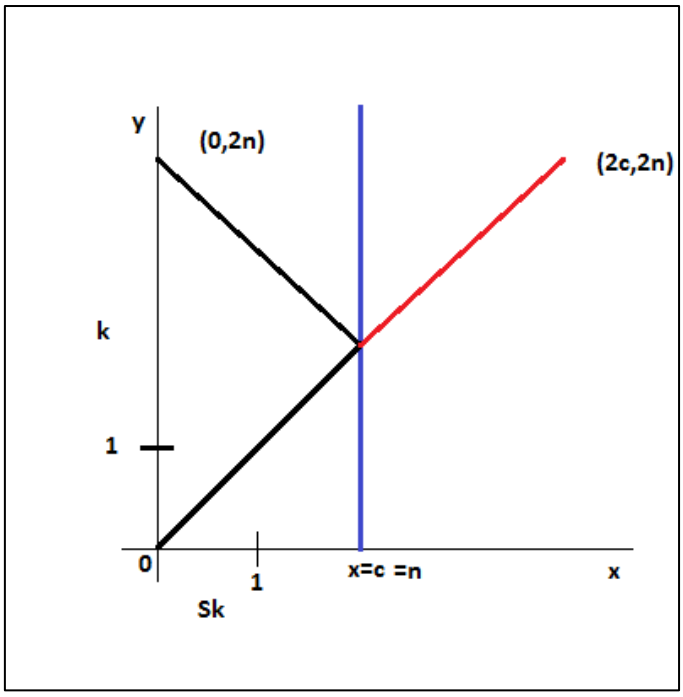
$$sup_{-\infty<x<\infty} n|F_n(x) - \ G_n(x)| = \max_{1\le k\le 2n} n|F_n(XY_k\ + 0) - \ G_n(XY_k\ + 0)| = \max_{1\le k\le 2n} |S_k|.$$

Nyní máme připravenou posloupnost $U$ z čísel 1 a $-1$ a zavedený vztah pro výraz $\max_{1\le k\le 2n} S_k$. Přistoupíme ke kombinatorickému postupu. Počet všech možností, jak takovou posloupnost $U$ lze získat, je vybráním z $2n$ prvků ($n$ jedniček a $n$ mínus jedniček) $n$ prvků, tj. $\binom{2n}{n}$. Je zřejmé, že každá taková vzniklá posloupnost je stejně pravděpodobná, pokud jsou výběry $X_1, \dots, X_n$
a $Y_1, \dots, Y_n$ navzájem nezávislé a stejně rozdělené. Pravděpodobnost každé takové posloupnosti je $\frac{1}{\binom{2n}{n}}$. Z Věty 2.6 potřebujeme nalézt pravděpodobnost $P(\max_{1\le k\le 2n} S_k < z\sqrt{2n})$. Pro zjištění této pravděpodobnosti musíme vědět, kolik posloupností $U$ tuto podmínku splňuje. Řešení lze zjistit pomocí grafického znázornění. Na osu $y$ vyneseme $k$ (počet sčítanců $S_k$) a na osu $x$ vyneseme hodnoty $S_k$ (součet prvních k-členů posloupnosti $U$). Vyneseme tedy body $[S_k, k]$ a spojíme je čárou. Tím každé možné posloupnosti
$U$ přiřadíme lomenou čáru v rovině vycházející z bodu $[0, 0]$, jejíž úseky svírají s osou $x$ $\pm 45°$. Úhel $45°$ vychází z nejjednodušší volby měřítka $1:1$ a v každém bodě $[S_k, k]$ máme právě dvě možnosti volby směru. Tyto lomené čáry nazveme cesty. Každá taková cesta bude vycházet z bodu $[0, 0]$ a bude končit v bodě $[0, 2n]$, jelikož součet všech $U$ musí být vždy 0 (nasčítáváme stejný počet 1 a $-1$). Celkový počet $k$ je rozsah posloupnosti $U$. Na Obrázku 2.2 splňují nerovnost $\max_{1\le k\le 2n} S_k < z\sqrt{2n}$ body nalevo přímky $x = z\sqrt{2n} = c$. Nerovnost je splněna, jestliže cesta nemá s přímkou $x = z\sqrt{2n} = c$ žádný společný bod, viz Obrázek 2.2. Počet takových cest (posloupností) zjistíme doplňkem k celkovému počtu cest, o kterém už víme, že je $\binom{2n}{n}$.

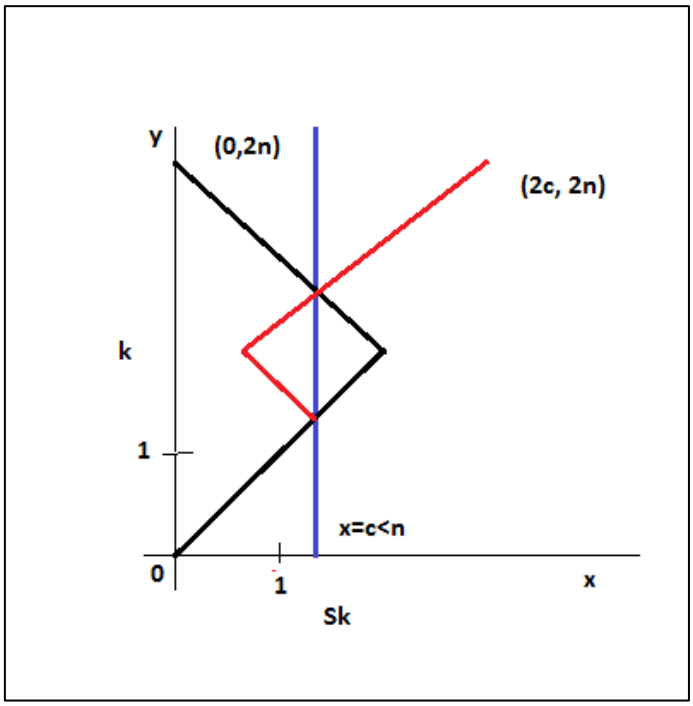



**Obrázek 2.2: Cesta neprotínající přímku $x = z\sqrt{2n} = c$**

Pro zjištění počtu cest, které nemají s přímkou $x = z\sqrt{2n} = c$ žádný společný bod, určíme nejprve počet cest, které naopak alespoň jeden společný bod mají. Na Obrázku 2.3 vidíme zobrazenou cestu, která má s přímkou $x = z\sqrt{2n} = c$ právě jeden společný bod, cesta je vyznačená černě. Tento případ je hraničním případem pro $c = n$ (jediná možnost vytvoření). Je to případ, kdy jde po sobě $n$-krát číslo 1 a $n$-krát číslo $-1$, aby byla splněna podmínka, že skončíme v $[0, 2n]$. Vytvořením zrcadlového obrazu části cesty podle přímky $c$ od jejich prvního společného bodu získáme novou cestu začínající v $[0, 0]$ a končící v bodě $[2c, 2n]$, pro hraniční případ to je přímka. Zrcadlením uměle změníme směr cesty (vyznačena červeně), nové úseky svírají s osou $x$ opět $\pm 45°$. Takto se vytvoří všechny cesty (příklad na Obrázku 2.4), které mají s přímkou $c$ alespoň jeden společný bod. Pokud si zvolíme např. $n - 1 = c$, tak cesta může změnit směr o jednou víckrát než pro $n = c$, počet cest bude $\binom{2n}{n-(n-1)}$, s dalším posunutím získáme více možností, tj. $\binom{2n}{n-c}$.




**Obrázek 2.3: Cesta protínající přímku $x = z\sqrt{2n} = c = n$, hraniční případ**




**Obrázek 2.4: Cesta protínající přímku $x = z\sqrt{2n} = c$**

Tudíž je získán doplněk k počtu cest, které žádný společný bod s $c$ nemají. Nyní již tento počet můžeme vyčíslit jako

$$\binom{2n}{n} - \binom{2n}{n-c}.$$

Pro nalezení pravděpodobnosti jevu $S_k < z\sqrt{2n}$ bylo zjištěno, kolik posloupností $U_1, \dots, U_{2n}$ tuto podmínku splňuje a vydělíme-li ho nyní počtem všech možných posloupností $U_1, \dots, U_{2n}$, dostaneme

$$P\left(\max_{1\le k\le 2n} S_k < z\sqrt{2n}\right) = \frac{\binom{2n}{n}-\binom{2n}{n-c}}{\binom{2n}{n}} = 1 - \frac{\binom{2n}{n-c}}{\binom{2n}{n}}.$$

Nyní je možno provést důkaz Věty 2.4 limitním přechodem Věty 2.6 pro $z\sqrt{2n} = c$, bude využito Stirlingova vzorce

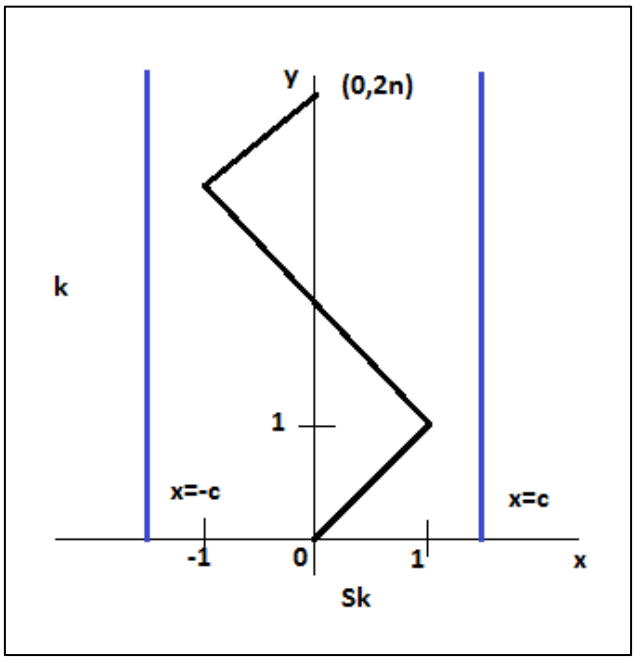
$$\lim_{n\to\infty} \frac{n!}{\sqrt{2\pi n}\left(\frac{n}{e}\right)^n} = 1.$$

*Důkaz Věty 2.4:*

$$\begin{aligned}
\lim_{n\to\infty} \frac{\binom{2n}{n+c}}{\binom{2n}{n}} &= \lim_{n\to\infty} \frac{(2n)!}{(n+c)!\,(n-c)!}\,\frac{n!\;n!}{(2n)!} \\
&= \lim_{n\to\infty} \frac{n!\,n!}{(n+c)!\,(n-c)!} \\
&= \lim_{n\to\infty} \frac{\sqrt{2\pi n}\left(\frac{n}{e}\right)^n \sqrt{2\pi n}\left(\frac{n}{e}\right)^n}{\sqrt{2\pi(n+c)}\left(\frac{n+c}{e}\right)^{n+c} \sqrt{2\pi(n-c)}\left(\frac{n-c}{e}\right)^{n-c}} \\
&= \lim_{n\to\infty} n^{1+2n}(n+c)^{-\frac{1}{2}-n-c}(n-c)^{-\frac{1}{2}-n+c} \\
&= \lim_{n\to\infty} \frac{n}{\sqrt{(n+c)(n-c)}}\left(\frac{n^2}{(n+c)(n-c)}\right)^n \left(\frac{n-c}{n+c}\right)^c \\
&= \lim_{n\to\infty} \frac{n}{\sqrt{\left(n+z\sqrt{2n}\right)\left(n-z\sqrt{2n}\right)}}\left(\frac{n^2}{\left(n+z\sqrt{2n}\right)\left(n-z\sqrt{2n}\right)}\right)^n \left(\frac{n-z\sqrt{2n}}{n+z\sqrt{2n}}\right)^{z\sqrt{2n}} \\
&= \lim_{n\to\infty} \frac{n}{\sqrt{(n+z\sqrt{2n})(n-z\sqrt{2n})}}\left(\left(1+\frac{2z^2}{n-2z^2}\right)^{2z^2}\right)^{\frac{n}{2z^2}} \frac{\left(\left(1-\frac{\sqrt{2}z}{\sqrt{n}}\right)^{\sqrt{n}}\right)^{\sqrt{2}z}}{\left(\left(1+\frac{\sqrt{2}z}{\sqrt{n}}\right)^{\sqrt{n}}\right)^{\sqrt{2}z}} \\
&= 1 \cdot e^{2z^2} \cdot \frac{e^{-2z^2}}{e^{2z^2}} = e^{-2z^2}
\end{aligned}$$

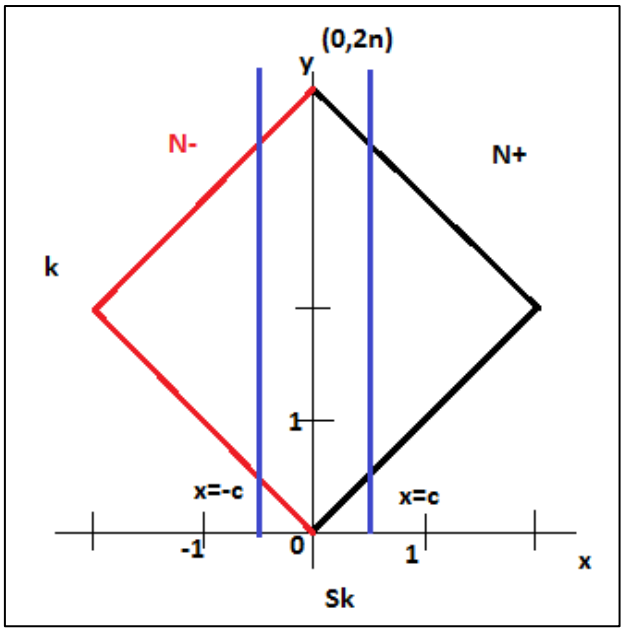
Pro ukázku důkazu Věty 2.7 bude uvedena alespoň grafická interpretace.

Myšlenka je obdobná jako u předešlého důkazu. Potřebujeme zjistit $P(\max_{1\leq k\leq 2n}|S_k| < z\sqrt{2n})$. Kvůli absolutní hodnotě nyní budou zkoumány cesty, které nemají žádný společný bod ani s jednou z přímek $x = \pm z\sqrt{2n} = \pm c$. Příklad takové cesty je uveden na Obrázku 2.5.



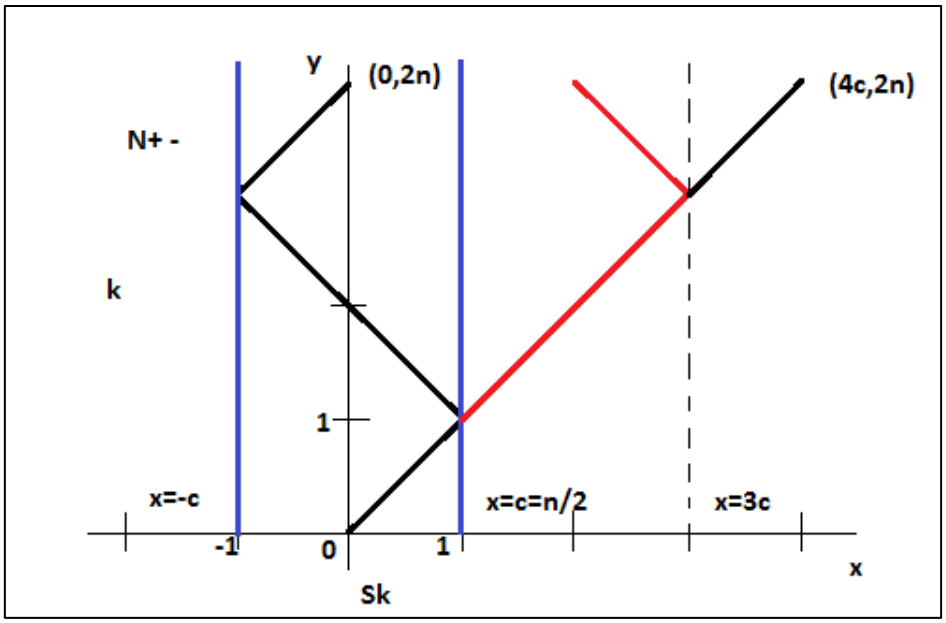

**Obrázek 2.5: Cesta neprotínající přímku $x = \pm z\sqrt{2n} = \pm c$**

Opět počet cest (označíme $N_0$) neprotínajících přímky $x = \pm z\sqrt{2n} = \pm c$ budeme hledat přes doplněk k celkovému počtu cest, který je $\binom{2n}{n}$. Celkem mohou nastat tři varianty, cesta s přímkou nemá žádný společný bod (Obrázek 2.5), cesta má společný bod jen s jednou z přímek (Obrázek 2.6), nebo cesta má společný bod s oběma přímkami (Obrázek 2.7).




**Obrázek 2.6: Cesta protínající právě jednu z přímek $x = \pm z\sqrt{2n} = \pm c$**

Na Obrázku 2.6 je vidět variantu, kdy cesta má jeden společný bod pouze s jednou z přímek (buď $x = c$ vyznačena černě, nebo $x = -c$ vyznačena červeně), počet takových cest označíme $N_+, N_-$. Příklad cesty, která má nejdříve společný bod s $x = c$ a poté s $x = -c$, je uveden na Obrázku 2.7, takovou cestu budeme značit $N_{+-}$. Nyní je hraničním případem varianta pro dotyk obou přímek $c = n/2$, neboť počet 1 a $-1$ musí být opět $2n$ a cesta je z úseků vždy po $\frac{1}{4}$ z $2n$, tj. $n/2$. Zrcadlení se provede nejprve podle $x = c$, za prvním společným bodem s touto přímkou (červeně) a poté takto vzniklou cestu zrcadlíme podle přímky $x = 3c$ od jejího prvního společného bodu s cestou (modře). Tím nám vznikla cesta začínající v $[0, 0]$ a končící v bodě $[4c, 2n]$. Celý důkaz lze najít např. v [4].




**Obrázek 2.7: Cesta protínající nejprve přímku $x = z\sqrt{2n} = c$ a potom přímku $x = -z\sqrt{2n} = -c$, hraniční případ**

Na závěr kapitoly je uvedena ještě Smirnovova věta pro různé rozsahy výběrů. Její důkaz lze najít např. v [6].

**Věta 2.8** Pokud $F(x) \equiv G(x)$, potom

$$\lim_{m,n\to\infty} P\left(\sqrt{\frac{nm}{m+n}}\, sup_{-\infty<x<\infty}|F_n(x) - \ G_m(x)| < y\right) = K(y),$$

kde $K(y) = 1 - 2\sum_{k=1}^{\infty}(-1)^{k+1}e^{-2k^2y^2}$.

Nelimitní případ věty lze nalézt např. v [5] na straně 175 nebo v [7] na straně 1452.

# 3 Simulace a zaokrouhlování dat

Záměrem práce bylo vyšetřit vliv zaokrouhlení dat na výsledky dvouvýběrového Kolmogorovova-Smirnovova testu. V [1] Kocandová testuje, zda se rozdělení dat narození šachistů shoduje s rozdělením dat české populace. Při testování používá zaokrouhlení vstupních dat na měsíc narození, tím je porušen předpoklad spojitosti rozdělení.

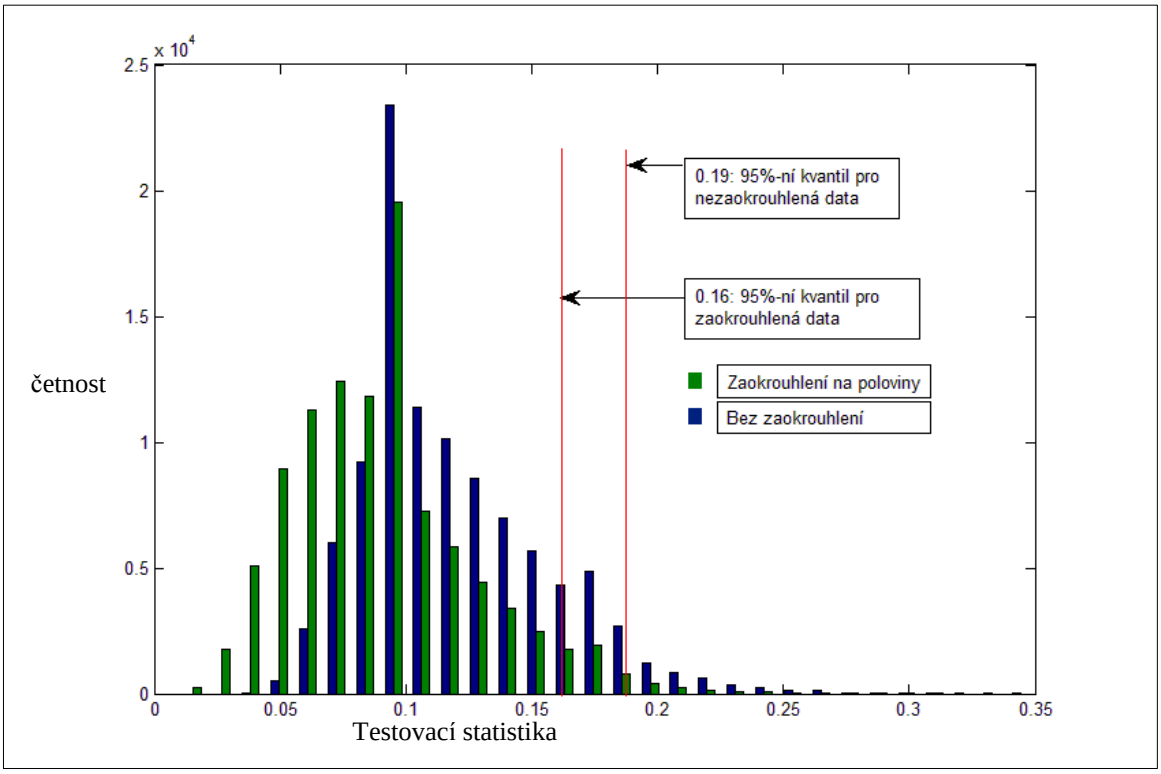
Původní model předpokládá, že $X_1, \dots, X_m$ je náhodný výběr pocházející ze spojitého rozdělení a $Y_1, \dots, Y_n$ je náhodný výběr ze spojitého rozdělení, my máme ale k dispozici zaokrouhlené hodnoty $X_1^d, \dots, X_m^d$ a $Y_1^d, \dots, Y_n^d$. Pokud bude vyhodnocen test s kritickou hodnotou $D_{m,n}^*(\alpha)$ pro spojitý případ, získá se však jiná hodnota pravděpodobnosti chyby 1. druhu než $\alpha$. Budeme se tedy zabývat vlivem zaokrouhlení na výsledek dvouvýběrového KS testu. Budou sledovány změny chyby 1. druhu, silofunkce, kritické hodnoty.

## 3.1 Kritická hodnota pro zaokrouhlená data

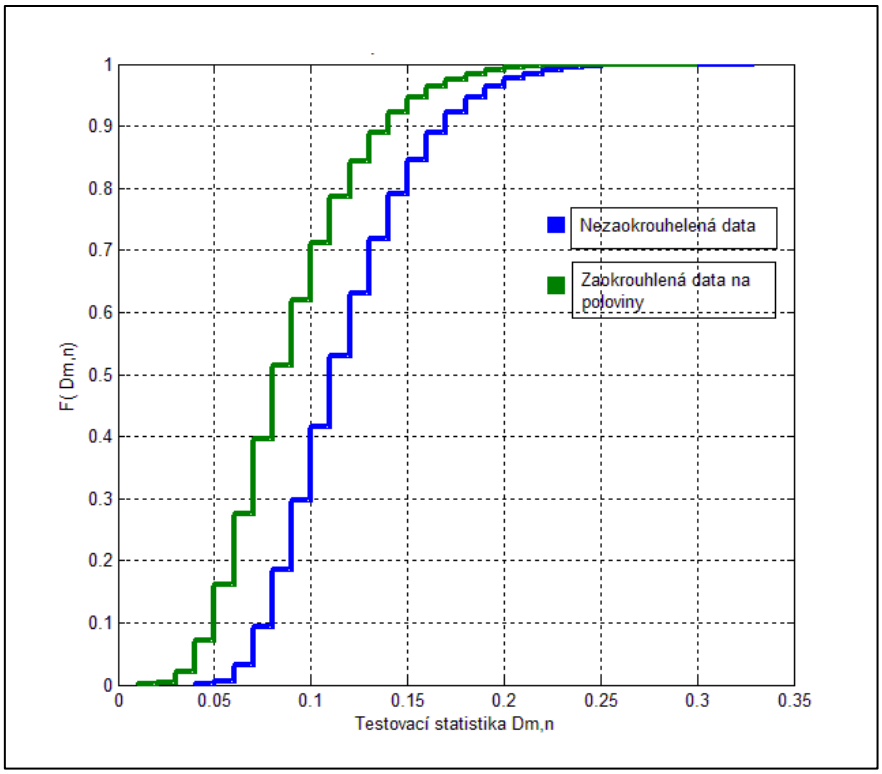
Kritickou hodnotu „správnou“, která zahrnuje skutečnost, že náhodný výběr obsahuje zaokrouhlená pozorování, označíme $D_{m,n}^!(\alpha)$. Tato „správná“ kritická hodnota se mění pro různé míry zaokrouhlení, rozsahy výběrů i pro zvolená rozdělení. Přibližná hodnota $D_{m,n}^!(\alpha)$ byla získána simulačně. Pro dvouvýběrový Kolomogorovův-Smirnovův test pro zaokrouhlená pozorování byla získána přibližná $D_{m,n}^!(\alpha)$ pomocí 100010 simulací. Při každé simulaci byla vygenerována testovací statistika, pomocí funkce v software Matlab $[D_{m,n}] = kstest2(x, y)$. Pro každou míru zaokrouhlení, rozdělení a velikost rozsahu byla provedena nová sada simulací. Tím se pro každou kombinaci rozsahu, rozdělení a míry zaokrouhlení získalo 100010 testovacích statistik $D_{m,n}$. Jelikož námi zvolená hladina významnosti testu byla $\alpha = 5\ \%$, hledali jsme pro určení kritické hodnoty $D_{m,n}^!(\alpha)$ 95% kvantil, který jsme odhadli 95% výběrovým kvantilem z 100010 nasimulovaných testovacích statistik. Počet simulací 100010 byl volen, aby hodnota $D_{m,n}^!(\alpha)$ byla určena jednoznačně. Pokud by simulací bylo přesně 100000, mohl by nastat případ, kdy by 95% výběrovému kvantilu odpovídal celý interval. Nelze tím však podchytit případ, že pro více simulací vyjde stejná hodnota $D_{m,n}$. Pokud se pro nezaokrouhlené výběry použije odhadnutá kritická hodnota $D_{m,n}^!(\alpha)$, měly by se získat srovnatelné výsledky jako při použití $D_{m,n}^*(\alpha)$.

Příklad získání jedné $D_{m,n}^!(\alpha)$ je uveden na Obrázku 3.1. Vyznačeny jsou histogramy četností testovacích statistik ze simulací Kolomogorvova-Smirnovova testu pro dva výběry

z normálního normovaného rozdělení o rozsahu 100, oba výběry jsou zaokrouhleny na poloviny (zeleně). Pro srovnání je vyznačen histogram testovacích statistik i pro nezaokrouhlená data (modře). V grafu je vyznačen 95% kvantil (červeně) pro nezaokrouhlená data ($D^{*}_{m,n}(\alpha)$) a pro zaokrouhlená data ($D^{!}_{m,n}(\alpha)$). Histogramy se přibližně liší posunutím. Zelený histogram pro zaokrouhlené pozorování je mírně přikloněn doleva. Hodnoty odhadnutých kritických hodnot se liší o 0,3. Pro porovnání uvádíme ještě empirické distribuční funkce na Obrázku 3.2, kde se také může pozorovat posun zhruba o 0,3.




**Obrázek 3.1: Simulace kritické hodnoty pro KS test mezi výběry z N (0, 1) o rozsahu 100 (bez zaokrouhlení a se zaokrouhlením na poloviny)**




**Obrázek 3.2: Empirické distribuční funkce pro nasimulované testovací statistiky**

Stejným postupem byla získána $D^{!}_{m,n}(\alpha)$ pro vyhodnocení dvouvýběrového Kolomogorvova-Smirnovova testu s různými rozsahy výběrů, rozdělením a mírou zaokrouhlení. Nyní se mohou porovnat výsledky, když je zanedbáváno zaokrouhlení, a když je použita správná kritická hodnota.

Jedním z kritérií pro porovnání výsledků byla zvolena pravděpodobnost chyby 1. druhu. Hodnoty se získaly následujícím postupem

1. Provede se 1000 simulací, a tím se dostane 1000 krát výběr $x$ a $y$ o určeném rozsahu $n$ z daného rozdělení.

2. Hodnoty výběrů se zaokrouhlí podle zvolené míry.

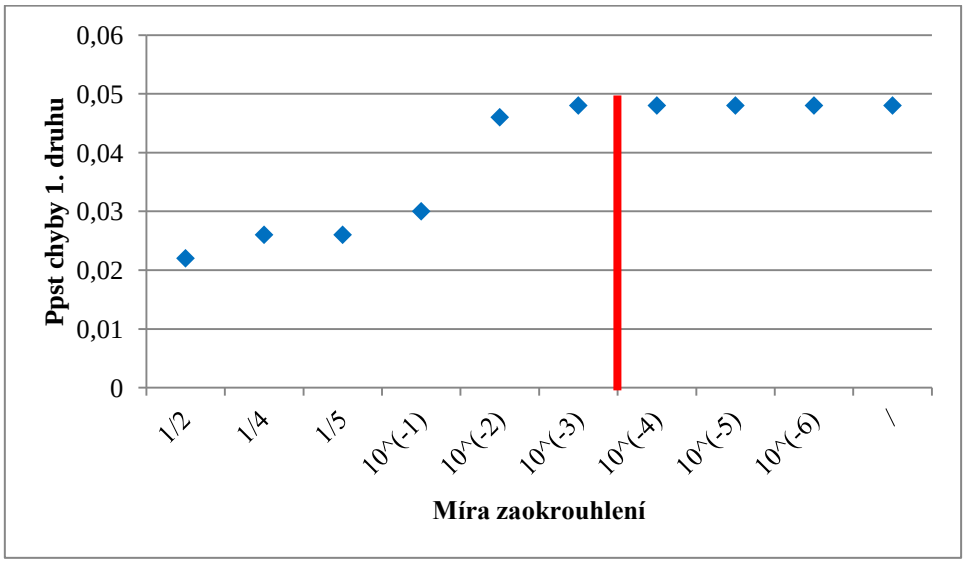
3. Pro každou míru zaokrouhlení zvlášť se vyhodnotí KS test, nejprve s hodnotou $D^{*}_{m,n}(\alpha)$ a pak s hodnotu $D^{!}_{m,n}(\alpha)$.

4. Sečte se počet případů, kdy byla $H_0$ zamítnuta, a počet se vydělí počtem simulací. Tím se získá odhad pravděpodobnosti chyby 1. druhu.

Další kritéria porovnání výsledků jsou například síla testu a změna kritické hodnoty.

## 3.2 Volba míry zaokrouhlení

Variant míry zaokrouhlení je mnoho. Bude ověřen předpoklad, že čím větší zaokrouhlení, tím větší vliv na výsledek budeme pozorovat. Je ovšem zbytečné volit příliš malé zaokrouhlení. Jako kritérium pro zvolenou míru byl vytvořen graf, který je uvedený v Obrázku 3.3. Testování proběhlo mezi náhodnými výběry s distribuční funkcí normálního normovaného rozdělení o rozsazích 100. Symbol „/“ odpovídá výběrům bez zaokrouhlení.




**Obrázek 3.3: Volba míry zaokrouhlení**

V grafu je vynesena pravděpodobnost chyby 1. druhu v KS testu s kritickou hodnotou $D^*_{m,n}(\alpha)$ v závislosti na míře zaokrouhlení. Jako hranice tedy byla zvolena míra zaokrouhlení $10^{-3}$, při zaokrouhlení na více desetinných míst předpokládáme obdobné výsledky jako při nezaokrouhlení. Způsob zaokrouhlení v software Matlab je uveden Tabulce 3.1.

| Míra zaokrouhlení | Zaokrouhlená hodnota | Funkce v Matlabu |
|---|---|---|
| **1/2** | 0 | round(x*2)/2 |
| **1/4** | 0,25 | round(x*4)/4; |
| **1/5** | 0,2 | round(x*5)/5; |
| **$10^{-1}$** | 0,2 | roundn(x, -1); |
| **$10^{-2}$** | 0,17 | roundn(x, -2); |
| **$10^{-3}$** | 0,166 | roundn(x, -3); |
| **/** | 0,165648 | / |

**Tabulka 3.1: Použití zaokrouhlení v software Matlab pro hodnotu $x = 0,165648$**

Pro zaokrouhlování byly využity funkce $round(x)$ a $roundn(x, a)$. První funkce zaokrouhlí číslo $x$ na celá čísla, druhá funkce číslo $x$ zaokrouhlí na nejbližší násobek $10^a$.

# 4 Rovnoměrné rozdělení

Pro simulační testy bylo zvoleno rovnoměrné a normální rozdělení. V této kapitole se budeme zabývat rovnoměrným rozdělením na intervalu $(a, b)$, dále jen $R(a, b)$. V simulacích jsme se omezili pouze na výběr z $R(0,1)$. Pro simulaci v software Matlab byla použita funkce $x = rand(n, 1)$, kde $n$ je zvolený rozsah. K libovolnému intervalu $(a, b)$ se lze dostat lineární transformací

$$Y = a + (b - a) * X,$$

kde $X \sim R(0, 1)$ a $Y \sim R(a, b)$.

Proto se budou jako reprezentativní případ simulovat data pouze z $R(0, 1)$. Výsledky dvouvýběrového KS testu pro $R(0, 1)$ a $R(a, b)$ budou shodné pro nezaokrouhlená pozorování. Dostane se shodná kritická hodnota i testovací statistika. Posouvají se jen hodnoty $x$ a $y$, ale hodnoty empirických distribuční funkcí $F_m$ a $G_n$ se nemění, protože platí

$$F_m(a + (b - a)x) = F_m(x).$$

Pokud se použijí zaokrouhlené výběry, transformací a následným zaokrouhlením už nemusí být odpovídající hodnoty původnímu výběru z $R(0, 1)$. (Např. Pro jedno pozorování z výběru z $R(0, 1)$ se zaokrouhlilo nahoru, ale transformované pozorování z výběru z $R(a, b)$ se zaokrouhlilo dolů.) Výsledky by zhruba měly být stejné, pokud se zvolí odpovídající zaokrouhlení a velikost intervalu $(a, b)$. (Např. Zaokrouhlí-li se výběry z $R(0, 1)$ na dvě desetinná místa a transformované výběry z $R(0,10)$ na jedno desetinné místo, získají se shodné výsledky. Analogicky dostaneme shodné výsledky pro výběry z $R(0, 1)$ zaokrouhlené na poloviny a transformované výběry z $R(0; 0,5)$ na čtvrtiny.)

Vždy bylo provedeno 1000 simulací, tj. 1000 krát se negenerovaly dva výběry z $R(0,1)$ a vyhodnotila se shoda rozdělení pomocí funkce $kstest2(x, y)$ v software Matlab. Pro každou sadu simulací byly výběry zaokrouhlovány různou mírou, získané výsledky ze zaokrouhlování jsou tedy vždy provedeny na stejnou sadu dat. V kapitole bude ukázán zároveň s vlivem zaokrouhlení také vliv velikosti rozsahu výběru a vliv transformování $R(0,1)$.

## 4.1 Změna rozsahu

Při simulacích byl zvolen rozsah výběrů od 10 do 1000. V Tabulce 4.1 jsou pro ukázku uvedeny nasimulované kritické hodnoty $D^{!}_{m,n}(\alpha)$. Z tabulky je vidět, že čím větší zaokrouhlení, tím je změna $D^{!}_{m,n}(\alpha)$ větší. Z toho lze vyvozovat, že pokud se vstupní data zaokrouhlí příliš, může nastat situace chybného vyhodnocení KS testu při použití $D^{*}_{m,n}(\alpha)$.

| Rozsah | Míra zaokrouhlení | | | | | | |
|---|---|---|---|---|---|---|---|
| | / | $10^{-3}$ | $10^{-2}$ | $10^{-1}$ | 1/5 | 1/4 | 1/2 |
| 10 | 0,6000 | 0,6000 | 0,6000 | 0,5000 | 0,5000 | 0,5000 | 0,4000 |
| 100 | 0,1900 | 0,1800 | 0,1700 | 0,1600 | 0,1500 | 0,1400 | 0,1900 |
| 200 | 0,1350 | 0,1350 | 0,1300 | 0,1150 | 0,1100 | 0,1050 | 0,0950 |
| 300 | 0,1100 | 0,1100 | 0,1067 | 0,0967 | 0,0900 | 0,0867 | 0,0800 |
| 500 | 0,0840 | 0,0840 | 0,0820 | 0,0740 | 0,0700 | 0,0680 | 0,0600 |
| 1000 | 0,0600 | 0,0600 | 0,0580 | 0,0530 | 0,0490 | 0,0480 | 0,0430 |

**Tabulka 4.1: Nasimulované kritické hodnoty pro výběry rovnoměrného rozdělení**

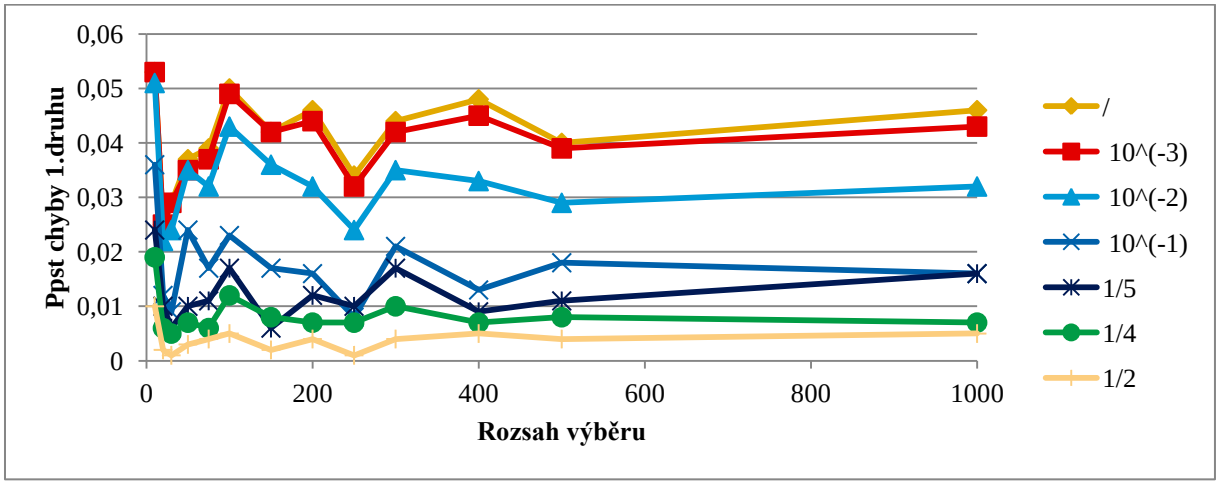
Pro rozsah výběru 10 se nasimulovaná kritická hodnota $D^{!}_{m,n}(\alpha)$ lišila od $D^{*}_{m,n}(\alpha)$ při zaokrouhlení $10^{-1}$. Ale například pro rozsah 200 se kritické hodnoty lišily již pro zaokrouhlení $10^{-2}$. V Tabulce 4.2 a v Tabulce 4.3 jsou uvedeny pravděpodobnosti chyby 1. druhu.

| Rozsah | Míra zaokrouhlení | | | | | | |
|---|---|---|---|---|---|---|---|
| | / | $10^{-3}$ | $10^{-2}$ | $10^{-1}$ | 1/5 | 1/4 | 1/2 |
| 10 | 0,053 | 0,053 | 0,051 | 0,036 | 0,024 | 0,019 | 0,01 |
| 100 | 0,05 | 0,049 | 0,043 | 0,023 | 0,017 | 0,012 | 0,005 |
| 200 | 0,046 | 0,044 | 0,032 | 0,016 | 0,012 | 0,007 | 0,004 |
| 300 | 0,044 | 0,042 | 0,035 | 0,021 | 0,017 | 0,01 | 0,004 |
| 500 | 0,04 | 0,039 | 0,029 | 0,018 | 0,011 | 0,008 | 0,004 |
| 1000 | 0,046 | 0,043 | 0,032 | 0,016 | 0,016 | 0,007 | 0,005 |

**Tabulka 4.2: Pravděpodobnosti chyb 1. druhu KS testu s $D^{*}_{m,n}(\alpha)$**

V Tabulce 4.2 vidíme sestupnou tendenci výsledků v závislosti na míře zaokrouhlení, tj. čím větší zaokrouhlení, tím menší pravděpodobnost chyby 1. druhu. Provedená sada testů zanedbávala zaokrouhlení vstupních dat. Pokud se výběry zaokrouhlily na poloviny, je

pravděpodobnost chyby 1. druhu přibližně 10 krát menší než pro nezaokrouhlené výběry.. Získané výsledky jsou graficky znázorněny na Obrázku 4.1. Každá lomená čára odpovídá jinému stupni zaokrouhlení. Při menších rozsazích výběrů se vliv zaokrouhlení jevil nepatrně menší.




**Obrázek 4.1: Graf s výsledky KS testu s $D^{*}_{m,n}(\alpha)$, $R(0,1)$**

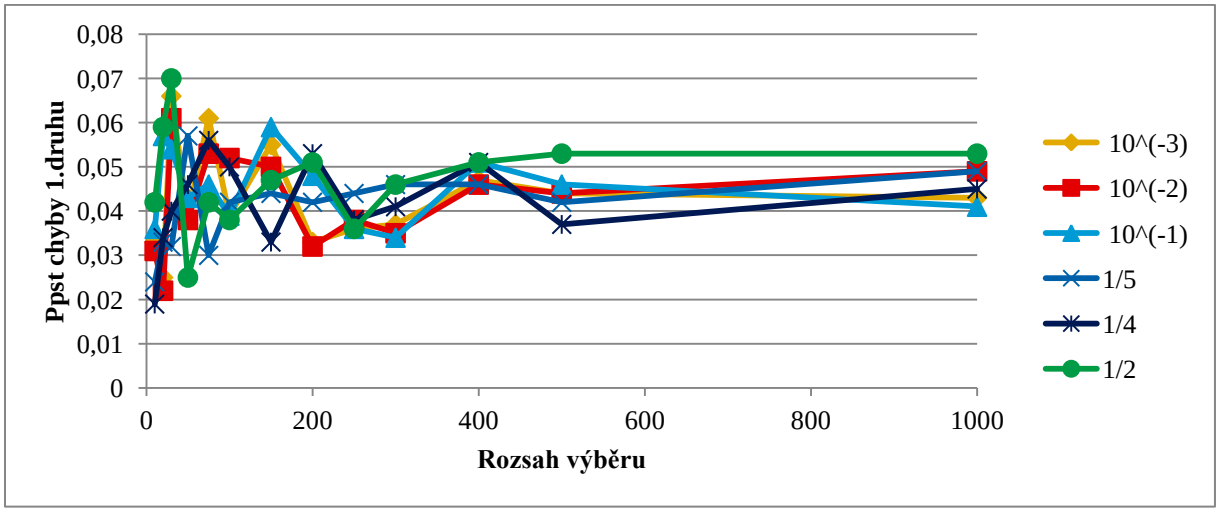
Nyní bude porovnán případ, kdy se vyhodnocení testu provede pomocí nasimulované $D^{!}_{m,n}(\alpha)$. Každé míře zaokrouhlení odpovídá vyhodnocení KS testu s příslušnou $D^{!}_{m,n}(\alpha)$, výsledky jsou uvedeny v Tabulce 4.3 a grafické znázornění na Obrázku 4.2.

| Rozsah | Míra zaokrouhlení | | | | | |
|---|---|---|---|---|---|---|
| | $10^{-3}$ | $10^{-2}$ | $10^{-1}$ | 1/5 | 1/4 | 1/2 |
| 10 | 0,033 | 0,031 | 0,036 | 0,024 | 0,019 | 0,042 |
| 100 | 0,039 | 0,052 | 0,039 | 0,042 | 0,050 | 0,038 |
| 200 | 0,033 | 0,032 | 0,048 | 0,042 | 0,053 | 0,051 |
| 300 | 0,037 | 0,035 | 0,034 | 0,046 | 0,041 | 0,046 |
| 500 | 0,044 | 0,044 | 0,046 | 0,042 | 0,037 | 0,053 |
| 1000 | 0,043 | 0,049 | 0,041 | 0,049 | 0,045 | 0,053 |

**Tabulka 4.3: Pravděpodobnosti chyb 1. druhu KS testu s $D^{!}_{m,n}(\alpha)$**

Z výsledků je patrné, že použitím $D^{!}_{m,n}(\alpha)$ jsme získaly výrazně odlišné výsledky než s použitím $D^{*}_{m,n}(\alpha)$. Při nižších rozsazích výsledky oscilují z důvodu, že při simulacích $D_{m,n}(\alpha)$ máme k dispozici málo pozorování, ze kterých se testovací statistika počítá. Při

rozsahu přibližně od 200 se ale výsledky ustalují přibližně na hladině pravděpodobnosti chyby 1. druhu 5 %. Pro výsledky s vyhodnocením s $D^{!}_{m,n}(\alpha)$ byly spočítány také intervaly spolehlivosti. Hodnoty v Tabulce 4.3 jsou přibližně srovnatelné, proto bude uveden příklad pro jednu hodnotu. Pro rozsah 300 a zaokrouhlení na čtvrtiny vyšel 95% interval spolehlivosti $(0{,}0287; 0{,}0533)$. Je zjevné, že při zvětšení počtu simulací se bude interval spolehlivosti zužovat.




**Obrázek 4.2: Graf s výsledky KS testu s $D^{!}_{m,n}(\alpha)$, $R(0,1)$**

## 4.2 Změna sklonu

Následující kapitola je zaměřena nejen na vliv zaokrouhlení vstupních dat, ale i na možnost, že druhý výběr nepochází přímo z $R(0,1)$, budeme tedy zjišťovat sílu testu při vybrané alternativě. Místo rovnoměrného rozdělení, bude mít druhý výběr lineární hustotu na intervalu $(0,1)$. Způsob získání jednoho takového rozdělení je popsán níže.

Nejprve uvedeme funkci hustoty pro rovnoměrné rozdělení, $X \sim R(0,1)$

$$f(x) = \begin{cases} 0 & pro\ x < 0, \\ 1 & pro\ 0 \leq x \leq 1, \\ 0 & pro\ x > 1. \end{cases}$$

Hledáme transformované rozdělení $R \sim R^{*}(0{,}1, a)$, aby platilo

$$f(r) = \begin{cases} 0 & pro\ r < 0, \\ ar + b & pro\ 0 \leq r \leq 1, (a, b\ kosnstanty), \\ 0 & pro\ r > 1. \end{cases}$$

$$F(r) = \begin{cases} 0 & pro\ r < 0, \\ \int_0^r (at + b)dt = \frac{ar^2}{2} + br & pro\ 0 \leq r \leq 1, \\ 1 & pro\ r > 1. \end{cases}$$

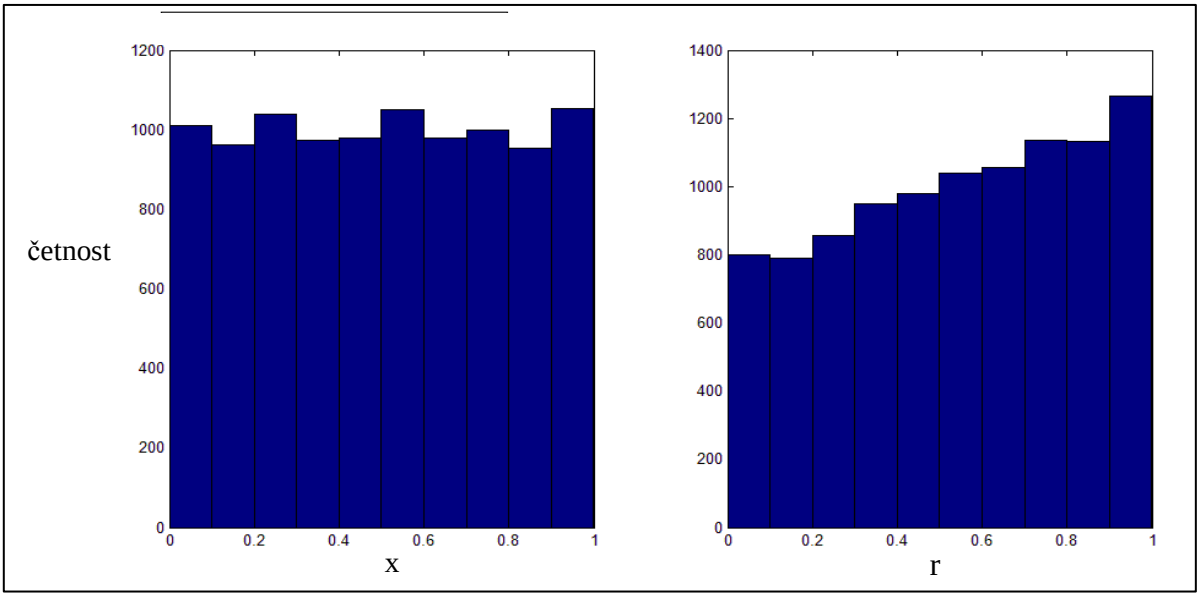
Pro simulaci hodnot, které se budou řídit distribuční funkcí z lineárního rozdělení $R^*(0, 1, a)$ je potřeba provést inverzní transformaci.

$$\frac{ar^2}{2} + br = x$$

$$(\sqrt{a}r + \frac{b}{\sqrt{a}})^2 = 2x + \frac{b^2}{a}$$

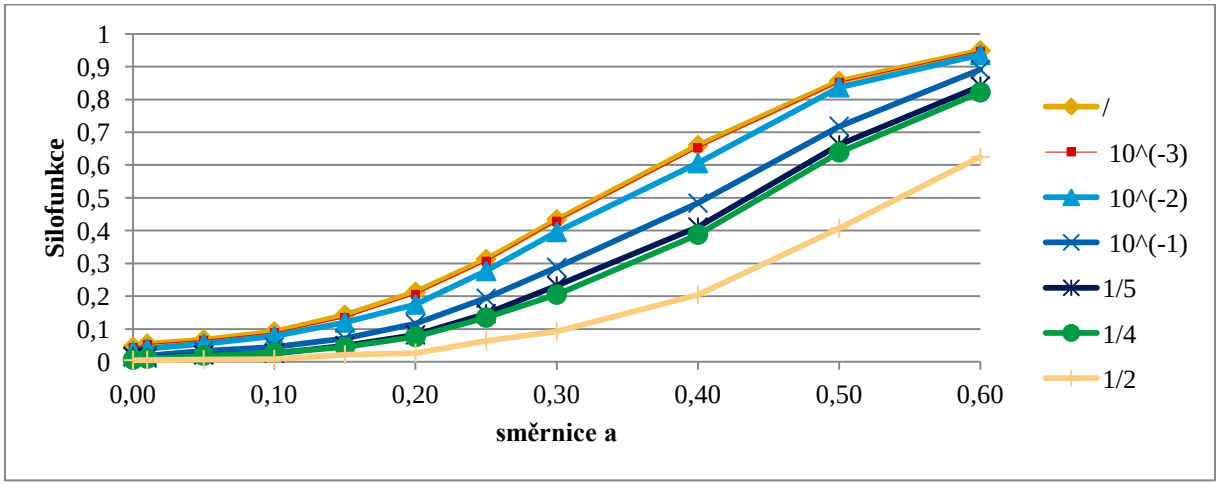
$$r = \frac{\sqrt{2x + \frac{b^2}{a}} - \frac{b}{\sqrt{a}}}{\sqrt{a}}$$

Kde $X \sim R(0, 1)$ a konstanty $a$, $b$ jsou konstanty lineární funkce. Konstanta $b$ závisí na volbě parametru $a$. Závislost lze vyjádřit jako $b = 1 - \frac{1}{2}a$. V dalším simulování se bude měnit právě směrnice („sklon“) $a$ (tím se mění i parametr $b$). Příklad simulace náhodného výběru z rozdělní s distribuční funkcí pro $R^*(0, 1, a)$ je uveden na Obrázku 4.3, pro srovnání je uveden i histogram náhodného výběru z rozdělní s distribuční funkcí pro $R(0, 1)$.



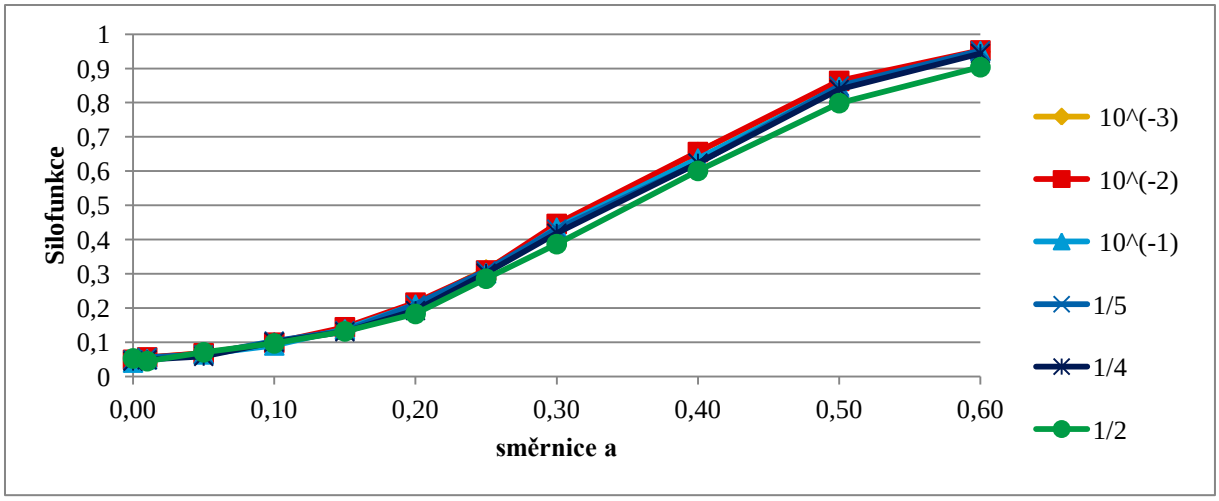

**Obrázek 4.3: Histogramy četností pro rozsah výběru 10000, vlevo simulace náhodného výběru z rozdělní s distribuční funkcí pro $R(0, 1)$, vpravo náhodný výběr z rozdělení s distribuční funkcí pro $R^*(0;\ 1; 0, 5)$**

Nyní si položíme otázku, zda zaokrouhlení obou výběrů zastře sklon jednoho z výběrů, který je upraven transformací na $R^*(0,1,a)$. Výsledky pokusu se zanedbáním zaokrouhlení jsou uvedeny na Obrázku 4.4, bylo provedeno 1000 simulací s rozsahy výběrů 1000.




**Obrázek 4.4: Výsledky KS testu s $D^*_{m,n}(\alpha)$ pro náhodný výběr z rozdělní s distribuční funkcí pro $R(0,1)$ a náhodný výběr z rozdělní s distribuční funkcí pro $R^*(0,1,a)$ se změnou sklonu**

Z grafu lze vyčíst, že pravděpodobnost zamítnutí $H_0$ pro zaokrouhlení na poloviny je nižší než pravděpodobnost zamítnutí u nezaokrouhlených výběrů. Například pro směrnici $a = 0{,}6$ je výsledek síly testu s nezaokrouhlenými daty o 34 % vyšší než se zaokrouhlením na poloviny. Test při zanedbání zaokrouhlení na poloviny má ve skutečnosti chybu 1. druhu nižší než $\alpha$. Z toho lze vyvozovat, že pokud se nebere v úvahu zaokrouhlení vstupních dat, tak vlivem zaokrouhlení se mohou zkreslit výsledky KS testu, a tím i částečně zastřít rozdíly v testovaných výběrech. Varianta s vyhodnocením KS testu s $D^!_{m,n}(\alpha)$ je uvedena v Obrázku 4.5.




**Obrázek 4.5: Výsledky KS testu s $D^{!}_{m,n}(\alpha)$ pro náhodný výběr z rozdělní s distribuční funkcí pro $R(0,1)$ a náhodný výběr z rozdělní s distribuční funkcí pro $R^{*}(0,1,\alpha)$**

Po použití odhadu kritické hodnoty $D^{!}_{m,n}(\alpha)$ jsou rozdíly výsledků s různými stupni zaokrouhlení minimální. Pokud se tedy bere v úvahu zaokrouhlení vstupních dat, tak vyhodnocení KS testu mezi výběry z $R(0,1)$ a z $R^{*}(0,1,\alpha)$ je téměř totožné jako pro nezaokrouhlený výběr.

# 5 Normální rozdělení

Další sada simulací byla provedena na náhodných výběrech z normálního rozdělení se střední hodnotou $\mu$ a rozptylem $\sigma^2$, dále $N(\mu, \sigma^2)$. V software Matlab byla použita funkce $random('normal', \mu, \sigma^2, n)$, kde $\mu$ je střední hodnota, $\sigma^2$ rozptyl a $n$ je rozsah výběru. Jako reprezentativní příklad rozdělení $N(\mu, \sigma^2)$ bylo zvoleno normální normované, tj. $N(0, 1)$. Lze totiž ukázat (viz níže), že k libovolnému $N(\mu, \sigma^2)$ se lze dostat transformací $N(0, 1)$.

Má-li náhodná veličina $X$ rozdělení $N(0, 1)$, pak pro náhodnou veličinu

$$Y = \mu + \sigma X$$

platí, že $Y \sim N(\mu, \sigma^2)$.

Zde nastává obdobný případ jako pro rovnoměrné rozdělení. Výsledky KS testu budou shodné i po transformaci $N(0, 1)$ na $N(\mu, \sigma^2)$ pro nezaokrouhlená pozorování. Vlivem zaokrouhlení transformovaného výběru může dojít opět k odlišným výsledkům, ale u zaokrouhlených výběrů by výsledky měly být zhruba stejné.

V této kapitole se budeme zabývat změnou rozsahu a odchylkou od $N(0, 1)$ pomocí parametrů $\mu$ a $\sigma^2$. Všechny variace simulací byly provedeny s různými mírami zaokrouhlení.

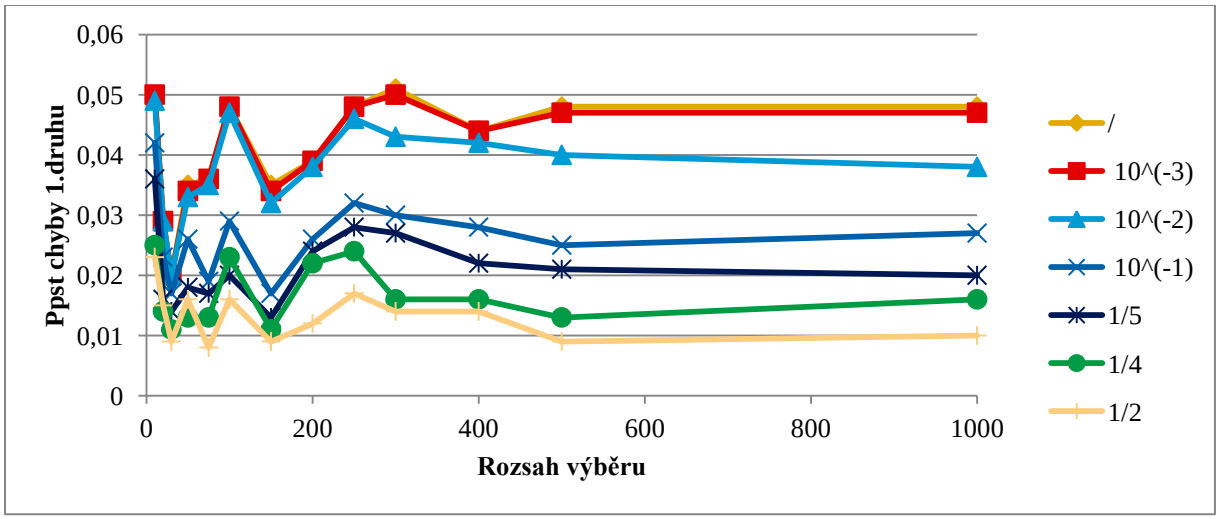
## 5.1 Změna rozsahu

Nejprve jsme opět nasimulovali odhad kritické hodnoty $D^{!}_{m,n}(\alpha)$ pro rozsahy výběrů od 10 do 1000. Příklad simulací odhadů je uveden v Tabulce 5.1.

| Rozsah | Míra zaokrouhlení | | | | | | |
|---|---|---|---|---|---|---|---|
| | / | $10^{-3}$ | $10^{-2}$ | $10^{-1}$ | 1/5 | 1/4 | 1/2 |
| 10 | 0,6000 | 0,6000 | 0,6000 | 0,5000 | 0,5000 | 0,5000 | 0,5000 |
| 100 | 0,1900 | 0,1900 | 0,1900 | 0,1800 | 0,1700 | 0,1700 | 0,1600 |
| 200 | 0,1350 | 0,1350 | 0,1300 | 0,1250 | 0,1200 | 0,1200 | 0,1100 |
| 300 | 0,1100 | 0,1100 | 0,1067 | 0,1033 | 0,0967 | 0,0967 | 0,0900 |
| 500 | 0,0840 | 0,0840 | 0,0840 | 0,0780 | 0,0760 | 0,0740 | 0,0700 |
| 1000 | 0,0600 | 0,0600 | 0,0590 | 0,0560 | 0,0540 | 0,0530 | 0,0500 |

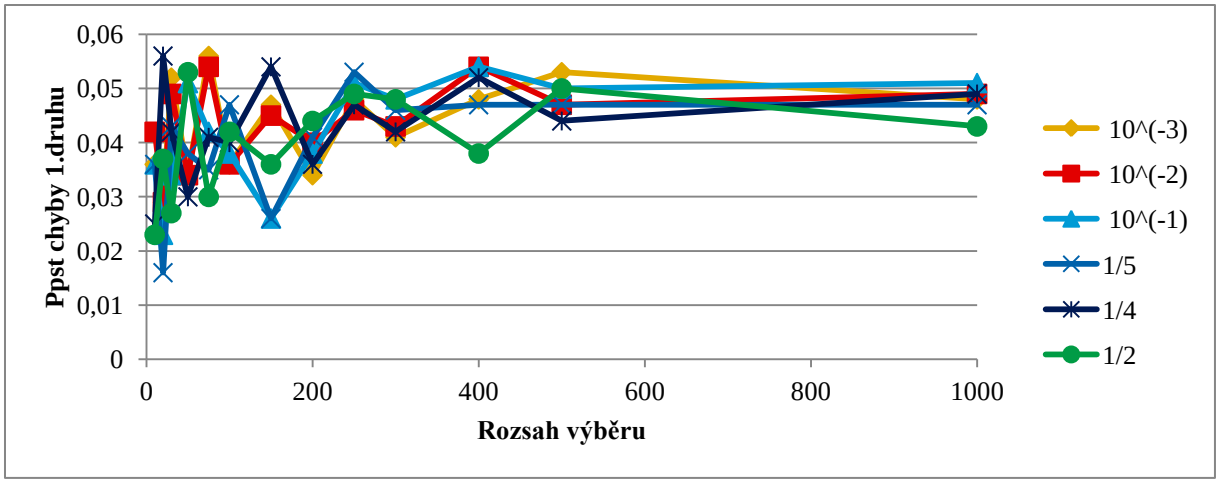
**Tabulka 5.1: Nasimulované kritické hodnoty pro výběry z normálního rozdělení**

Kritická hodnota $D^*_{m,n}(\alpha)$ pro zaokrouhlení $10^{-3}$ se pro všechny délky rozsahů jevila totožná jako bez zaokrouhlení. Z toho lze usuzovat, že zaokrouhlení na tři desetinná místa už je zanedbatelné. Z výsledků je vidět, že pro rozsah 300 se kritická hodnota nyní mění pro každý další stupeň zaokrouhlení. Grafické znázornění výsledků KS testu při zanedbání zaokrouhlení pro různé rozsahy náhodných výběrů z rozdělní s distribuční funkcí $N(0,1)$ je uvedeno na Obrázku 5.1. Každá lomená čára odpovídá jinému stupni zaokrouhlení.



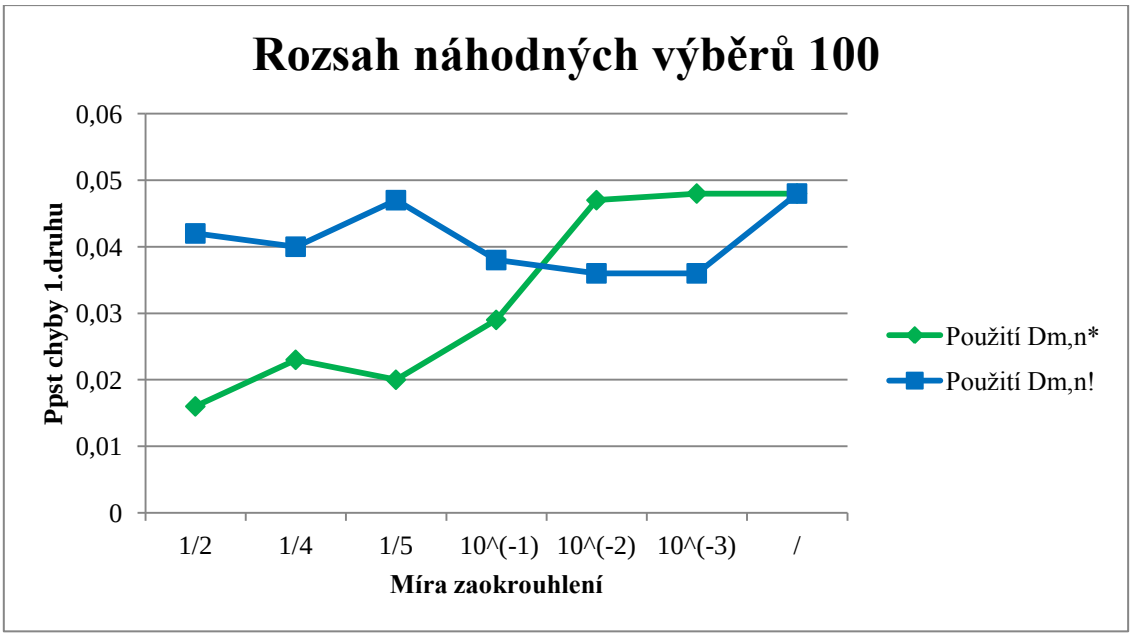

**Obrázek 5.1: Graf s výsledky KS testu s $D^*_{m,n}(\alpha)$, $N(0,1)$**

Z Obrázku 5.1 je možné vyčíst, že míra zaokrouhlení může mít velký vliv na výsledky dvouvýběrového KS testu. Pro dva náhodné výběry z rozdělní s distribuční funkcí $N(0,1)$ je vliv rozsahu výběrů menší přibližně do rozsahu 50. Od rozsahu 300 se hodnoty pravděpodobností chyb 1. druhu přibližně ustalují. V grafu lze pozorovat seřazení výsledků v závislosti na míře zaokrouhlení (největší zaokrouhlení odpovídá nejmenší pravděpodobnosti chyby 1. druhu). Všechny sady simulací KS testů uvedené na Obrázku 5.1 byly také vyhodnoceny se simulovanou hodnotou $D^!_{m,n}(\alpha)$. Výsledky jsou uvedeny na Obrázku 5.2.




**Obrázek 5.2: Graf s výsledky KS testu s $D^{!}_{m,n}(\alpha)$, $N(0,1)$**

V grafu nyní pozorujeme ustálení výsledků přibližně na hladině $\alpha = 0{,}05$. Při velikosti rozsahu do 200 je vidět stále kmitání kolem hodnoty 0,05. Nicméně všechny výsledky již jsou srovnatelné s výsledky KS testu bez zaokrouhlených hodnot. Použitím hodnoty $D^{!}_{m,n}(\alpha)$ jsme vzali v úvahu zaokrouhlení vstupních dat, a tím jsme dostali odpovídající vyhodnocení testu. Zvýraznění rozdílu použití mezi $D^{*}_{m,n}(\alpha)$ a $D^{!}_{m,n}(\alpha)$ je ukázáno na Obrázku 5.3.

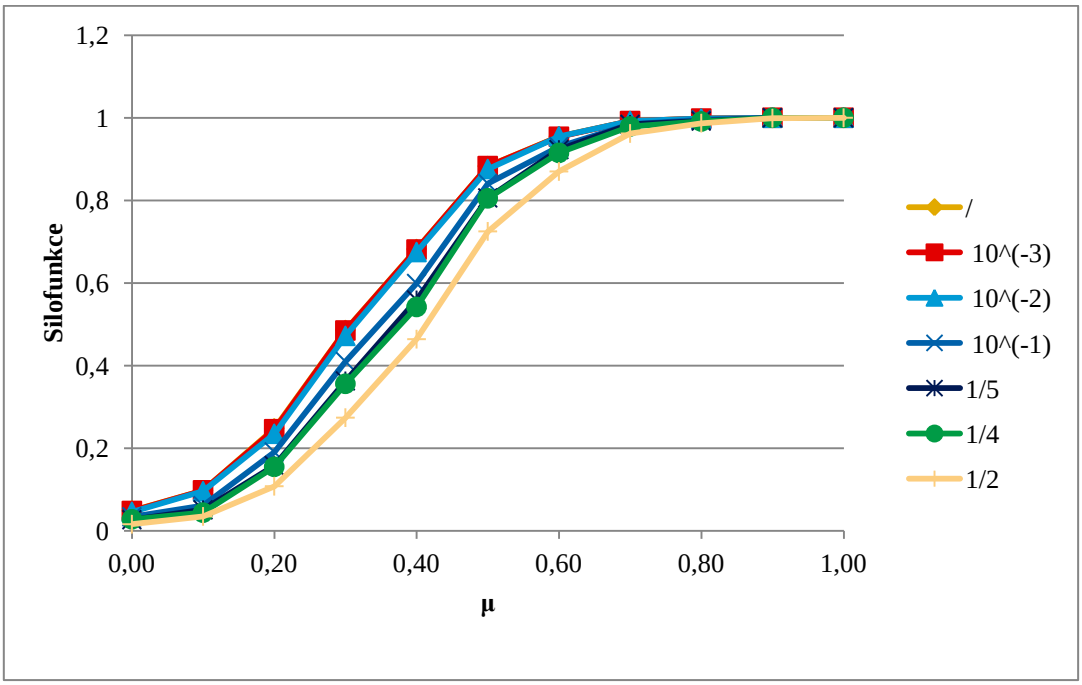



**Obrázek 5.3: Rozdíl výsledku KS testu s použitím $D^{*}_{m,n}(\alpha)$ a $D^{!}_{m,n}(\alpha)$, $N(0,1)$**

Rozdíl výsledků je ukázán na příkladě velikosti rozsahu výběrů 100. Nyní je možné pozorovat snižování rozdílu mezi výsledky se zmenšující se mírou zaokrouhlení.

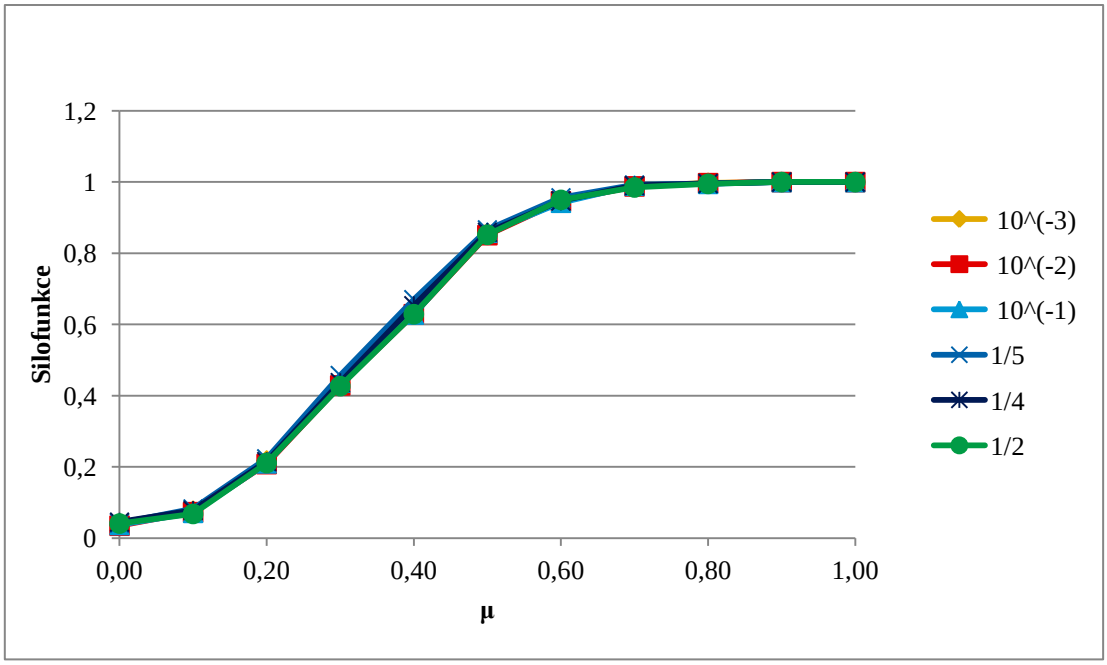
## 5.2 Změna $\mu$

V následujícím odstavci se bude aplikovat dvouvýběrový KS test pro jeden náhodný výběr z rozdělní s $N(0,1)$ a druhý s $N(\mu,1)$, kde parametr $\mu$ budeme měnit. Budeme sledovat, jak míra zaokrouhlení ovlivní výsledky testování dvou alternativ, které se liší posunutím $\mu$. Zvolený krok změny $\mu$ byl nastaven na 0,1. Při jiné volbě kroku se dosahovalo přibližně totožných výsledků. Pokud se testovala změna $\mu$ od $-1$ do 1, výsledky byly téměř symetrické (vlivem zaokrouhlení může dojít k nepatrným rozdílům), proto se v práci uvádí pouze výsledky pro interval $\langle 0,1\rangle$ a krok změny 0,1. Opět bude porovnávána varianta KS testu s $D^{*}_{m,n}(\alpha)$ (viz Obrázek 5.4) a s $D^{!}_{m,n}(\alpha)$ (viz Obrázek 5.5).



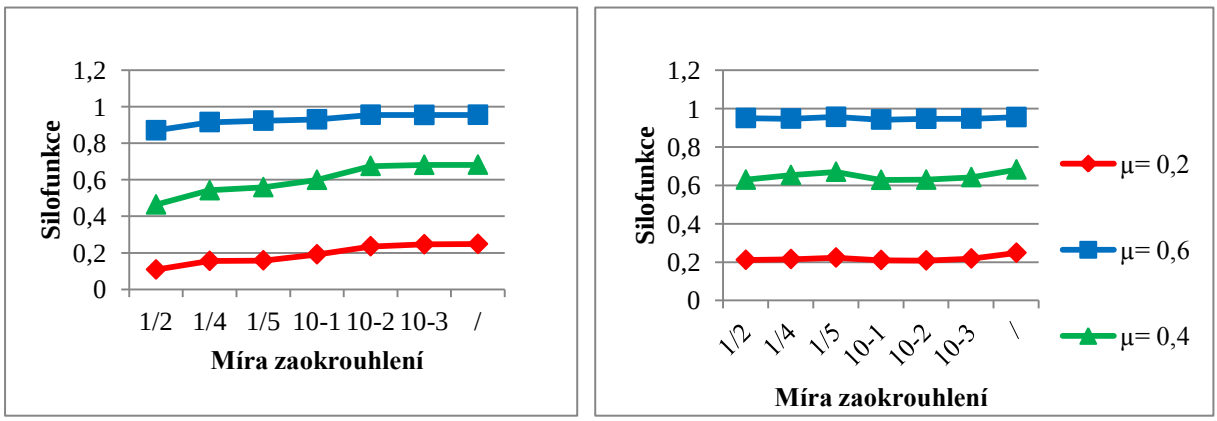

**Obrázek 5.4: Výsledky KS testu s $D^{*}_{m,n}(\alpha)$ pro náhodný výběr z rozdělní s distribuční funkcí pro $N(0,1)$ a náhodný výběr z rozdělní s distribuční funkcí pro $N(\mu,1)$ se změnou $\mu$**

Na Obrázku 5.4 je vidět vyhodnocení KS testu pro dva náhodné výběry z rozdělní s různou distribuční funkcí $N(0,1)$ a $N(\mu,1)$. V grafu lze pozorovat, že například pro hodnotu parametru $\mu = 0{,}4$ při zanedbání zaokrouhlení na poloviny správně zamítáme $H_0$ na hladině významnosti 5 % ve 46,4 % případů, přestože pro nezaokrouhlená vstupní data se zamítá celkem pro 68,2 % případů.




**Obrázek 5.5: Výsledky KS testu s $D^{!}_{m,n}(\alpha)$ pro náhodný výběr z rozdělní s distribuční funkcí pro $N(0,1)$ a náhodný výběr z rozdělní s distribuční funkcí pro $N(\mu,1)$ se změnou $\mu$**

Nyní po vyhodnocení testu s $D^{!}_{m,n}(\alpha)$ jsou výsledky všech sad testů prakticky totožné s výsledky testů s nezaokrouhlenými pozorováními. Konkrétně pro hodnotu parametru $\mu = 0{,}4$ při zaokrouhlení na poloviny správně zamítáme $H_0$ na hladině významnosti 5 % ve 62,9 % případů a pro nezaokrouhlená vstupní data zamítáme celkem ve 68,2 % případů. Výsledky se liší v řádech setin. Ještě se podíváme na výsledky podrobněji, viz Obrázek 5.6.

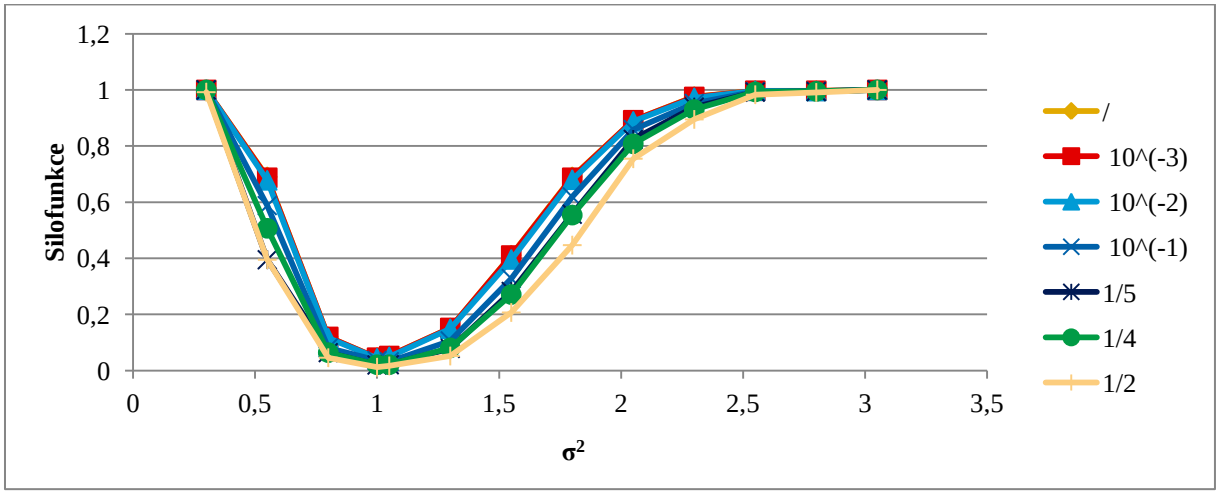



**Obrázek 5.6: Srovnání výsledků pro střední hodnoty $0,2; 0,4; 0,6$, vlevo výsledky s $D^{*}_{m,n}(\alpha)$ a vpravo s $D^{!}_{m,n}(\alpha)$**

Na Obrázku 5.6 lze porovnat rozdíl mezi vyhodnocením s $D^*_{m,n}(\alpha)$ a s $D^!_{m,n}(\alpha)$. Vlevo mají lomené čáry sklon v závislosti na míře zaokrouhlení. Vpravo naopak jsou lomené čáry téměř konstantní a vliv zaokrouhlení už není znatelný.

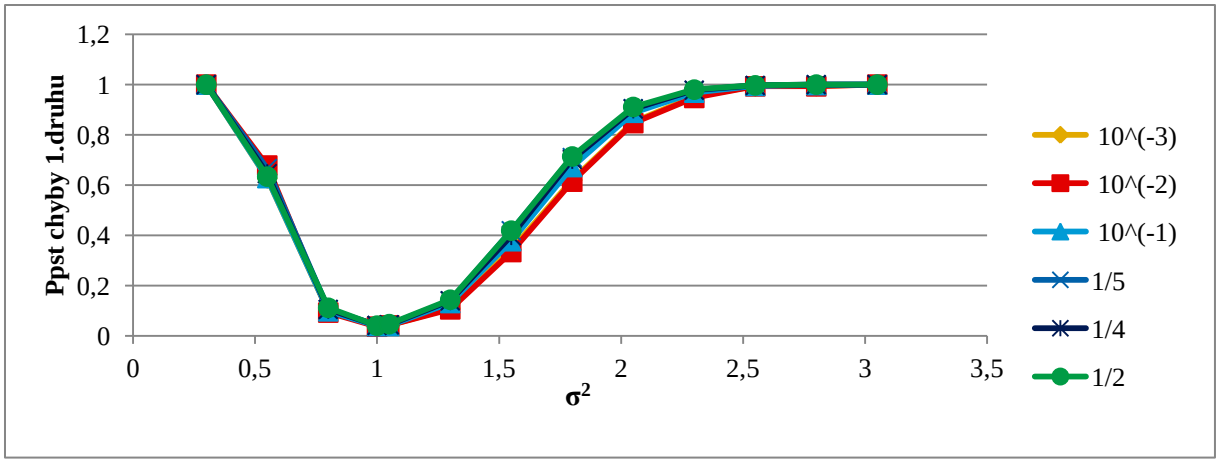
## 5.3 Změna $\sigma^2$

Jako další sledovaný parametr normálního rozdělení byl zvolen rozptyl. Dvouvýběrový KS test bude proveden pro jeden náhodný výběr z rozdělní s distribuční funkcí $N(0,1)$ a druhý s distribuční funkcí $N(0,\sigma^2)$, kde parametr $\sigma^2$ budeme měnit. Výsledky jsou uvedeny pro parametr od 0,3 do 3,05, krok byl zvolen 0,25.



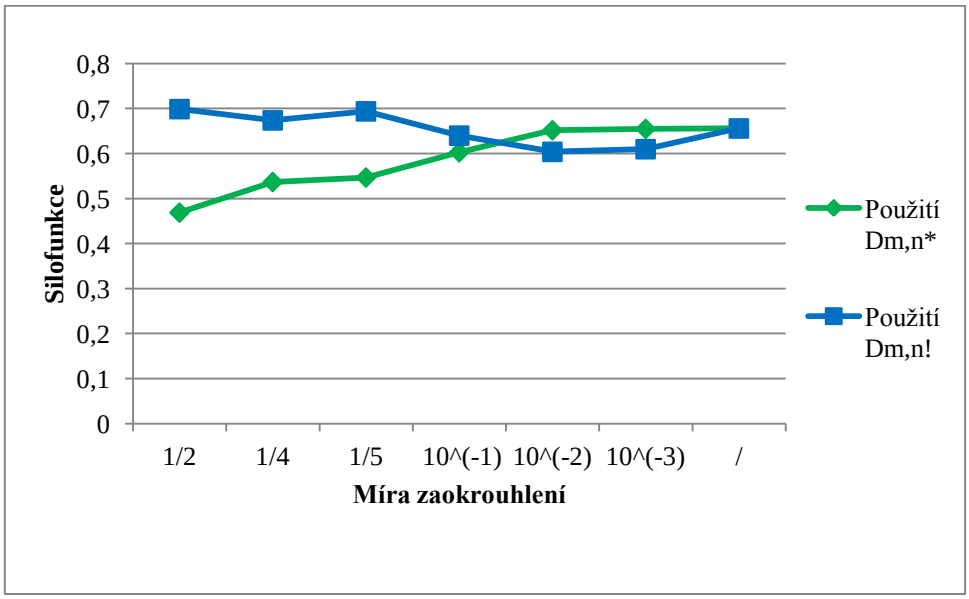

**Obrázek 5.7: Výsledky KS testu s $D^*_{m,n}(\alpha)$ pro náhodný výběr z rozdělní s distribuční funkcí pro $N(0,1)$ a náhodný výběr z rozdělní s distribuční funkcí pro $N(0,\sigma^2)$ se změnou $\sigma^2$**

Z Obrázku 5.7 vyplývá následující. Pokud se zanedbává zaokrouhlení na poloviny pro $\sigma^2 = 1{,}8$, tak se hodnota silofunkce od případu bez zaokrouhlení liší o 0,243. Vyhodnotí-li se však test s $D^!_{m,n}(\alpha)$, dostáváme rozdíl v silofunkci pro stejný případ už jen 0,024, tedy po přihlédnutí k zaokrouhlení téměř nulový, viz Obrázek 5.8.




**Obrázek 5.8: Výsledky KS testu s $D^{!}_{m,n}(\alpha)$ pro náhodný výběr z rozdělní s distribuční funkcí pro $N(0,1)$ a náhodný výběr z rozdělní s distribuční funkcí pro $N(0,\sigma^2)$ se změnou $\sigma^2$**

Na závěr kapitoly ještě bude ukázán příklad, kdy se změní oba parametry normálního rozdělení. Takových kombinací je ovšem nespočetně. V práci na Obrázku 5.9 je uvedena alespoň jedna varianta. Rozsah výběrů byl zvolen 100.




**Obrázek 5.9: Rozdíl výsledku KS testu s použitím $D^{*}_{m,n}(\alpha)$ a $D^{!}_{m,n}(\alpha)$, jeden výběr pochází z $N(0,1)$, druhý výběr pochází z $N(0,2;1,7)$**

V grafu jsou vyneseny silofunkce pro dvouvýběrový KS test. Jeden výběr pocházel z normálního normovaného rozdělení, druhý výběr z normálního rozdělení s parametry $\mu = 0{,}2$ a $\sigma^2 = 1{,}7$. Při použití kritické hodnoty se zanedbáním zaokrouhlením na poloviny, čtvrtiny, pětiny a $10^{-1}$ se dosahuje nižších hodnot silofunkce než při vyhodnocení testu

s $D^{!}_{m,n}(\alpha)$. Je možné také vyvozovat, že při snižování míry zaokrouhlení se rozdíl mezi vyhodnocením s $D^{!}_{m,n}(\alpha)$ a $D^{*}_{m,n}(\alpha)$ minimalizuje.

# 6 Šachisté

Motivací pro vznik práce byly výpočty z bakalářské práce Kocandové [1]. Studentka se snažila identifikovat vliv relativního věku u šachistů, tj. vysledovat vliv data narození na sportovní výsledky. V práci mimo jiné také uvádí výsledky pro hráče hokeje a fotbalu. K dispozici bylo však málo vstupních dat, proto i diplomová práce se zaměřuje na data z prostředí Šachového svazu České republiky. Všechna vstupní data uvedena v následující kapitole jsou čerpána z [1] a z [8]. K dispozici byly údaje ze dvou databází, z roku 2010 a z roku 2015. Šachisté jsou řazeni do různých mládežnických kategorií. Pro hochy to jsou kategorie $H10, H12, H14, H16, H18$ a $H20$, pro dívky $D10, D12, D14$ a $D16$. V textu jsou uvedeny výsledky pro kategorie $H10$, $H20$ (databáze z roku 2015) a pro smíšenou kategorii $HD10$ (databáze z roku 2010). Do kategorie $H10$ patří chlapci ve věku do 10 let. V kategorii $H20$ soutěží chlapci ve věku 19 a 20 let a kategorie $HD10$ vznikla spojením $H10$ a $D10$.

K testování vlivu relativního věku u šachistů byl použit dvouvýběrový KS test. Testovanými výběry o shodě rozdělení jsou šachisté a česká populace. V práci Kocandové při užití KS testu se předpokládá, že $X_1, \dots, X_m$ (měsíc narození šachistů v daném roce) je náhodný výběr pocházející ze spojitého rozdělení a $Y_1, \dots, Y_n$ (měsíc narození českých dětí v daném roce) je náhodný výběr ze spojitého rozdělení. Ovšem ve skutečnosti do testování vstupují data narození již zaokrouhlená právě na měsíce, tím dostáváme zaokrouhlené hodnoty $X_1^d, \dots, X_m^d$ a $Y_1^d, \dots, Y_n^d$. Proto pro správné vyhodnocení dvouvýběrového KS testu by měla být použita kritická hodnota $D_{m,n}^{!}(\alpha)$, která bere v úvahu vliv zaokrouhlení vstupních dat. V kapitole se budeme zajímat o rozdíl výsledku testování s použitím odhadnuté $D_{m,n}^{!}(\alpha)$ a s $D_{m,n}^{*}(\alpha)$.

Přibližná „správná“ kritická hodnota $D_{m,n}^{!}(\alpha)$ pro zaokrouhlená data bude získána opět simulačně obdobným postupem jako v kapitole 3.1. Nyní budou však různé rozsahy výběrů. Opět bylo provedeno $100010$ simulací. Při každé simulaci byly vygenerovány dva výběry z $R(0, 12)$ o rozsahu $m$ a $n$ (jednotlivé rozsahy jsou uvedeny v Tabulce 6.1). Předpokládá se totiž, že data narození dětí se řídí přibližně rovnoměrným rozdělením. Interval $(0, 12)$ byl zvolen, neboť vstupních data nabývají pouze hodnot $1, 2, \dots, 12$, kde 1 odpovídá měsíci narození leden, 2 odpovídá měsíci únor atd. Vygenerovaný výběr byl na „měsíce“ zaokrouhlen v software Matlab pomocí funkce $ceil(x)$, která reálné číslo $x$ zaokrouhlí na

nejbližší vyšší celé číslo nahoru. Ze sady 100010 testovacích statistik $D_{m,n}$ byla opět odhadnuta kritická hodnota $D^{!}_{m,n}(\alpha)$ 95% výběrovým kvantilem.

Vybrané kategorie a počty narozených dětí v daném roce jsou uvedeny v následující tabulce.

| Kategorie | Roky narození | Počet šachistů | Počet narozených dětí v ČR |
|---|---|---|---|
| **H10** | 2005, 2006 | 752 | 208042 |
| **H20** | 1995, 1996 | 232 | 186543 |
| **HD10** | 2000, 2001 | 572 | 181625 |

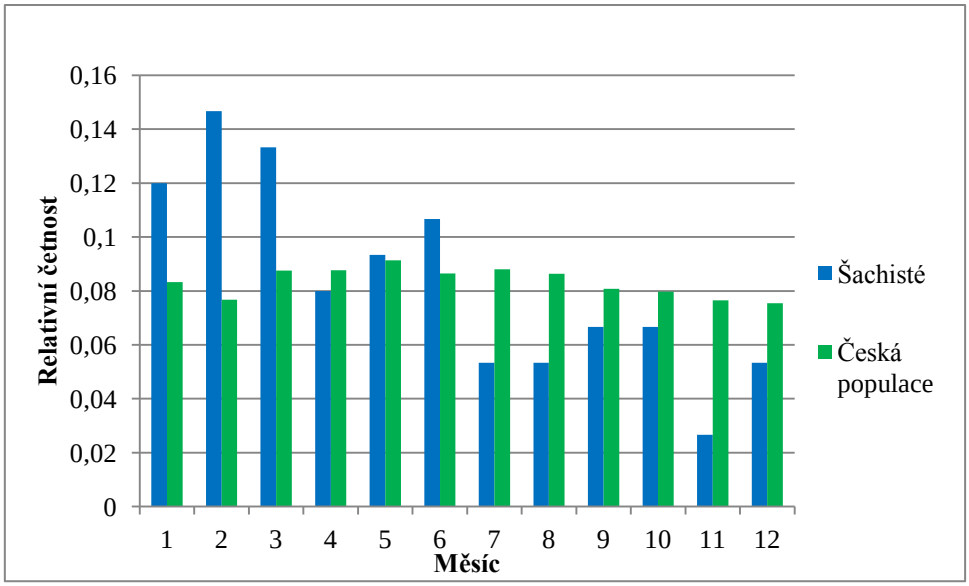
**Tabulka 6.1: Počty šachistů a všech českých dětí narozených v daném roce**

Z Tabulky 6.1 lze vyčíst, kolik šachistů patřilo v roce 2015 do kategorie $H10$ a $H20$, popřípadě v roce 2010 do kategorie $HD10$. Druhý výběr je vždy česká populace, odpovídající počty k roku narození jsou rovněž uvedeny v tabulce. Kocandová ve své práci při testování v některých případech vybrala prvních 75 nejlepších šachistů podle národního ela ([1] strana 3) v dané kategorii a ty podrobila testování. Při porovnávání budeme postupovat stejným způsobem.

Dále je kapitola členěna na podkapitoly podle testované kategorie šachistů.

## 6.1 Kategorie HD10

Jako první byla vybraná kategorie k porovnání výsledků $HD10$. Jedná se o chlapce a dívky ve věku do 10 let. Byla vybrána data narození nejlepších 75 šachistů a šachistek v dané kategorii. Histogram relativních četností narození šachistů a české populace je ukázán na Obrázku 6.1.




**Obrázek 6.1: Histogram relativních četností narození šachistů a české populace v letech 2000/2001 v daném měsíci**

Relativní četnosti uvedené v histogramu byly získány výpočtem $\frac{p}{k}$, kde $p$ je počet narození v daném měsíci a $k$ je počet narození dětí v celém roce. Pro šachisty je $k = 75$ a pro českou populaci $k = 181625$.

Pro získání odhadu kritické hodnoty $D^{!}_{m,n}(\alpha)$ byly tedy pro každou simulaci vygenerovány dva výběry z $R(0,12)$, jeden o rozsahu 75 a druhý o rozsahu 181625. Oba výběry byly zaokrouhleny výše popsaným způsobem. Odhad kritické hodnoty $D^{!}_{m,n}(\alpha)$ pro $\alpha = 0{,}05$ vyšel $0{,}1373$. Kritická hodnota pro nezaokrouhlená pozorování vyšla $D^{*}_{m,n}(\alpha) = 0{,}1543$. Testovací statistika pro dvouvýběrový KS test byla stanovena na hodnotu $D_{m,n} = 0{,}1668$. Platí tedy následující

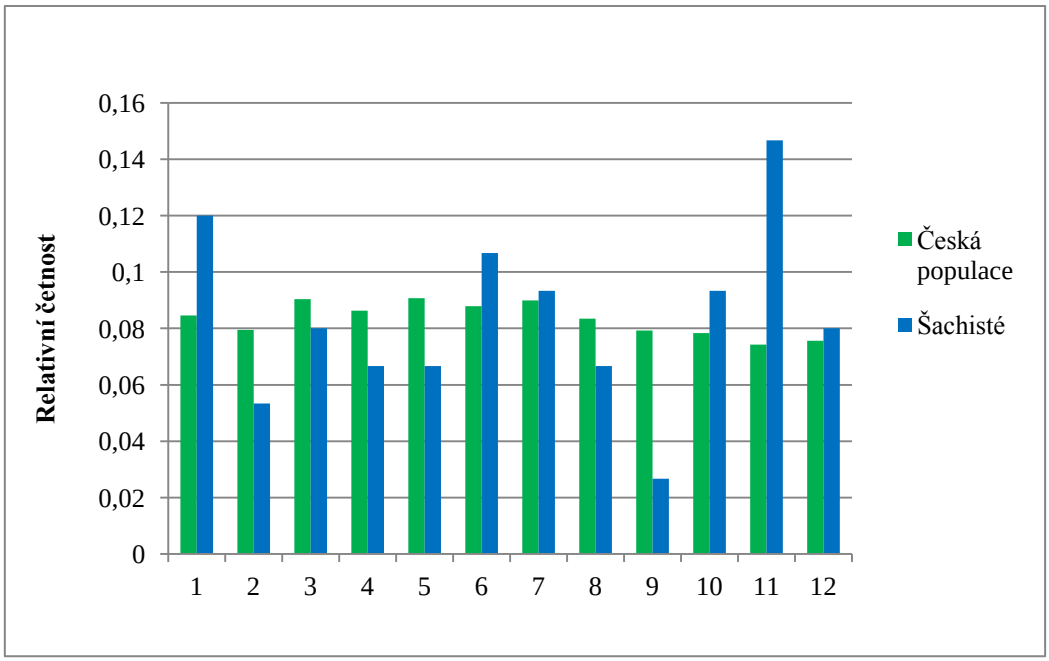
$$D^{*}_{75,181625}(0{,}05) = 0{,}1543 < D_{75,181625} = 0{,}1668,$$

$$D^{!}_{75,181625}(0{,}05) = 0{,}1373 < D_{75,181625} = 0{,}1668.$$

P-hodnota testu při vyhodnocení s $D^{*}_{m,n}(\alpha)$ vyšla $0{,}0271$ a při užití $D^{!}_{m,n}(\alpha)$ $0{,}0107$. Z dosažených výsledků lze tvrdit, že hypotéza o shodě rozdělení je zamítána na hladině významnosti 5 % v obou případech. Rozdíl v kritické hodnotě a testovací statistice pro případ se zanedbáním zaokrouhlení je v řádu setin.

## 6.2 Kategorie H20

Poslední mládežnickou kategorií je $H20$. V této kategorii soutěží nejstarší děti. Porovnání měsíce jejich narození s českou populací je uvedeno v Obrázku 6.2.




**Obrázek 6.2: Histogram relativních četností narození šachistů a české populace v letech 1995/1996 v daném měsíci**

Počet vybraných nejlepších šachistů je opět 75 a počet všech narozených dětí v české populaci v letech 1995 a 1996 je 186543. První výběr jsou měsíce narození 75 nejlepších šachistů z kategorie $H20$ a druhý výběr jsou měsíce narození 186543 českých dětí. Výběry se otestují dvouvýběrovým KS testem o shodě rozdělení. Bylo dosaženo následujících výsledků.

Odhadnutá kritická hodnota pro $\alpha = 0{,}05$ vyšla $D^{!}_{m,n}(\alpha) = 0{,}1370$, kritická hodnota pro nezaokrouhlená pozorování vyšla $D^{*}_{m,n}(\alpha) = 0{,}1543$. Testovací statistika je $D_{m,n} = 0{,}0918$. Je vidět, že platí následující

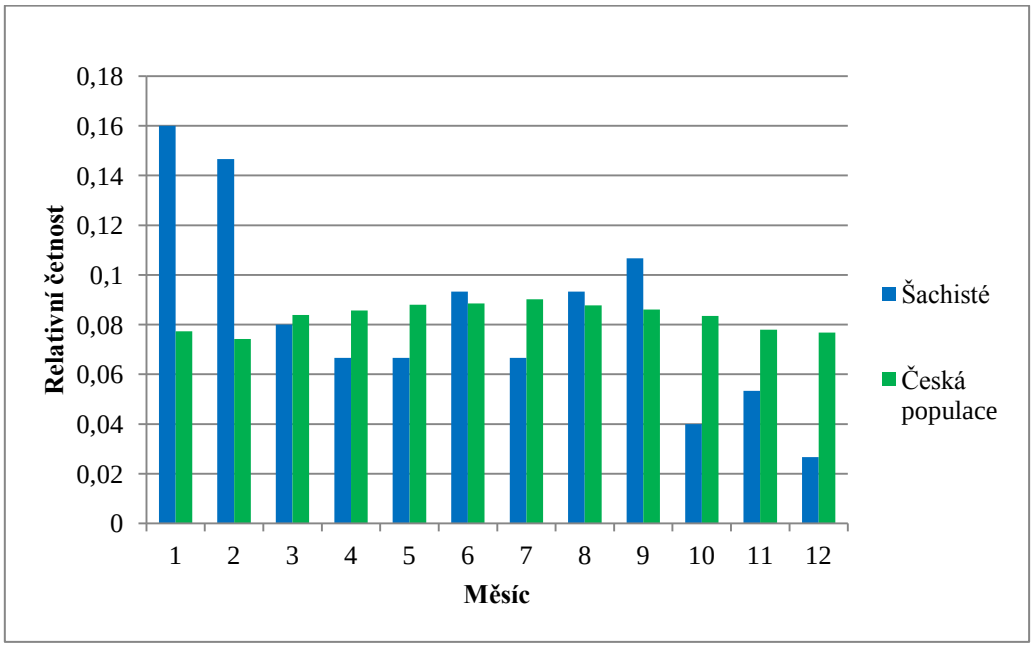
$$D^{*}_{75,186543}(0{,}05) = 0{,}1543 > D_{75,186543} = 0{,}0918,$$

$$D^{!}_{75,186543}(0{,}05) = 0{,}1370 > D_{75,186543} = 0{,}0918.$$

P-hodnota testu při aplikaci $D^{*}_{m,n}(\alpha)$ vyšla 0,5322 a při užití $D^{!}_{m,n}(\alpha)$ 0,3125. Z výsledků vyplývá, že v obou případech nulovou hypotézu o shodě rozdělení výběrů na hladině významnosti 5 % nezamítáme.

## 6.3 Kategorie H10

V následující kapitole budou porovnány výsledky dvouvýběrového KS testu pro kategorii $H10$. Byla opět vybrána data narození nejlepších 75 šachistů v dané kategorii. U nižších věků šachistů se předpokládá větší vliv věku na výsledek než u seniorských kategorií. Porovnání relativních četností narození šachistů a české populace v daném měsíci je uvedeno v Obrázku 6.3.




**Obrázek 6.3: Histogram relativních četností narození šachistů a české populace v letech 2005/2006 v daném měsíci**

Relativní četnosti v histogramu byly získány stejným způsobem jako v předchozím případě. Počet šachistů je 75 a dětí narozených v České republice v letech 2005 a 2006 je 208042.

Postup při testování je stejný jako u předchozích kategorií. Odhadnutá kritická hodnota vyšla pro $\alpha = 0{,}05$ $D^{!}_{m,n}(\alpha) = 0{,}1375$, ale kritická hodnota pro nezaokrouhlená pozorování vyšla $D^{*}_{m,n}(\alpha) = 0{,}1545$.

Vidíme, že platí následující

$$D^{*}_{75,208042}(0{,}05) = 0{,}1545 < D_{75,208042} = 0{,}1551,$$

$$D^{!}_{75,208042}(0{,}05) = 0{,}1375 < D_{75,208042} = 0{,}1551.$$

P-hodnota testu při vyhodnocení s $D^{*}_{m,n}(\alpha)$ vyšla 0,0485 a při užití $D^{!}_{m,n}(\alpha)$ 0,0205, proto v obou případech zamítáme hypotézu o shodě rozdělení na hladině významnosti 5 %.

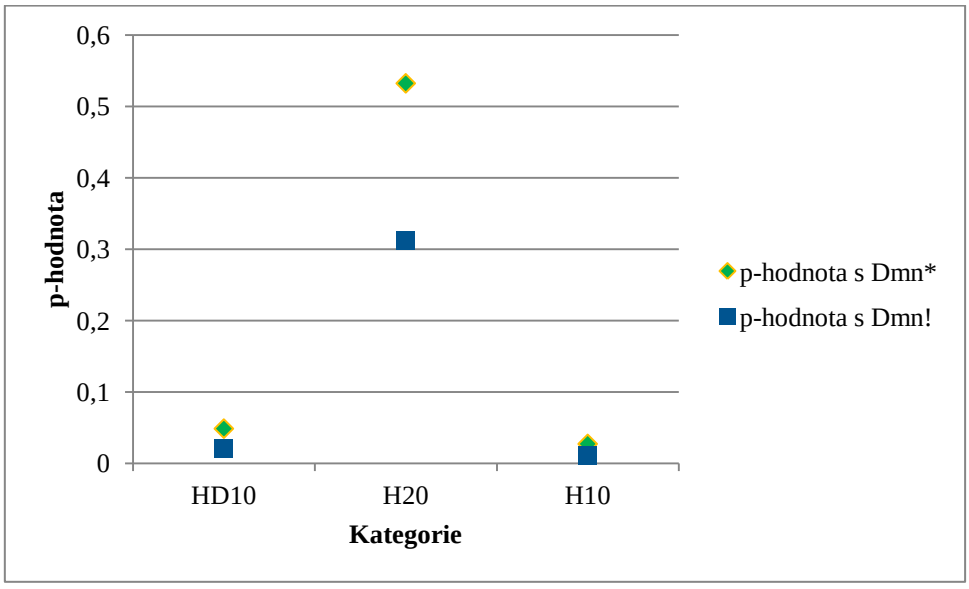
I v tomto případě je finální výsledek testování shodný pro oba případy, zamítá se nulová hypotéza. Avšak rozdíl v nerovnosti při užití $D^*_{m,n}(\alpha)$ je již velmi malý. Příslušná p-hodnota je také téměř na hranici 5 %, proto se podíváme na výsledky, sníží-li se hladina významnosti na 4 %. Bylo nutné znovu odhadnout kritickou hodnotu $D^!_{m,n}(\alpha)$ (v tomto případě však 96% výběrovým kvantilem). Výsledky byly následující

$$D^*_{75,208042}(0{,}04) = \ 0.1585 > D_{75,208042} = 0{,}1551,$$

$$D^!_{75,208042}(0{,}04) = 0.1413 < D_{75,208042} = 0{,}1551,$$

Z výsledků lze vyvozovat, že pří snížení hladiny významnosti testu z 5 % na 4 %, dochází ke změně vyhodnocení. Pokud se zanedbává zaokrouhlení vstupních dat, nulovou hypotézu o shodě rozdělení na hladině významnosti 4 % přijímáme. Tento výsledek se neshoduje s variantou, použije-li se odhadnutá kritická hodnotu $D^!_{m,n}(\alpha)$. Na příkladě lze pozorovat, jaký vliv může mít počáteční zanedbání zaokrouhlení výběru.

Na závěr uvedeme Obrázek 6.4 se souhrnnými výsledky pro všechny kategorie.




**Obrázek 6.4: Srovnání p-hodnot pro všechny kategorie šachistů**

Uvedený graf znázorňuje rozdíly v p-hodnotách testu, vyhodnotí-li se s $D^!_{m,n}(\alpha)$ nebo s $D^*_{m,n}(\alpha)$. Nejvíce nás zajímají p-hodnoty pohybující se kolem hodnoty 0,05, popřípadě kolem hodnoty 0,04. V těchto případech může dojít k odlišnému vyhodnocení testů (jako pro kategorii H10). Bylo by dobré tento graf sestrojit pro všechny výpočty uvedené v bakalářské

práci [1], ale k tomu by byly zapotřebí všechna zdrojová data, k dispozici byla data uvedená pouze v textu.

# Závěr

Cílem práce bylo vyšetřit vliv zaokrouhlení vstupních dat na výsledky dvouvýběrového Kolmogorovova-Smirnovova testu o shodě rozdělení. V první řadě jsme se zaměřili na definování a zavedení testu. Pro ukázku byl zmíněn i důkaz Smirnovovy věty, která pojednává o rozdělení testovací statistiky testu.

Nejdříve testování probíhalo na simulovaných datech. Pozorovanými parametry byla chyba 1. druhu, kritická hodnota a síla testu. Vždy byly porovnávány získané výsledky při zanedbání zaokrouhlení na vstupu a nezanedbání zaokrouhlení vstupních pozorování. Z výsledků simulací lze vyvozovat, že vliv zaokrouhlení vstupních dat má vliv na vyhodnocení dvouvýběrového Kolmogorovova-Smirnovova testu. Při simulaci odhadu kritické hodnoty pro zaokrouhlené výběry z rovnoměrného i normálního rozdělení byly zjištěny rozdílné hodnoty od kritické hodnoty pro nezaokrouhlený případ. Odhadnuté kritické hodnoty byly ve většině případů nižší než kritická hodnota pro nezaokrouhlený případ. Rozdíly kritických hodnot byly znatelné pro určité případy již od zaokrouhlení na dvě desetinná místa.

Ve čtvrté kapitole proběhlo testování hypotézy o shodě rozdělení, kdy jeden výběr pocházel z rovnoměrného rozdělení a druhý z lineárního rozdělení. Bylo zjištěno, že míra zaokrouhlení může zastřít rozdíl v rozděleních. Největší vliv na vyhodnocení testu měla nejhrubší volba zaokrouhlení. Z toho lze vyvozovat, že čím větší zaokrouhlení zanedbáme, tím větší nepřesnosti výsledku testu můžeme získat. Naopak z výsledků je možné vyvozovat, že zaokrouhlení na tři desetinná místa již bylo prakticky totožné jako při nezaokrouhlení.

Jako druhé zkoumané rozdělení bylo normální. Vliv zaokrouhlení (při vyhodnocení testu se zanedbáním zaokrouhlení) byl nepatrně menší pro malé rozsahy, výsledky se ustálily přibližně od rozsahu $300$. Zkoumání posunutí střední hodnoty a rozptylu od $N(0,1)$ potvrdilo závěry získané pro rovnoměrné rozdělení. Opět pokus při zanedbání zaokrouhlení vycházel rozdílně než pro nezanedbání. Hodnoty silofunkce (případ zanedbání zaokrouhlení) pro výběry zaokrouhlené na poloviny byly znatelně nižší než pro nezaokrouhlené. Naopak pro případ vyhodnocení s příslušnou odhadnutou kritickou hodnotou byly rozdíly v silofunkci již prakticky nulové.

V závěrečné kapitole byly poznatky ze simulací aplikované na reálná data. Omezili jsme se na testování tří kategorií šachistů $H10, HD10$ a $H20$. Postup byl analogický jako u simulací. Pro kategorie $HD10$ a $H20$ se vyhodnocení testu na hladině významnosti $5\ \%$ shodovalo s vlivem zaokrouhlení i bez něj. U poslední kategorie bylo vyhodnocení obou variant na hranici. Pokud se snížila hladina významnosti na $4\ \%$, tak pro zanedbání zaokrouhlení vstupních dat jsme hypotézu o shodě rozdělení nezamítali, ale pro vyhodnocení testu s odhadnutou kritickou hodnotou byla nulová hypotéza zamítnuta.

Z dosažených výsledků lze vyvozovat, že zanedbání vlivu zaokrouhlení vstupních dat může mít velký vliv na vyhodnocení dvouvýběrového Kolmogorovova-Smirnovova testu, a tím lze získat zkreslené závěry. Největší vliv byl pozorován při největší míře zaokrouhlení výběrů.

# Použitá literatura

[1] Kocandová, M.: *Srovnání vlivu relativního věku ve sportu*, bakalářská práce, Západočeská univerzita, 2015.

[2] Anděl, J.: *Základy matematické statistiky*, MATFYZPRESS, 2007.

[3] Likeš, J., Machek, J.: *Matematická statistika*, Praha SNTL, 1983.

[4] Rényi, A.: *Teorie pravděpodobnosti,* Academia Praha, 1972.

[5] Hájek, J., Šidák, Z., Sen, P., K.: *Theory of Rank Tests (Second Edition),* Academic Press, 1999.

[6] Hájek, J., Šidák, Z.: *Theory of Rank Tests (First Edition),* Academic Press, 1967.

[7] Steck, G., P.: *The Smirnov Two Sample Tests as Rank Tests,* 1969, [online, 20-04-2017], dostupné z: *https://projecteuclid.org/euclid.aoms/1177697516*

[8] Český statistický úřad, *Živě narozené děti podle kalendářních měsíců v letech 1950–2015*, [online, 26-2-2017], dostupné z: *https://www.czso.cz/csu/czso/demograficka-prirucka-2015*