UNIVERZITA KOMENSKÉHO V BRATISLAVE
FAKULTA MATEMATIKY, FYZIKY A INFORMATIKY

ODHADY MEDIÁNU V MODELOCH TEÓRIE NÁHODNÉHO VÝBERU

DIPLOMOVÁ PRÁCA

Bratislava 2014 Bc. Zuzana BIELAKOVÁ

UNIVERZITA KOMENSKÉHO V BRATISLAVE
FAKULTA MATEMATIKY, FYZIKY A INFORMATIKY

ODHADY MEDIÁNU V MODELOCH TEÓRIE NÁHODNÉHO VÝBERU

DIPLOMOVÁ PRÁCA

Študijný program:	Ekonomická a finančná matematika
Študijný odbor:	1114 Aplikovaná matematika
Školiace pracovisko:	Katedra aplikovanej matematiky a štatistiky
Vedúci práce:	doc. RNDr. Katarína Janková, CSc.

Bratislava 2014 **Bc.Zuzana BIELAKOVÁ**

Univerzita Komenského v Bratislave
Fakulta matematiky, fyziky a informatiky

ZADANIE ZÁVEREČNEJ PRÁCE

Meno a priezvisko študenta: Bc. Zuzana Bielaková
Študijný program: ekonomická a finančná matematika (Jednoodborové štúdium, magisterský II. st., denná forma)
Študijný odbor: 9.1.9. aplikovaná matematika
Typ záverečnej práce: diplomová
Jazyk záverečnej práce: slovenský

Názov: Odhady mediánu v modeloch teórie náhodného výberu / *Median estimation in sampling theory*

Cieľ: Spracovať známe výsledky pre bodový a intervalový odhad kvantilov používané v teórii náhodného výberu. Analyzovať situácie pre rôzne výberové schémy pravdepodobnostného náhodného výberu a simulačne overiť kvalitu používaných odhadov.

Vedúci: doc. RNDr. Katarína Janková, CSc.
Katedra: FMFI.KAMŠ - Katedra aplikovanej matematiky a štatistiky
Vedúci katedry: prof. RNDr. Daniel Ševčovič, CSc.

Dátum zadania: 25.01.2013

Dátum schválenia: 04.02.2013 prof. RNDr. Daniel Ševčovič, CSc.
garant študijného programu

.....
študent

.....
vedúci práce

Čestné prehlásenie Čestne prehlasujem, že som túto prácu vypracovala samostatne s využitím teoretických vedomostí a použitím uvedenej literatúry.

. . . . . .

Bratislava 28.04.2014 Zuzana Bielaková

Poďakovanie Chcem poďakovať predovšetkým svojej vedúcej práce, doc. RNDr. Kataríne Jankovej za jej nápady, pripomienky, cenné rady, ale aj za ľudský prístup, ústretovosť a čas, ktorý mne a mojej práci venovala. Veľká vďaka tiež patrí mojim blízkym za ich podporu a vytvorenie vhodného prostredia na písanie.

Abstrakt v štátnom jazyku

Bielaková, Zuzana: Odhady mediánu v modeloch teórie náhodného výberu [Diplomová práca]. Univerzita Komenského v Bratislave, Mlynská dolina 84248, Bratislava, Fakulta matematiky, fyziky a informatiky, Katedra aplikovanej matematiky a štatistiky.
Vedúci práce: doc. RNDr. Katarína Janková, CSc.
Bratislava, 2014.

Táto diplomová práca sa zaoberá skúmaním vlastností odhadov mediánu základného súboru použitím viacero prístupov. Ide o bodový a intervalový odhad pri viacerých typoch výberových schém, ktorými sa zaoberáme, ale aj o rôzne prístupy k odhadom pre medián (cez distribučnú funkciu, pomocou hustoty normálneho rozdelenia, cez hypergeometrické rozdelenie pri jednoduchom náhodnom výbere bez návratu). Ohľad berieme aj na rozsah základného súboru a prípadné rozdelenie základného súboru na oblasti. Ďalší prístup je cez kontingenčnú tabuľku a pohľad cez poradové štatistiky výberového súboru. Výsledky získané výpočtom porovnáme s hodnotami získanými simulovaním výberov z exponenciálneho rozdelenia a overíme spoľahlivosť získaných intervalov spoľahlivosti pre medián.

Kľúčové slová: medián, odhad, základný súbor, výberový súbor, interval spoľahlivosti

Abstract

Bielaková, Zuzana : Median estimation sampling theory [Master's thesis]. Comenius University Bratislava, Mlynska dolina 84248, Bratislava, Faculty of Mathematics, Physics, and Informatics, Department of Applied Mathematics and Statistics.
Supervisor: doc. RNDr. Katarína Janková, CSc.
Bratislava, 2014.

The main content of this thesis is to analyse properties of estimation of population median using more approaches. We deal with point and interval estimate for more forms of sampling designs and several approaches to estimate median (by distribution function, density of normal distribution, by hypergeometric distribution for simple random sampling without replacement). We take into consideration size of population and it's division into strata. Another view is examining confidence intervals of median based on contingence table and order statistics of the sample. We compare calculated results with results obtained from simulation of sample from exponential distribution and we verify confidence of obtained confidence intervals.

Keywords: median, estimation, population, sample, confidence interval

Obsah

Úvod

Základným súborom rozumieme množinu objektov, na ktorých sledujeme vlastnosti (znaky), ktoré vieme číselne ohodnotiť (napríklad pevnosť materálu u súčiastok, podiel osôb, rodín s určitou vlastnosťou pri mzdovom ohodnotení, ...). Z praktických, nákladových či časových dôvodov sa informácia o znaku získava iba z podmnožiny objektov, z tzv. výberového súboru.

Teória náhodného výberu sa zaoberá vlastnosťami odhadov získaných pomocou výberového súboru pre charakteristiky základného súboru pri viacerých spôsoboch konštrukcie výberového súboru.

Na začiatku zavedieme pojmy pravdepodobnostného výberu, pričom sa zameriame na odhad úhnu, ktorý ďalej použijeme na odhad distribučnej funkcie a kvantilu základného súboru. K tejto úlohe pristúpime z hľadiska odlišných konštrukcií výberového súboru. Zvlášť budeme pristupovať k úlohe pre súbory malého rozsahu a veľkého rozsahu, pri ktorom popíšeme asymptotické vlastnosti odhadov. V niektorých prípadoch sú odvodené výsledky približné, preto kvalitu odhadov preveríme simuláciami.

Pri prístupe k odhadom chceme brať ohľad okrem rozsahu základného súboru, aj na jeho rozdelenie do oblastí, a preto odvodenia pri výbere z jednej oblasti rozšírime aj na oblastný výber. Pre rôzne výberové schémy odvodíme interval spoľahlivosti skúmanej charakteristiky súboru. Pokúsime sa popísať viacero spôsobov určovania intervalu spoľahlivosti pre často pozorovaný druhý kvartil, medián, a to aj pri špeciálnych predpokladoch oblastného výberu.

Získané teoretické vzťahy aplikujeme pri výberoch zo základného súboru, daného konečným počtom prvkov z exponenciálneho rozdelenia, a porovnáme tak rôzne prístupy k odhadom. Zaujimavé môže byť aj overenie miery spoľahlivosti odvodených intervalov pre odhady.

1 Problematika výberu z konečného základného súboru

Uvažujme základný súbor (ZS) jednotiek ktorých je konečný počet, matematicky označíme túto množinu jednotiek $\mathbf{U} = \{1, 2, \ldots, N\}$ (ide napr. o kusy vyrobených súčiastok, osoby v určitej vekovej skupine, ...). Na týchto objektoch súboru konečného rozsahu, $\{y_k : k \in \mathbf{U}\}$, budeme pozorovať určitý znak (napr. pevnosť súčiastok, mzdové ohodnotenie osôb, ...). U znakov sa budeme zaujímať o ich charakteristiky (stupeň odolnosti voči tlaku, podiel osôb s danou vlastnosťou, ...).

Informácie o znaku sa už zo spomínaných dôvodov získavajú iba na danej podmnožine súboru, výberovom súbore (VS), ktorého spôsob tvorenia sa určí výberovými schémami. Medzi základné výberové schémy patrí:

- Systematický výber:
  Vychádzame už z vytvoreného usporiadania jednotiek v základnom súbore. Vo všeobecnosti je charakterizovaný dĺžkou kroku a, podľa ktorého vyberáme objekty do výberového súboru. Prvý objekt je vybraný náhodne spomedzi prvých a objektov, každý ďalší je vzdialený dĺžku kroku od predchádzajúceho. Takto realizujeme výber až po koniec základného súboru.

- Jednoduchý náhodný výber:
  Jednotky sa náhodne vyberajú zo základného súboru. Môže ísť o jednoduchý náhodný výber s návratom alebo častejšie používaný, bez návratu, kedy sa jednotka po vybratí naspäť do základného súboru už nevracia.

- Oblastný výber:
  ZS rozdelíme na disjunktné oblasti a v každej nezávisle uskutočníme jednoduchý náhodný výber.

Tieto postupy sú podrobnejšie popísané v [7] pričom v tejto publikácii ako aj vo väčšine iných sa autor sústreďuje na odhad priemeru základného súboru, prípadne úhrnu. V porovnaní s priemerom v mnohých prípadoch, najmä v prípade výberu zo šikmého rozdelenia sa dáva prednosť použitiu mediánu. V našej práci sa zaoberáme práve týmito odhadmi.

2 Pravdepodobnostný výber a rôzne výberové schémy

Prístup pravdepodobnostného náhodného výberu, [7], rozumie pod výberovým súborom náhodnú podmnožinu základného súboru, označme S, pričom $P(S = s) = p(s)$ pre $s \in 2^{\mathbf{U}}$. Teda pravdepodobnostný výber, je daný rozdelením náhodnej veličiny S, ktorá je daná výberovou funkciou $p(.)$. Realizáciou výberu potom získame podmnožinu s, rozsahu n, ktorá predstavuje VS zo základného súboru $\mathbf{U}$, rozsahu N, s pravdepodobnosťou $p(s)$, kde:

$$p(.) : 2^{\mathbf{U}} \Rightarrow [0, 1],$$

$\sum_{s \in 2^{\mathbf{U}}} p(s) = 1$, $p(s) \geq 0$.

Indikátorom výberu k-tej jednotky rozumieme funkciu náhodnej premennej S:

$$I_k(S) = \begin{cases} 1 & k \in S, \\ 0 & k \notin S. \end{cases} \tag{1}$$

Vektorovo zapíšeme N týchto indikátorov zahrnutia objektov základného súboru:

$$I_S = (I_1(S), \ldots, I_k(S), \ldots, I_N(S))'.$$

Udalosť $S = s$ tak vieme prepísať ako $I_S = i_s$, kde :

$$i_s = (i_{1s}, \ldots, i_{ks}, \ldots, i_{Ns})'$$

a

$$i_{ks} = \begin{cases} 1 & k \in s, \\ 0 & k \notin s, \end{cases}$$

preto $p(s) = P(S = s) = P(I_S = i_s)$.

Ďalej označme pravdepodobnosti zahrnutia [1] :

$$\pi_k = P(k \in s) = P(I_k(S) = 1) = \sum_{s \ni k} p(s), \tag{2}$$

$$\pi_{kl} = \begin{cases} P(k \in s, l \in s) = P(I_k(S) I_l(S) = 1) = \sum_{s \ni (k,l)} p(s) & pre\, k \neq l, \\ P(k \in s, k \in s) = P(I_k(S)^2 = 1) = P(I_k(S) = 1) = \pi_k & pre\, k = l. \end{cases} \tag{3}$$

[1] $s \ni k$ označuje všetky také výbery, ktoré zahŕňajú y_k,

$s \ni (k, l)$ označuje všetky také výbery, ktoré zahŕňajú y_k a y_l

Rôznou voľbou $p(s)$ dostaneme už nami spomenuté schémy, ale aj nové výberové schémy:

- Bernoulliho výber

  Ide o výber v ktorom sú indikátory $I_k(S)$ nezávislé, rovnako rozdelené s parametrom $\pi \in [0,1]$, ktorý predstavuje pravdepodobnosť zahrnutia každého objektu v základnom súbore do výberového súboru. Pre pravdepodobnosti zahrnutia pri Bernoulliho výbere platí:

  $$\pi_k = \pi, \tag{4}$$

  $$\pi_{kl} = \begin{cases} \pi^2 & k \neq l, \\ \pi & k = l. \end{cases} \tag{5}$$

- Modifikovaný systematický výber

  Je daný veľkosťou kroku $a = \frac{N}{n}$, aby $a \in Z$. Ďalej definujeme jeden z typov systematického výberu. Náhodne zvolíme $m > 1$, aby $\frac{n}{m} \in Z$ a zrealizujeme náhodný výber m prvkov $\{y_{k_1}, \ldots, y_{k_m}\}$ spomedzi prvkov základného súboru $\{y_1, \ldots, y_{ma}\}$. Do výberového súboru ďalej vstupuje každý ma-ty prvok od prvých, náhodne zvolených, až po koniec základného súboru. Výber s potom zahŕňa objekty:

  $$s = \{Y_k : k = k_i + (j-1)ma, i = 1, \ldots m, j = 1, \ldots n/m\}$$

  Pre pravdepodobnosti zahrnutia objektov do výberového súboru platí:

  $$\pi_k = \frac{m}{ma} = \frac{1}{a}, \tag{6}$$

  $$\pi_{kl} = \begin{cases} \frac{1}{a} & |k-l| = amb\,, b = \{1, 2, \ldots\}, \\ \frac{1}{a}\left(\frac{m-1}{ma-1}\right) & inak. \end{cases} \tag{7}$$

- Jednoduchý náhodný výber bez návratu

  Pri jednoduchom náhodnom výbere bez vrátenia rozsahu n, je pre ľubovoľnú podmnožinu $s \in 2^S$ výberová funkcia : $p(s) = P(S = s) = \frac{1}{\binom{N}{n}}$ a pre pravdepodobnosti zahnutia platí:

  $$\pi_k = \frac{\binom{N-1}{n-1}}{\binom{N}{n}} = \frac{n}{N}, \tag{8}$$

$$\pi_{kl} = \begin{cases} \frac{\binom{N-2}{n-2}}{\binom{N}{n}} = \frac{n}{N}\frac{n-1}{N-1} & k \neq l, \\ \frac{n}{N} & k = l. \end{cases} \tag{9}$$

Príklad 2.1. *Pre výber rozsahu $n = 2$ z množiny $S = \{1, 2, 3\}$ dostávame podmnožiny $s = \{1, 2\}, s = \{2, 3\}, s = \{1, 3\}$ každú s pravdepodobnosťou $\frac{1}{3}$.*

Ďalej uvedieme vlastnosti indikátorov funkcií.

Lema 2.1. *Pre ľubovoľnú výberovú funkciu $p(s)$ a pre každé $k = \{1, \ldots, N\}$ platí:*

$$E(I_k(S)) = \pi_k.$$

$$V(I_k(S)) = \pi_k(1 - \pi_k)$$

$$C(I_k(S), I_l(S)) = \pi_{kl} - \pi_k\pi_l = \Delta_{kl}$$

Dôkaz. *Využitím vlastností pravdepodobnosti indikátorov :*
$P(I_k(S) = 1) = \pi_k$, $P(I_k(S) = 0) = 1 - \pi_k$,
vieme nasledovne odvodiť požadované charatkteristiky:
$E(I_k(S)) = 1\pi_k + 0(1 - \pi_k) = \pi_k$,
$E(I_k^2(S)) = E(I_k(S)) = \pi_k$, $V(I_k(S)) = E(I_k^2(S)) - (E(I_k(S)))^2 = \pi_k(1 - \pi_k)$,
$C(I_k(S), I_l(S)) = E(I_k(S)I_l(S)) - E(I_k(S))E(I_l(S)) = \pi_{kl} - \pi_k\pi_l$. □

Indikátory zahrnutia sú najjednoduchšie prípady štatistík, ktoré sa vyskytujú v pravdepodobnostnom náhodnom výbere. Spomenieme ešte niektoré ďalšie a uvedieme ich základné vlastnosti. Jednou z nich je náhodná veličina predstavujúca veľkosť výberu, $n_s = \sum_{k \in \mathbf{U}} I_k(S)$, ďalej výberový úhrn $\hat{t}_\pi = \sum_{k \in s} \frac{y_k}{\pi_k}$, výberový priemer, výberová distribučná funkcia či výberový medián ktorým sa budeme zaoberať podrobnejšie neskôr.

Lema 2.2. *Pre štatistiku n_s, veľkosť výberu, platí :*

$$E(n_s) = \sum_{k \in \boldsymbol{U}} \pi_k,$$

$$V(n_s) = \sum_{k \in \boldsymbol{U}} \pi_k - \left(\sum_{k \in \boldsymbol{U}} \pi_k \right)^2 + \sum_{k \in \boldsymbol{U}} \sum_{l \in \boldsymbol{U}, k \neq l} \pi_{kl}.$$

Dôkaz. *Podľa vlastnosti* $I_k(S)$*:*
$E(n_s) = E(\sum_{k\in \boldsymbol{U}} I_k(S)) = \sum_{k\in \boldsymbol{U}} E(I_k(S)) = \sum_{k\in \boldsymbol{U}} \pi_k$,
$V(n_s) = V(\sum_{k\in \boldsymbol{U}} I_k(S)) = \sum_{k\in \boldsymbol{U}} \pi_k(1-\pi_k) + \sum_{k\in \boldsymbol{U}} \sum_{l\in \boldsymbol{U}, k\neq l} (\pi_{kl} - \pi_k\pi_l) = \sum_{k\in \boldsymbol{U}} \pi_k -$
$(\sum_{k\in \boldsymbol{U}} \pi_k)^2 + \sum_{k\in \boldsymbol{U}} \sum_{l\in \boldsymbol{U}, k\neq l} \pi_{kl}$. □

Pre naše ďalšie úvahy budú zaujimavé vlastnosti úhrnu a jeho odhadu, $\hat{t}_\pi = \sum_{k\in s} \frac{y_k}{\pi_k}$.

Lema 2.3. *Nevychýlený odhad úhrnu:*

$$t = \sum_{k\in \boldsymbol{U}} y_k,$$

predstavuje štatistika:

$$\hat{t}_\pi = \sum_{k\in s} \check{y}_k$$

s varianciou:

$$V(\hat{t}_\pi) = \sum_{k\in \boldsymbol{U}} \sum_{k\in \boldsymbol{U}} \Delta_{kl} \check{y}_k \check{y}_l,$$

ktorej nevychýlený odhad je:

$$\hat{V}(\hat{t}_\pi) = \sum_{k\in s} \sum_{l\in s} \check{\Delta}_{kl} \check{y}_k \check{y}_l,$$

kde $\check{\Delta}_{kl} = \frac{\Delta_{kl}}{\pi_{kl}}$ *a* $\check{y}_k = \frac{y_k}{\pi_k}$.

Dôkaz. *Nevychýlenosť úhrnu vyplýva z:*
$E\left(\sum_{k\in s} \frac{y_k}{\pi_k}\right) = E\left(\sum_{k\in \boldsymbol{U}} \frac{y_k}{\pi_k} I_k(S)\right) = \sum_{k\in \boldsymbol{U}} \frac{y_k}{\pi_k} E(I_k(S)) = \sum_{k\in \boldsymbol{U}} \frac{y_k}{\pi_k} \pi_k = \sum_{k\in \boldsymbol{U}} y_k$.
Pre varianciu úhrnu platí:
$V(\hat{t}_\pi) = V(\sum_{k\in s} \check{y}_k) = V(\sum_{k\in \boldsymbol{U}} \check{y}_k I_k(S)) = \sum_{k\in \boldsymbol{U}} \check{y}_k^2 V(I_k) + \sum_{k\in \boldsymbol{U}} \sum_{l\in \boldsymbol{U}, k\neq l} \Delta_{kl} \check{y}_k \check{y}_l =$
$\sum_{k\in \boldsymbol{U}} \check{y}_k^2 \Delta_{kk} + \sum_{k\in \boldsymbol{U}} \sum_{l\in \boldsymbol{U}, k\neq l} \Delta_{kl} \check{y}_k \check{y}_l = \sum_{k\in \boldsymbol{U}} \sum_{l\in \boldsymbol{U}} \Delta_{kl} \check{y}_k \check{y}_l$.
Ďalej platí : $\hat{V}(\hat{t}_\pi) = \sum_{k\in \boldsymbol{U}} \sum_{l\in \boldsymbol{U}} I_k(S) I_l(S) \check{\Delta}_{kl} \check{y}_k \check{y}_l$.
Podľa rovností:

$$E(\hat{V}(\hat{t}_\pi)) = \sum_{k\in \boldsymbol{U}} \sum_{l\in \boldsymbol{U}} \check{\Delta}_{kl} \check{y}_k \check{y}_l E(I_k(S) I_l(S)) = \sum_{k\in \boldsymbol{U}} \sum_{l\in \boldsymbol{U}} \check{\Delta}_{kl} \check{y}_k \check{y}_l \pi_{kl} = V(\hat{t}_\pi)$$

ide o nevychýlený odhad variancie. □

Definícia 2.1. *Budeme hovoriť, že rozsah výberu je pevne daný, ak existuje* n_s *také, že* $p(s) = 0$ *pre také* s *, kde* $|s| \neq n_s$.

Lema 2.4. *Za predpokladu, že rozsah výberu je pevne daný, môžme alternatívne varianciu pre odhad daný pravdepodobnosťami π určiť ako:*

$$\hat{V}(\hat{t}_\pi) = -\frac{1}{2}\sum_{k\in \boldsymbol{U}}\sum_{l\in \boldsymbol{U}} \check{\Delta}_{kl}(\check{y}_k - \check{y}_l)^2.$$

Dôkaz. *Za predpokladu, že veľkosť výberu n je pevne daná môžme rozpísať vzťah (2.4):*

$$\hat{V}(\hat{t}_\pi) = \sum_{k\in \boldsymbol{U}}\sum_{l\in \boldsymbol{U}} \check{\Delta}_{kl}\check{y}_k\check{y}_l - \sum_{k\in \boldsymbol{U}}\sum_{l\in \boldsymbol{U}} \check{\Delta}_{kl}(\check{y}_k)^2.$$

Z toho vieme časť upraviť ako:

$$\sum_{k\in \boldsymbol{U}}\sum_{l\in \boldsymbol{U}} \check{\Delta}_{kl}(\check{y}_k)^2 = \sum_{k\in \boldsymbol{U}}(\check{y}_k)^2\sum_{l\in \boldsymbol{U}} \check{\Delta}_{kl}.$$

Po dosadení:

$$\sum_{l\in \boldsymbol{U}} \check{\Delta}_{kl} = \sum_{l\in \boldsymbol{U}} \pi_{kl} - \sum_{l\in \boldsymbol{U}} \pi_k\pi_l = n\pi_k - n\pi_k = 0,$$

získavame pôvodný vzťah (2.3). □

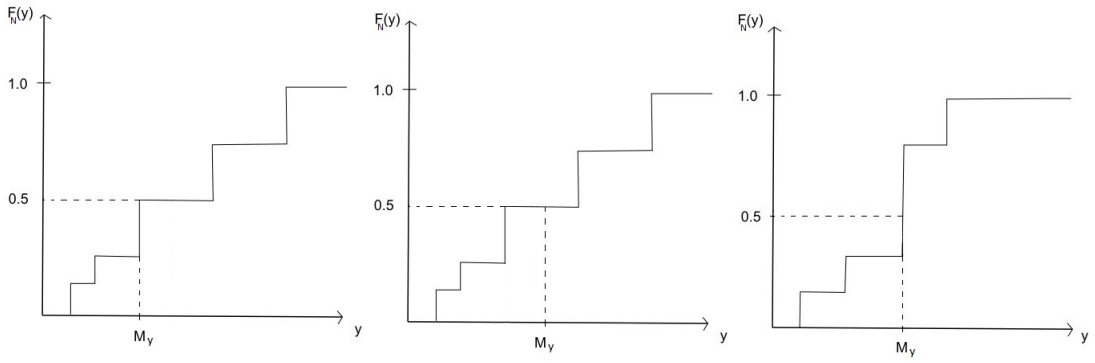
Lema 2.5. *Nevychýlený odhad veľkosti ZS, N, je $\sum_{k\in s}\frac{1}{\pi_k}$.*

Dôkaz. $E(\sum_{k\epsilon s}\frac{1}{\pi_k}) = E(\sum_{k\in \boldsymbol{U}}\frac{1}{\pi_k}I_k(S)) = \sum_{k\in \boldsymbol{U}}\frac{1}{\pi_k}E(I_k(S)) = \sum_{k\in \boldsymbol{U}}\frac{1}{\pi_k}\pi_k = N$ □

3 Odhad pre p kvantil

3.1 P kvantil pri výbere z jednej oblasti

Medián M_y základného súboru, je definovaný ako polovičný kvantil, $p = 0.5$. Je to prvok základného súboru, ktorý rozdeľuje prvky súboru na polovicu, polovica je menších ako M_y a polovica väčších. Ak medián nie je určený jednoznačne a danej podmienke vyhovuje interval objektov, prvé dva grafy na Obr.1, medián určíme buď ako infimum, obrázok vľavo, alebo ako jeho stred, obrázok v strede. Obrázok vpravo zobrazuje medián určený jednoznačne.



Obr. 1: Medián základného súboru

Vo všeobecnosti platí, že funkcia $F_N(y)$ vyjadruje, aký podiel prvkov v celom súbore je menších alebo rovných ako pevne dané y.
Označme:

$$Z_{k,y} = \begin{cases} 1 & ak\, y_k \leq y, \\ 0 & ak\, y_k > y. \end{cases} \tag{10}$$

Potom distribučnú funkciu $F_N(y)$ v bode y vieme zapísať:

$$F_N(y) = \frac{1}{N} \sum_{k \in \mathbf{U}} Z_{k,y} \tag{11}$$

a jej nevychýlený odhad pri pravdepodobnostnom výbere s pravdepodobnosťami zahrnutia $\pi_k, k \in \mathbf{U}$:

$$F_n(y) = \frac{\sum_{k \in s} \frac{Z_{k,y}}{\pi_k}}{\sum_{k \in s} \frac{1}{\pi_k}}. \tag{12}$$

Pomocou realizácie pravdepodobnostného výberu s rozsahu n, ktorý je daný prvkami základného súboru $\{y_k : k \in s\}$, vieme odhadnúť ľubovoľný p kvantil základného súboru rozsahu N, označme $r_{y,p}$ kde $p \in [0,1]$. Ak pri nejednoznačnosti zvolíme hodnotu ako infimum z nadobúdajúcich hodnôt, tak z vlastností distribučnej funkcie vieme p kvantil vyjadriť nasledovne:

$$r_{y,p} = F_N^{-1}(p), \tag{13}$$

kde

$$F_N^{-1}(p) = inf\{y : F_N(y) \geq p\}. \tag{14}$$

Na odhad $r_{y,p}$, označme $\hat{r}_{y,p}$, preto potrebujeme odhad distribučnej funkcie $F_N(y)$, označme $F_n(y)$ a preto:

$$\hat{r}_{y,p} = F_n^{-1}(p), \tag{15}$$

kde:

$$F_n^{-1}(p) = inf\{y : F_n(y) \geq p\} \tag{16}$$

predstavuje p kvantil výberového súboru.

Pri výpočtoch odhadu mediánu môžeme použiť postup, kedy najskôr zoradíme prvky výberového súboru $\{Y_k : k \in s\}$ do neklesajúcej postupnosti:

$$Y_{(1)} \leq Y_{(2)} \leq \cdots \leq Y_{(n)}$$

a priradíme im príslušné pravdepodobnosti zahrnutia,označme: $\pi_{1,s}, \pi_{2,s}, \ldots, \pi_{n,s}$.
Potom pomocou postupnosti $B_0 = 0, B_l = \sum_{j=1}^{l} \frac{1}{\pi_{j,s}}, B_n = \hat{N}$ vieme určiť odhad mediánu z objektov výberu s, v prípade nejednoznačnosti ako infimum:

$$\hat{M}_y = \begin{cases} Y_{(t)} & B_{t-1} < 0.5\hat{N} < B_t \\ Y_{(t-1)} & B_{t-1} = 0.5\hat{N} \end{cases} \tag{17}$$

Príklad 3.1. *Náhodný výber, rôzne pravdepodobnosti zahrnutia.*
Výber rozsahu $n = 5$ *zo súboru rozsahu* $N = 50$ *zahŕňa* $s = \{17, 2, 33, 29, 28\}$*:*

k	y_k	π_k	$\frac{1}{\pi_k}$
17	8	0.10	10
2	4	0.10	10
33	15	0.05	20
29	12	0.10	10
28	8	0.20	5
$\sum$			55

Podľa vzťahu (12) potom vieme určiť:

k	Y_k	$F_n(Y_k)$
1	8	$(10+10+5)/55 = 25/55$
2	4	$10/55$
3	15	$(10+10+5+10+20)/55 = 1$
4	12	$(10+10+5+10)/55 = 35/55$
5	8	$(10+10+5)/55 = 25/55$

$$\hat{r}_{y,0.5} = F_n^{-1}(0.5) = inf\{Y_k : F_n(Y_k) \geq 0.5\} = 12$$

Alternatívne:

k	$Y_{(k)}$	$\frac{1}{\pi_{k,s}}$	B_k
1	4	10	10
2	8	10	$10+10 = 20$
3	8	5	$10+10+5 = 25$
4	12	10	$10+10+5+10 = 35$
5	15	20	$10+10+5+10+20 = 55$

$B_3 = 25 < 0.5\hat{N} = 55/2$ *a* $B_4 = 35 > 55/2$ *z čoho dostávame:* $\hat{r}_{y,0.5} = Y_{(4)} = 12$.

Príklad 3.2. *Jednoduchý náhodný výber bez návratu, rovnaké pravdepodobnosti zahrnutia.*

Výber rozsahu $n = 5$ *zo súboru rozsahu* $N = 50$ *zahŕňa* $s = \{3, 20, 11, 4, 31\}$*:*

k	y_k	π_k	$\frac{1}{\pi_k}$
3	2	0.10	10
20	28	0.10	10
11	16	0.10	10
4	5	0.10	10
31	46	0.10	10
$\sum$			50

Podľa vzťahu (12) potom vieme určiť:

k	Y_k	$F_n(Y_k)$
1	2	$10/50$
2	28	$(10+10+10+10)/50 = 40/50$
3	16	$(10+10+10)/50 = 30/50$
4	5	$(10+10)/50 = 20/50$
5	46	$(10+10+10+10+10)/50 = 1$

$$\hat{r}_{y,0.5} = F_n^{-1}(0.5) = inf\{Y_k : F_n(Y_k) \geq 0.5\} = 16$$

Alternatívne:

k	$Y_{(k)}$	$\frac{1}{\pi_{k,s}}$	B_k
1	2	10	10
2	5	10	$10+10 = 20$
3	16	10	$10+10+10 = 30$
4	28	10	$10+10+10+10 = 40$
5	46	10	$10+10+10+10+10 = 50$

$B_2 = 20 < 0.5\hat{N} = 50/2$ *a* $B_3 = 30 > 50/2$ *z čoho dostávame:* $\hat{r}_{y,0.5} = Y_{(3)} = 16$.

3.2 P kvantil pri oblastnom jednoduchom náhodnom výbere bez návratu

Uvažujme súbor N objektov rozdelených do K oblastí rozsahu $N_1, N_2, \dots N_K$. Náhodným výberom bez vrátenia získame n objektov, rozsahy výberov z jednotlivých oblastí označme $n_1, n_2, \dots n_K$. Pravdepodobnosti zahrnutia jednotlivých objektov základného súboru potom závisia od oblasti do ktorej sú zahrnuté. Pravdepodobnosť zahrnutia objektu $y_k, k \in \mathbf{U}$, z oblasti $i, i \in \{1, 2, \dots, K\}$, vieme napísať ako:

$$\pi_k = \frac{n_i}{N_i}.$$

Odhad p kvantilu, $\hat{r}_{y,p}$, potom získame rovnakým postupom, priradením pravdepodobností zahrnutia, ako sme uviedli v časti 3.1, pričom vychádzame z usporiadanej postupnosti výberového súboru zjednoteného z výberových súborov každej oblasti.

Ďalej priraďme každej oblasti váhu v_i:

$$v_i = \frac{N_i}{N}. \tag{18}$$

Potom distribučná funkcia súboru rozdeleného na K oblastí má tvar:

$$F_N(y) = \sum_{i=1}^{K} v_i F_{N_i}(y), \tag{19}$$

a jej odhad:

$$F_n(y) = \sum_{i=1}^{K} v_i F_{n_i}(y), \tag{20}$$

kde $F_{N_i}(y)$ je distribučná funkcia pozorovaného znaku v oblasti i, $F_{n_i}(y)$ je jej odhad a pre každú oblasť i sú definované rovnako ako pri náhodnom výbere v časti 3.1, pričom $N = N_i$, $n = n_i$. $F_{N_i}(y)$ teda predstavuje podiel prvkov oblasti i , pre ktoré platí, že sú menšie alebo rovné ako y a $F_{n_i}(y)$ podiel prvkov vo výbere z oblasti i, ktoré sú menšie alebo rovné ako y.

Intuitívne, ak sa hodnota n_i približuje k N_i pre $i = \{1, \dots K\}$, tak sa odhad $\hat{r}_{y,p}$ blíži k hodnote p kvantilu základného súboru $r_{y,p}$.

4 Interval spoľahlivosti pre p kvantil pri rôznych výberových schémach

4.1 Prístup cez empirickú distribučnú funkciu

Podľa Woodruffa v [7], [8] je zavedený nasledujúci postup pre získanie intervalu spoľahlivosti pre medián. Postup pre medián zovšeobecníme na ľubovoľný prvok základného súboru $r_{y,p}$, ktorý určuje p kvantil základného súboru (14). Ak pre nejaké dve konštanty c_1, c_2 platí:

$$Pr\{c_1 \leq F_n(r_{y,p}) \leq c_2\} = Pr\{F_n^{-1}(c_1) \leq r_{y,p} \leq F_n^{-1}(c_2)\} = 1 - \alpha,$$

tak

$$\left[F_n^{-1}(c_1), F_n^{-1}(c_2)\right] \tag{21}$$

určuje $(1-\alpha)100\%$-ný interval spoľahlivosti pre $r_{y,p}$.

Neskôr ukážeme, že $F_n(r_{y,p})$ má aproximačne normálne rozdelenie so strednou hodnotou $F_N(r_{y,p}) = p$ a disperziou $V[F_n(r_{y,p})]$. Preto pre výpočet $(1-\alpha)100\%$ -ného intervalu spoľahlivosti pre p kvantil použijeme vzťah:

$$Pr\left(\frac{c_1 - p}{V[F_n(r_{y,p})]^{1/2}} \leq \frac{F_n(r_{y,p}) - p}{V[F_n(r_{y,p})]^{1/2}} \leq \frac{c_2 - p}{V[F_n(r_{y,p})]^{1/2}}\right) = 1 - \alpha.$$

Z toho vidieť, že treba zvoliť:

$$c_1 = p + u_{\frac{\alpha}{2}}(V[F_n(r_{y,p})])^{1/2}, \tag{22}$$

$$c_2 = p - u_{\frac{\alpha}{2}}(V[F_n(r_{y,p})])^{1/2}. \tag{23}$$

Neznámou ostáva hodnota variancie distribučnej funkcie $V[F_n(r_{y,p})]$. Hodnotu variancie distribučnej funkcie základného súboru vieme napísať ako podiel:

$$V[F_N(r_{y,p})] = V\left[\frac{\sum_{k\epsilon\mathbf{U}} Z_{k,r_{y,p}}}{N}\right] = V\left[\frac{\sum_{k\epsilon\mathbf{U}} Z_{k,r_{y,p}}}{\sum_{k\epsilon\mathbf{U}} 1}\right].$$

Pre ďalší postup uvedieme metódu používanú na odhad variancie nelineárnej funkcie dvoch náhodných veličín. V našom prípade ide o funkciu podielu dvoch úhrnov t_1, t_2, ktoré budeme odhadovať. Označme túto nelineárnu funkciu charakteristík základného súboru t_1, t_2 parameterom $\rho = f(t_1, t_2) = \frac{t_1}{t_2}$, pričom :

$$t_1 = \sum_{k\epsilon\mathbf{U}} Z_{k,r_{y,p}},$$

$$t_2 = \sum_{k \epsilon \mathbf{U}} 1.$$

Potom odhad parametra $\hat{\rho}$ predstavuje podiel: $\hat{\rho} = f(\hat{t}_1, \hat{t}_2) = \frac{\hat{t}_1}{\hat{t}_2}$, kde :

$$\hat{t}_1 = \sum_{k \epsilon s} \frac{Z_{k,r_{y,p}}}{\pi_k},$$

$$\hat{t}_2 = \sum_{k \epsilon s} \frac{1}{\pi_k}.$$

Z toho hľadaná variancia $V[F_n(r_{y,p})] = V[\hat{\rho}]$.

Odhad $\hat{\rho}$ aproximujeme lineárnym odhadom $\hat{\rho}_0$, ktorý získame z Taylorovho rozvoja prvého rádu:

$$\begin{aligned}
\hat{\rho} \doteq \hat{\rho}_0 &= \rho + \frac{d\hat{\rho}}{d\hat{t}_1}|_{(t_1,t_2)}(\hat{t}_1 - t_1) + \frac{d\hat{\rho}}{d\hat{t}_2}|_{(t_1,t_2)}(\hat{t}_2 - t_2) \\
&= \rho + \frac{1}{t_2}(\hat{t}_1 - t_1) - \frac{t_1}{t_2^2}(\hat{t}_2 - t_2) \\
&= \rho + \frac{1}{t_2}\left(\hat{t}_1 - \frac{t_1}{t_2}\hat{t}_2\right) \\
&= \rho + \frac{1}{t_2}\sum_{k \epsilon s} \frac{(Z_{k,r_{y,p}} - \rho)}{\pi_k}.
\end{aligned}$$

Potom variancia $\hat{\rho}_0$, aproximácia variancie $\hat{\rho}$, označme $AV[\hat{\rho}]$, má tvar:

$$\begin{aligned}
AV[\hat{\rho}] &= \frac{1}{t_2^2} AV\left[\sum_{k \epsilon s} \frac{(Z_{k,r_{y,p}} - \rho)}{\pi_k}\right] = \frac{1}{t_2^2} AV\left[\sum_{k \epsilon \mathbf{U}} \frac{(Z_{k,r_{y,p}} - \rho)}{\pi_k} I_k(S)\right] \qquad (24) \\
&= \frac{1}{t_2^2}\sum_{k \epsilon \mathbf{U}} \frac{(Z_{k,r_{y,p}} - \rho)^2}{\pi_k^2} V[I_k(S)] \\
&+ \frac{1}{t_2^2}\sum_{k \epsilon \mathbf{U}} \sum_{l \epsilon \mathbf{U}, k \neq l} C[I_k(S), I_l(S)] \frac{(Z_{k,r_{y,p}} - \rho)}{\pi_k} \frac{(Z_{l,r_{y,p}} - \rho)}{\pi_l} \\
&= \frac{1}{t_2^2}\sum_{k \epsilon \mathbf{U}} \sum_{l \epsilon \mathbf{U}} \Delta_{kl} \frac{(Z_{k,r_y} - \rho)}{\pi_k} \frac{(Z_{l,r_{y,p}} - \rho)}{\pi_l}.
\end{aligned}$$

$r_{y,p}$ predstavuje p kvantil základného súboru, preto:

$$\rho = \frac{t_1}{t_2} = \frac{\sum_{k \epsilon \mathbf{U}} Z_{k,r_{y,p}}}{N} = p.$$

Pre odhad variancie $\hat{\rho}$ v mediáne potom platí:

$$\hat{V}[F_n(r_{y,p})] = \frac{1}{\hat{t}_2^2}\sum_{k \epsilon s} \sum_{l \epsilon s} \check{\Delta}_{kl} \frac{(Z_{k,r_{y,p}} - p)}{\pi_k} \frac{(Z_{l,r_{y,p}} - p)}{\pi_l}. \qquad (25)$$

Navyše hodnotu $r_{y,p}$ nepoznáme, preto s premennou $Z_{k,r_{y,p}}$ nevieme pracovať a použijeme odhad objektu $\hat{r}_{y,p}$ a $Z_{k,\hat{r}_{y,p}}$. Dostávame tak vzťah:

$$\hat{V}[F_n(\hat{r}_{y,p})] = \frac{1}{\hat{t}_2^2} \sum_{k\epsilon s} \sum_{l\epsilon s} \check{\Delta}_{kl} \frac{(Z_{k,\hat{r}_{y,p}} - p)}{\pi_k} \frac{(Z_{l,\hat{r}_{y,p}} - p)}{\pi_l}. \tag{26}$$

Následne už vieme určiť aproximačne hodnoty na výpočet hraníc intervalu spoľahlivosti pre medián dané vzťahmi (22), (23).

Pozrieme sa na disperzie (25) a (26) pri rôznych výberových schémach.

4.1.1 Bernoulliho výber s parametrom π

Podľa vzťahov (4), (5) platí:

$$\Delta_{kl} = \begin{cases} 0 & k \neq l, \\ \pi(1-\pi) & k = l. \end{cases} \tag{27}$$

Preto pre disperziu (24) pri Bernouliho výbere s náhodným rozsahom výberu n_s platí:

$$V[F_n(r_{y,p})] = \frac{1}{N^2} \frac{(1-\pi)}{\pi} \sum_{k\in\mathbf{U}} (Z_{k,r_{y,p}} - p)^2, \tag{28}$$

a na jej odhad použijeme podľa (26):

$$\hat{V}[F_n(\hat{r}_{y,p})] = \frac{(1-\pi)}{n_s^2} \sum_{k\in s} (Z_{k,\hat{r}_{y,p}} - p)^2. \tag{29}$$

4.1.2 Systematický, modifikovaný, výber s krokom a

Podľa vzťahov (6), (7) platí:

$$\Delta_{kl} = \begin{cases} \frac{1}{a} - \frac{1}{a^2} & |k-l| = amb, b = \{1, 2, ...\}, \\ \frac{1}{a^2} \frac{(1-a)}{(ma-1)} & inak. \end{cases} \tag{30}$$

Preto pre disperziu (24) pri systematickom výbere platí:

$$\begin{aligned} V[F_n(r_{y,p})] = \frac{1}{N^2} \sum_{k\in\mathbf{U}} \Big[& \sum_{l\in\mathbf{U}, |k-l| = amb} (a-1)(Z_{k,r_{y,p}} - p)(Z_{l,r_{y,p}} - p) + \\ & \sum_{l\in\mathbf{U}, |k-l| \neq amb} \frac{(1-a)}{(ma-1)} (Z_{k,r_{y,p}} - p)(Z_{l,r_{y,p}} - p) \Big], \end{aligned} \tag{31}$$

a pre jej odhad podľa (26) :

$$\hat{V}[F_n(\hat{r}_{y,p})] = \frac{1}{(na)^2}\sum_{k\in s}[\sum_{l\in s,|k-l|=amb} a(a-1)(Z_{k,\hat{r}_{y,p}}-p)(Z_{l,\hat{r}_{y,p}}-p)+ \sum_{l\in s,|k-l|\neq amb} a\frac{(1-a)}{(m-1)}(Z_{k,\hat{r}_{y,p}}-p)(Z_{l,\hat{r}_{y,p}}-p)] \tag{32}$$

V skutočnosti, systematický výber predstavuje na rozdiel od Bernoulliho výberu, typ výberu s pevne daným, nenáhodným, rozsahom výberu n. Preto by sme na odvodenie variancie distribučnej funkcie mohli použiť jednoduchší postup:

$$\begin{aligned} V[F_n(r_{y,p})] &= V\left[\frac{\sum_{k\in s}\frac{Z_{k,r_{y,p}}}{\pi_k}}{\sum_{k\in s}\frac{1}{\pi_k}}\right] = V\left[\frac{\sum_{k\in s}\frac{Z_{k,r_{y,p}}}{\pi_k}}{an}\right] \\ &= \frac{1}{(an)^2}V\left[\sum_{k\in s}\frac{Z_{k,r_{y,p}}}{\pi_k}\right] = \frac{1}{(an)^2}V\left[\sum_{k\in \mathbf{U}}\frac{Z_{k,r_{y,p}}}{\pi_k}I_k\right] \\ &= \frac{1}{(an)^2}\sum_{k\in\mathbf{U}}\frac{Z^2_{k,r_{y,p}}}{\pi_k^2}V[I_k] + \frac{1}{N^2}\sum_{k\in\mathbf{U}}\sum_{l\in\mathbf{U},k\neq l}C[I_k,I_l]\frac{Z_{k,r_{y,p}}}{\pi_k}\frac{Z_{l,r_{y,p}}}{\pi_l} \\ &= \frac{1}{(an)^2}\sum_{k\in\mathbf{U}}\frac{1-\pi_k}{\pi_k}Z_{k,r_{y,p}} + \frac{1}{N^2}\sum_{k\in\mathbf{U}}\sum_{l\in\mathbf{U},k\neq l}\Delta_{kl}\frac{Z_{k,r_{y,p}}}{\pi_k}\frac{Z_{l,r_{y,p}}}{\pi_l} \\ &= \frac{1}{(an)^2}\sum_{k\in\mathbf{U}}\sum_{l\in\mathbf{U}}\Delta_{kl}\frac{Z_{k,r_y}}{\pi_k}\frac{Z_{l,r_{y,p}}}{\pi_l}. \end{aligned} \tag{33}$$

Z toho potom pre varianciu výberovej distribučnej fukcie pri systematickom výbere dostávame vzťah:

$$V[F_n(r_{y,p})] = \frac{1}{(an)^2}\sum_{k\in\mathbf{U}}[\sum_{l\in\mathbf{U},|k-l|=amb}(a-1)Z_{k,r_{y,p}}Z_{l,r_{y,p}}+ \sum_{l\in\mathbf{U},|k-l|\neq amb}\frac{(1-a)}{(ma-1)}Z_{k,r_{y,p}}Z_{l,r_{y,p}}], \tag{34}$$

a pre jej odhad:

$$\hat{V}[F_n(\hat{r}_{y,p})] = \frac{1}{(an)^2}\sum_{k\in s}[\sum_{l\in s,|k-l|=amb} a(a-1)Z_{k,\hat{r}_{y,p}}Z_{l,\hat{r}_{y,p}}+ \sum_{l\in s,|k-l|\neq amb} a\frac{(1-a)}{(m-1)}Z_{k,\hat{r}_{y,p}}Z_{l,\hat{r}_{y,p}}]. \tag{35}$$

4.1.3 Jednoduchý náhodný výber bez návratu

Pri jednoduchom náhodnom výbere bez vrátenia vieme $V[F_n(r_y)]$ zjednodušiť. Z vlastností pravdepodobností zahrnutia pri jednoduchom náhodnom výbere (8) dostávame

zo vzťahu pre odhad distibučnej funkcie v bode $r_{y,p}$, podľa (12):

$$nF_n(r_{y,p}) = \sum_{k \in s} Z_{k,r_{y,p}}.$$

Z toho $nF_n(r_{y,p})$ vyjadruje počet prvkov vo výbere, ktoré sú menšie alebo rovné ako prvok $r_{y,p}$. Náhodná veličina $nF_n(r_{y,p})$ má preto hypergeometrické rozdelenie so strednou hodnotou a varianciou:

$$E[F_n(r_{y,p})] = F_N(r_{y,p}) = p,$$

$$V[F_n(r_{y,p})] = \frac{N-n}{N-1}\frac{1}{n}F_N(r_{y,p})\{1-F_N(r_{y,p})\} = \frac{1-\frac{n}{N}}{N-1}\frac{N}{n}p(1-p) \doteq \frac{1-\frac{n}{N}}{n}p(1-p). \tag{36}$$

4.1.4 Oblastný jednoduchý náhodný výber bez návratu

Vrátime sa teraz k oblastnému výberu zavedenému v časti (3.2).

Z hypergeometrického rozdelenia náhodnej premennej $n_i F_{n_i}(r_{y,p})$ pre $i = \{1, \ldots, K\}$ a z ich nezávislosti:

$$E[F_{n_i}(r_{y,p})] = F_{N_i}(r_{y,p}),$$

$$V[F_{n_i}(r_{y,p})] = \frac{N_i - n_i}{N_i - 1}\frac{1}{n_i}F_{N_i}(r_{y,p})\{1-F_{N_i}(r_{y,p})\} \doteq \frac{1-\frac{n_i}{N_i}}{n_i}F_{N_i}(r_{y,p})[1-F_{N_i}(r_{y,p})]. \tag{37}$$

Z toho pre celkový súbor:

$$E[F_n(r_{y,p})] = \sum_{i=1}^{K} v_i F_{N_i}(r_{y,p}) = p,$$

$$V[F_n(r_{y,p})] = \sum_{i=1}^{K} v_i^2 \frac{1-\frac{n_i}{N_i}}{n_i}F_{N_i}(r_{y,p})\{1-F_{N_i}(r_{y,p})\} \tag{38}$$

a pre odhad variacie distribučnej funkcie v odhade pre medián:

$$\hat{V}[F_n(\hat{r}_{y,p})] = \sum_{i=1}^{K} v_i^2 \frac{1-\frac{n_i}{N_i}}{n_i}F_{n_i}(\hat{r}_{y,p})\{1-F_{n_i}(\hat{r}_{y,p})\}. \tag{39}$$

4.1.5 Výpočet a simulácie variancií distribučnej funkcie pri rôznych výberových schémach

Získané vzťahy sú približné (vzťahy pre varianciu odvodené z aproximácie nelineárnej funkcie náhodných veličín lineárnou), preto cieľom je porovnať odhady variancií

empirickej distribučnej funkcie v mediáne základného súboru a v mediáne výberového súboru získané výpočtom, s hodnotami získanými simuláciami pri zvolených výberových schémach.

Pri výpočtoch variancií distribučnej funkcie a jej odhadoch budeme vychádzať z už nami definovaných výberových schém. Na začiatku generujeme N dát základného súboru z exponenciálneho rozdelenia s parametrom $\lambda = 0.2$ a určíme hodnotu mediánu základného súboru, $r_{y,0.5}$.

Zvoľme náhodné rozdelenie prvkov základného súboru pomocou diskrétnej náhodnej veličiny O do K oblastí rozsahu $N_1, N_2, \ldots, N_K$:

$$O = \begin{cases} 1 & \text{s pravdepodobnosťou } p_1, \\ 2 & \text{s } p_2, \\ \vdots & \\ K & \text{s } p_K, \end{cases} \tag{40}$$

kde $\sum_{i=1}^{K} p_i = 1$.

Z dát základného súboru určíme výpočtom, z už odvodených vzťahov v tejto kapitole, hodnoty variancií distribučnej funkcie v mediáne základného súboru, $r_{y,0.5}$,pre jednotlivé výberové schémy, hodnota $V[F_n(r_{y,0.5})]$.

Z vygenerovaných dát realizujeme výbery podľa výberovej schémy. Pri Bernoulliho výbere vychádzame z celkového základného súboru pričom pravdepodobnosť zahrnutia určíme tak, aby stredná hodnota rozsahu výberového súboru bola rovnaká ako pri ostatných výberových schémach, kvôli možnosti porovnania získaných výsledkov, t.j. $\pi = \frac{n}{N}$.

Pri systematickom a jednoduchom náhodnom výbere bez návratu vyberáme n prvkov zo základného súboru rozsahu N, pri systematickom nastavením správnej dĺžky kroku, $a = \frac{N}{n}, a \in Z$.

Pri oblastnom proporcionálnom výbere (Oblastný p) realizujeme jednoduchý náhodný výber bez návratu zo všetkých oblastí, do ktorých sme rozdelili prvky základného súboru pomocou diskrétnej náhodnej veličiny. Rozsah výberu z jednotlivých oblastí je $n_1, n_2, \ldots, n_K$, pričom platí:

$$\frac{n_1}{N_1} = \frac{n_2}{N_2} = \cdots = \frac{n_K}{N_K} = \frac{n}{N}. \tag{41}$$

Pri neproporcionálnom oblastnom výbere (Oblastný np) s pevne daným rozsahom výberového súboru n, určíme ľubovoľné rozsahy výberov z jednotlivých oblastí $n_1, n_2, \ldots, n_K$, aby platilo $\sum_{i=1}^{K} n_i = n$.

Ďalej zo získaných výberov určíme hodnotu výberovej distribučnej funkcie v mediáne základného súboru, $F_n(r_{y,0.5})$, a v jeho odhade, $F_n(\hat{r}_{y,0.5})$. Odhad mediánu je z jedného zrealizovaného výberu každou výberovou schémou, pri jednooblastnom výbere pomocou postupnosti váh z výberového súboru ako uvádzame v časti 3.1, pri oblastnom výbere z výberového súboru pomocou postupnosti váh podľa 3.2. Z opakovaných simulácií ďalej určíme odhady variancií, označme $\hat{V}_s[F_n(r_{y,0.5})]$ a $\hat{V}_s[F_n(\hat{r}_{y,0.5})]$.

Výsledky získané simuláciou porovnáme s odhadom disperzií výpočtom, určené ako priemer z opakovaní, podľa vzťahov odvodených v predchádzajúcej časti, pričom opäť používame hodnotu mediánu základného súboru, $\hat{V}[F_n(r_{y,0.5})]$, a jeho odhad, $\hat{V}[F_n(\hat{r}_{y,0.5})]$.

Výpočty a simulácie variancií (získané z opakovania 10000 krát) pre dáta z exponenciálneho rozdelenia s parametrom $\lambda = 0.2$(zadelené do oblastí s pravdepodobnosťami $p_i = 1/K$ pre $\forall i$), pri oblastnom výbere s dvoma oblasťami, pre rôzne prípady rozsahov základného súboru a rozsahov výberu, zobrazujú Tabuľky 1,2,3,4. Pri neproporcionálnom oblastnom výbere sme pri $n = 0.1N$ použili rozsah výberu s prvej oblasti $n_1 = 0.05N_1$ a pri rozsahu $n = 0.05N$ rozsah $n_1 = 0.01N_1$. Pri systematickom výbere sme získali výpočtom rovnaké výsledky variancií oboma odvodenými vzťahmi v časti 4.1. Tabuľka 5 predstavuje 95%-ný interval spoľahlivosti pre varianciu v odhade pre medián získanú simuláciami.

Z porovnania variancií získaných simuláciami a výpočtom vidieť, že kvalita výpočtov závisí od rozsahu základného a výberového súboru ale výsledky môžu byť ovplyvnené aj tvorbou základného súboru.

$N = 1000$	výber n=0.1N				
variancia	Bernouliho	Systematický	Jednoduchý	Oblastný p	Oblastný np
$\hat{V}_s[F_n(r_{y,0.5})]$	0.002303815	0.003122789	0.002276133	0.002248896	0.003333930
$\hat{V}_s[F_n(\hat{r}_{y,0.5})]$	0.002169312	0.003355094	0.002181522	0.002196792	0.003321399
$V[F_n(r_{y,0.5})]$	0.002250000	0.003195918	0.002252252	0.002249423	0.003323748
$\hat{V}[F_n(r_{y,0.5})]$	0.002269508	0.003185595	0.002252252	0.002208346	0.003209008
$\hat{V}[F_n(\hat{r}_{y,0.5})]$	0.002269508	0.003404758	0.002252252	0.002170244	0.003207874

Tabuľka 1

$N = 100$	výber n=0.1N				
variancia	Bernouliho	Systematický	Jednoduchý	Oblastný p	Oblastný np
$\hat{V}_s[F_n(r_{y,0.5})]$	0.02605165	0.02548449	0.02253325	0.02293128	0.02922419
$\hat{V}_s[F_n(\hat{r}_{y,0.5})]$	0.02605165	0.02397637	0.02198509	0.02271550	0.02654872
$V[F_n(r_{y,0.5})]$	0.02250000	0.02571429	0.02272727	0.02125622	0.03546756
$\hat{V}[F_n(r_{y,0.5})]$	0.02512016	0.02592045	0.02272727	0.01833169	0.02044928
$\hat{V}[F_n(\hat{r}_{y,0.5})]$	0.02512016	0.02458665	0.02272727	0.01829140	0.01866342

Tabuľka 2

$N = 1000$	výber n=0.05N				
variancia	Bernouliho	Systematický	Jednoduchý	Oblastný p	Oblastný np
$\hat{V}_s[F_n(r_{y,0.5})]$	0.004793774	0.005664842	0.004725943	0.004759187	0.014787268
$\hat{V}_s[F_n(\hat{r}_{y,0.5})]$	0.004749796	0.006078337	0.004706222	0.004496200	0.014742153
$V[F_n(r_{y,0.5})]$	0.004750000	0.005680808	0.004754755	0.006998205	0.017870257
$\hat{V}[F_n(r_{y,0.5})]$	0.004849473	0.005615773	0.004754755	0.004567982	0.011820413
$\hat{V}[F_n(\hat{r}_{y,0.5})]$	0.004849473	0.006078556	0.004754755	0.004364592	0.011802284

Tabuľka 3

$N = 100$	výber n=0.05N				
variancia	Bernouliho	Systematický	Jednoduchý	Oblastný p	Oblastný np
$\hat{V}_s[F_n(r_{y,0.5})]$	0.06278202	0.04738782	0.04773070	0.04935461	0.07917928
$\hat{V}_s[F_n(\hat{r}_{y,0.5})]$	0.06288095	0.03568716	0.02001061	0.04294837	0.05253348
$V[F_n(r_{y,0.5})]$	0.04750000	0.04797980	0.04797980	0.05575221	0.09724926
$\hat{V}[F_n(r_{y,0.5})]$	0.06211286	0.04812130	0.04797980	0.02857774	0.01015443
$\hat{V}[F_n(\hat{r}_{y,0.5})]$	0.06211286	0.03598030	0.04797980	0.02496538	0.00767184

Tabuľka 4

$N = 100$	výber n=0.05N
výberová schéma	$\hat{V}_s[F_n(r_{y,0.5})]$
Bernouliho	$(0.06123821, 0.06464886)$
Systematický	$(0.04677165, 0.04901598)$
Jednoduchý	$(0.04671190, 0.04920305)$
Oblastný p	$(0.04695963, 0.05103527)$
Oblastný np	$(0.05896502, 0.10162933)$

Tabuľka 5: 95% IS

4.2 Prístup pri veľkom rozsahu súboru pomocou hustoty rozdelenia, asymptotické rozdelenie mediánu

Iný postup ako získať IS pre medián je odvodiť asymptotické rozdelenie výberových kvantilov použitím aproximácie hustoty. Pre túto úlohu existuje viacero postupov, v tejto kapitole opíšeme jeden z nich podľa [5], využitím viacero tvrdení.

4.2.1 Jednoduchý náhodný výber bez návratu

Nech $\{\mathbf{U}_N, N = 1, 2, \dots\}$ tvorí postupnosť konečných základných súborov. Každé $\mathbf{U}_N$ je charakterizované postupnosťou prvkov základného súboru, $y_{1,N}, y_{2,N}, \dots, y_{N,N}$, s príslušnou distribučnou funkciou $F_N(y)$. Pre každé N uvažujeme výberový súbor s prvkami $\{Y_{1,N}, Y_{2,N}, \dots, Y_{n,N}\}$ rozsahu n s empirickou distribučnou funkciou $F_n(y)$. Označme p kvantil základného súboru:

$$r_{y,p,N} = F_N^{-1}(p) = inf\{y \in \mathbf{U}_N : F_N(y) \geq p\} \tag{42}$$

a p kvantil výberového súboru:

$$\hat{r}_{y,p,N} = F_n^{-1}(p) = inf\{y \in \mathbf{U}_N : F_n(y) \geq p\}. \tag{43}$$

Predpoklady:

(P1) Postupnosť $\{r_{y,p,N}\}$ je ohraničená.

(P2) Existuje postupnosť funkcií$\{f_N\}$ taká, že:

$$\lim_{N\to\infty}\left[\frac{F_N(r_{y,p,N} + \delta_N) - p}{\delta_N} - f_N(r_{y,p,N})\right] = 0,$$

ak

$$\{\delta_N\} \sim O(n^{-1/2}),$$

kde $\{\delta_N\} \sim O(n^{-1/2})$ znamená, že $\exists\varepsilon > 0$: pre $\forall n > 0 : n^{-1/2} - \varepsilon < \{\delta_N\} < n^{-1/2} + \varepsilon$.

Uvedené predpoklady sa používajú vo viacerých prácach. Lema 4.3 aj s dôkazom je z článku [5], pre úplnosť uvádzame aj známu Lemu 4.1 používanú pri jej dôkaze.

Lema 4.1. *(Centrálne limitná veta pre výbery z konečného súboru [6]).*
Označme pre prvky množiny $\boldsymbol{U}_N$*:*

$$Z_{k,y,N} = \begin{cases} 1 & y_{k,N} \leq y \\ 0 & y_{k,N} > y \end{cases}$$

Ďalej označme:

$$E_N = \sum_{i \in \boldsymbol{U}} Z_{i,y,N} = NF_N(y),$$

$$D_N^2 = \sum_{i \in \boldsymbol{U}} (Z_{i,y,N} - F_N(y))^2 = NF_N(y)(1 - F_N(y)),$$

Pri pevnom N *zrealizujme jednoduchým náhodným výberom bez vrátenia výber* s*, obsahujúci* $n = n(N) < N$ *prvkov. Ďalej označme náhodnú pemennú:*

$$Y_{N,n} = \sum_{i \in s} Z_{i,y,N} = nF_n(y),$$

ktorá má hypergeometrické rozdelenie s disperziou:

$$D_{N,n} = \sqrt{NF_N(y)(1 - F_N(y)) \frac{n}{N} \frac{(N-n)}{(N-1)}} \doteq D_N \sqrt{\frac{n}{N}\left(1 - \frac{n}{N}\right)},$$

a

$$Y_{N,n}^* = Y_{N,n} - nF_N(y) = nF_n(y) - nF_N(y).$$

Položme:

$$d_{N,n}(\varepsilon) = \frac{1}{D_N^2} \sum_{|Z_{i,y,N} - F_N(y)| > D_{N,n}\varepsilon} (Z_{i,y,N} - F_N(y))^2.$$

Potom ak pre každé $\varepsilon > 0$ *:*

$$\lim_{N \to \infty} d_{N,n}(\varepsilon) = 0, \tag{44}$$

tak:

$$\lim_{N \to \infty} P\left(\frac{Y_{N,n}^*}{D_{N,n}} < y\right) = \Phi(y). \tag{45}$$

Lema 4.2. *(Sluckého lema. [1])*

Nech $\{X_N\}$, $\{Y_N\}$ *predstavujú postupnosti náhodných veličín, kde:* $\{X_N\} \xrightarrow{d} X$, [2] $\{Y_N\} \xrightarrow{p} c$. [3] *Potom* $\{X_N + Y_N\} \xrightarrow{d} X + c$.

Lema 4.3. *Za platnosti predpokladov* $(P1)$ *a* $(P2)$ *platí:*

$$P\left(\frac{\hat{r}_{y,p,N} - r_{y,p,N}}{a_{N,r_{y,p,N}}} \leq y\right) \xrightarrow{d} \Phi(y),$$

kde :

$\Phi(y)$ *je distribučná funkcia* $N(0,1)$ *v bode* y,

$n, N - n \to \infty$,

$0 \leq \frac{n}{N} < \delta < 1$,

$a_{N,r_{y,p,N}} = \sqrt{\left(\frac{(N-n)p(1-p)}{(N-1)nf_N^2(r_{y,p,N})}\right)}$.

Dôkaz. *Označme pre jednoduchosť* $a_{N,r_{y,p,N}} = a$, $r_{y,p,N} = r_{y,p}$ *a upravme pravdepodobnosť:*

$$\begin{aligned} P\left(\frac{\hat{r}_{y,p} - r_{y,p}}{a} \leq y\right) &= P\left(\hat{r}_{y,p} \leq r_{y,p} + ya\right) \\ &= P\left(p \leq F_n(r_{y,p} + ya)\right) \\ &= P\left(\frac{p - F_N(r_{y,p} + ya)}{af_N(r_{y,p})} \leq \frac{F_n(r_{y,p} + ya) - F_N(r_{y,p} + ya)}{af_N(r_{y,p})}\right) \end{aligned}$$

Označme teraz :

$$b_N = \frac{p - F_N(r_{y,p} + ya)}{af_N(r_{y,p})}, \tag{46}$$

$$W_N = \frac{F_n(r_{y,p} + ya) - F_N(r_{y,p} + ya)}{af_N(r_{y,p})}.$$

Výraz (46) *môžme vyjadriť aj:* $b_N = -\frac{y(F_N(r_{y,p}+ya)-p)}{af_N(r_{y,p})y}$ *a podľa predpokladu* $(P2)$: $\lim_{n,N-n\to\infty} b_N = -y$.

Preto ak dokážeme, že $W_N \xrightarrow{d} W$ *pre* $N \to \infty$, *kde* $W \sim N(0,1)$, *tak*

$$\lim_{N\to\infty} P(b_N \leq W_N) = 1 - \Phi(-y) = \Phi(y),$$

[2] Postupnosť náhodných veličín $X_1, X_2, \ldots$ konverguje k náhodnej veličine X podľa distribučnej funkcie, ozn. $\{X_N\} \xrightarrow{d} X$ ak $lim_{N\to\infty} F_{X_N}(x) = F_X(x)$ v každom bode spojitosti funkcie $F_X(x)$.

[3] Postupnosť náhodných veličín $Y_1, Y_2, \ldots$ konverguje k náhodnej veličine Y podľa pravdepodobnosti, ozn. $\{Y_N\} \xrightarrow{p} Y$ ak pre $\forall\varepsilon$: $lim_{N\to\infty} P[|Y_N - Y| \geq \varepsilon] = 0$.

čo je požadované tvrdenie.

Rozšírme $W_N = W_N^* + (W_N - W_N^*)$, *kde* $W_N^* = \frac{F_n(r_{y,p}) - F_N(r_{y,p})}{a f_N(r_{y,p})}$ *a ďalej:*

1. *Dokážeme, že* $W_N^* \xrightarrow{d} W$ *pomocou zovšeobecnenej Centrálnej limitnej vety pre výbery z konečného súboru.*

2. *Dokážeme:* $|W_N - W_N^*| \xrightarrow{p} 0$ *využitím Čebyševovej nerovnosti.*

3. *Aplikujeme Sluckého lemu na získané výsledky.*

Prvý krok:

Platí, že $nF_n(r_{y,p})$ *má hypergeometrické rozdelenie, preto pre* y *také, že:* $max(0, n - N(1 - F_N(r_{y,p}))) \leq y \leq min(n, NF_N(r_{y,p}))$ *platí:*

$$P(nF_n(r_{y,p}) = y) = \frac{\binom{NF_N(r_{y,p})}{y}\binom{N(1-F_N(r_{y,p}))}{n-y}}{\binom{N}{n}}.$$

Aplikáciou Centrálnej limitnej vety pre výbery z konečného súboru, ktorú uvádzame v špeciálnom tvare pre náš prípad v Leme 4.1*, chceme dokázať normálne rozdelenie* W_N^*. *Ukážeme platnosť predpokladu* (44)*:*

$$\begin{aligned} \lim_{N\to\infty} d_{N,n}(\varepsilon) &= \lim_{N\to\infty} \frac{\sum_{|Z_{i,y,N} - F_N(y)| > D_{N,n}\varepsilon} (Z_{i,y,N} - F_N(y))^2}{\sum_{\boldsymbol{U}} (Z_{i,y,N} - F_N(y))^2} \qquad (47) \\ &= \lim_{N\to\infty} \frac{\sum_{|Z_{i,y,N} - F_N(y)| > \sqrt{NF_N(y)(1-F_N(y))\frac{n}{N}\frac{(N-n)}{(N-1)}}\varepsilon} (Z_{i,y,N} - F_N(y))^2}{NF_N(y)(1 - F_N(y))} = 0. \end{aligned}$$

Podmienka (47) *je splnená ak:*

$$\lim_{N\to\infty} NF_N(y)(1 - F_N(y))\frac{n}{N}\frac{(N-n)}{(N-1)} = +\infty. \qquad (48)$$

Bez ujmy na všeobecnosti môžme pre $N \to \infty$ *predpokladať, že* $NF_N(y) \leq \frac{N}{2}$ *a* $n \leq \frac{N}{2}$.

Potom je ale rovnosť (48) *ekvivalentná s rovnosťou*

$$\lim_{N\to\infty} n = +\infty.$$

Preto ak $N \to \infty$ *tak aj* $n \to \infty$ *a z vlastnosti* $0 < \frac{n}{N} < \delta < 1$ *vyplýva platnosť podmienky* (48) *a teda platnosť predpokladu* (47).

Preto dostáveme pre W_N^**:*

$$\lim_{N\to\infty} P\left(W_N^* < y\right) = \Phi(y).$$

Druhý krok:

Platí, že:

$$|W_N - W_N^*| = \left| \frac{F_n(r_{y,p} + ya) - F_n(r_{y,p})}{af_N(r_{y,p})} - \frac{F_N(r_{y,p} + ya) - F_N(r_{y,p})}{af_N(r_{y,p})} \right|.$$

Ďalej vieme, že

$$n|F_n(r_{y,p} + ya) - F_n(r_{y,p})|,$$

pre $y = 0, 1, \ldots, min(n, N|F_N(r_{y,p} + ya) - F_N(r_{y,p})|)$, *má hypergeometrické rozdelenie s varianciou:*

$$Var(n|F_n(r_{y,p} + ya) - F_n(r_{y,p})|) = \frac{(N-n)}{(N-1)} n$$
$$\times |F_N(r_{y,p} + ya) - F_N(r_{y,p})|(1 - |F_N(r_{y,p} + ya) - F_N(r_{y,p})|). \tag{49}$$

Potom z Čebyševovej nerovnosti [1] a použitím vzťahu (49) *pre každé* $\varepsilon > 0$*:*

$$\begin{aligned} P(|W_N - W_N^*| \geq \varepsilon) &\leq \frac{1}{\varepsilon^2} Var(W_N - W_N^*) \\ &= \frac{1}{\varepsilon^2} \frac{(N-n)}{(N-1)} \frac{1}{n} \frac{1}{a^2 f_N^2(r_{y,p})} \\ &\quad \times |F_N(r_{y,p} + ya) - F_N(r_{y,p})|(1 - |F_N(r_{y,p} + ya) - F_N(r_{y,p})|). \end{aligned} \tag{50}$$

Ďalej využijeme nasledujúce limitné vlastnosti plynúce z $n, N - n \to \infty$*:*

(V1) predpoklad vety $\{\delta_N\} \sim O(n^{-1/2})$ *a* $0 < \frac{n}{N} < \delta_N < 1$,

(V2) z predpokladu $(P2)$ *:* $\frac{F_N(r_{y,p}+ya) - F_N(r_{y,p})}{a} = O(n^0)$,

(V3) $a = O(n^{-1/2})$.

Použitím $(V1), (V2), (V3)$ *pre* $n, N - n \to \infty$ *je limita pravej strany nerovnosti* (50) *rovná nule.*

Z toho plynie : $|W_N - W_N^*| \xrightarrow{p} 0$.

Tretí krok:

Aplikáciou Sluckého lemy 4.2 *na doterajšie výsledky dôkazu :*

$W_N^* \xrightarrow{d} W$,

$|W_N - W_N^*| \xrightarrow{p} 0$,

dostávame

$$\lim_{N\to\infty} P\left(\frac{\hat{r}_{y,p} - r_{y,p}}{a} \leq y\right) = \lim_{N\to\infty} P(b_N \leq W_N) = \Phi(y)$$

čím je veta dokázaná. □

Položme vo vzťahu (42), (43) $p = 0.5$. Potom aplikáciou Lemy (4.3) pre medián základného súboru $r_{y,0.5}$ a jeho odhad pomocou výberového súboru $\hat{r}_{y,0.5}$ platí:

$$P\left(\frac{\hat{r}_{y,0.5} - r_{y,0.5}}{a_{N,r_{y,0.5}}} \leq y\right) \xrightarrow{d} \Phi(y), \tag{51}$$

kde :

$$a_{N,r_{y,0.5}} = \frac{1}{f(r_{y,0.5})}\sqrt{\frac{N-n}{N-1}\frac{1}{n}0.25}, \tag{52}$$

a $f(r_{y,0.5})$ predstavuje podľa $(P2)$ hustotu jednotiek základného súboru v mediáne.

Hodnotu $\frac{1}{f(r_{y,0.5})}$ budeme odhadovať pomocou inverznej distribučnej funkcie nasledovne:

$$\widehat{\left(\frac{1}{f(r_{y,0.5})}\right)} = \frac{F_n^{-1}(0.5+h) - F_n^{-1}(0.5-h)}{2h}, \tag{53}$$

kde $h = n^{-1/2}$.

Preto pre $(1-\alpha)100\%$ -ný interval spoľahlivosti použijeme vzťah:

$$Pr\left(\hat{r}_{y,0.5} + u_{\frac{\alpha}{2}}\hat{a}_{N,r_{y,0.5}} \geq r_{y,0.5} \geq \hat{r}_{y,0.5} - u_{\frac{\alpha}{2}}\hat{a}_{N,r_{y,0.5}}\right) = 1-\alpha, \tag{54}$$

kde:

$$\hat{a}_{N,r_{y,0.5}} = \widehat{\left(\frac{1}{f(r_{y,0.5})}\right)}\sqrt{\frac{N-n}{N-1}\frac{1}{n}0.25}. \tag{55}$$

4.2.2 Oblastný jednoduchý náhodný výber bez návratu

Podľa článku [2] pre varianciu mediánu pri oblastnom výbere platí:

$$a_{N,r_{y,0.5}} = \frac{1}{f(r_{y,0.5})}\sqrt{\sum_{i=1}^{K} v_i^2 \frac{1-\frac{n_i}{N_i}}{n_i} F_{N_i}(r_{y,0.5})\{1 - F_{N_i}(r_{y,0.5})\}} \tag{56}$$

a $f(r_{y,0.5})$ predstavuje hustotu v mediáne.

Hodnotu $\frac{1}{f(r_{y,0.5})}$ budeme odhadovať pomocou inverznej distribučnej funkcie rovnako ako v (53). Odhad intervalu spoľahlivosti je daný (54), pričom

$$\hat{a}_{N,r_{y,0.5}} = \widehat{\left(\frac{1}{f(r_{y,0.5})}\right)}\sqrt{\sum_{i=1}^{K} v_i^2 \frac{1-\frac{n_i}{N_i}}{n_i} F_{n_i}(r_{y,0.5})\{1 - F_{n_i}(r_{y,0.5})\}}. \tag{57}$$

4.3 Prístup v prípade malého rozsahu súboru

V tejto časti sa zameriame na prístup k odhadu mediánu v prípade malého rozsahu základného súboru pri jednoduchom náhodnom výbere bez návratu. Pozrieme sa na ZS rozdelený na jednu a viac oblastí. Pôjde o odvodenie odhadu variancie mediánu cez hypergeometrické rozdelenie, vychádať budeme z článku [2].

4.3.1 Jednoduchý náhodný výber bez návratu

Pri jednoduchom náhodnom výbere z celkového súboru platí, že pravdepodobnosť zahrnutia ľubovoľného objektu $y_k, k \in \mathbf{U}$, do výberového súboru je rovnaká pre každý objekt vo výbere, $\pi_k = \pi$, podľa (8). Označme potom váhu každého objektu vo výberovom súbore $w = \frac{\frac{1}{\pi}}{N} = \frac{1}{n}$.

Usporiadaním objektov základného súboru dostávame postupnosť:

$$y_{(1)} \leq y_{(2)} \leq \cdots \leq y_{(l)} \leq y_{(l+1)} \leq \cdots \leq y_{(N)}$$

a objektov výberového súboru postupnosť:

$$Y_{(1)} \leq Y_{(2)} \leq \cdots \leq Y_{(l)} \leq Y_{(l+1)} \leq \cdots \leq Y_{(n)}.$$

Takto usporiadané objekty $y_{(1)}, \ldots, y_{(N)}$ a $Y_{(1)}, \ldots, Y_{(n)}$ tvoria poradové štatistiky.

Náhodnou premennou $T(l)$ označme počet objektov vo výberovom súbore, ktoré sú menšie alebo rovné ako hodnota objektu $y_{(l)}$.

Lema 4.4. *Pre základný súbor rozsahu N, s prvkami $y_{(l)}$, $l \in [1, N]$ a výberový súbor rozsahu n s prvkami $Y_{(l)}$, $l \in [1, n]$, mediánom $Y_{(med)}$ a váhami objektov $w = \frac{1}{n}$, platí:*

$$(Y_{(med)} > y_{(l)}) \Longleftrightarrow (T(l)w < 0.5) \tag{58}$$

Dôkaz.

$$\begin{aligned}(T(l)w < 0.5) &= (nF_n(y_{(l)})\frac{1}{n}) < 0.5) \\ &= (y_{(l)} < F_n^{-1}(0.5)) = (y_{(l)} < Y_{(med)})\end{aligned}$$

□

Lema 4.5. *Náhodná veličina $T(l)$ má pre pevne dané l, $l \in [1, N]$, Hypergeometrické rozdelenie s parametrami $NF_N(y_{(l)})$, $N(1 - F_N(y_{(l)}))$, n, kde $F_N(y_{(l)})$ označuje distribučnú funkciu základného súboru v bode $y_{(l)}$, $NF_N(y_{(l)})$ počet jednotiek základného súboru, ktoré nadobúdajú hodnotu menšiu alebo rovnú ako $y_{(l)}$ a n, rozsah výberu. Pre také t spĺňajúce podmienky:*

- $tw < 0.5$,
- $Max(0, n + N(F_N(y_{(l)}) - 1)) \leq t \leq Min(n, NF_N(y_{(l)}))$,

platí:

$$p_l = P(T(l)w < 0.5) = \sum_{(t)} \frac{\binom{NF_N(y_{(l)})}{t}\binom{N(1-F_N(y_{(l)}))}{n-t}}{\binom{N}{n}}, \tag{59}$$

kde $p_0 = 1$.

Dôkaz. *Z Hypergeometrického rozdelenia náhodnej premennej $T(l)$ pre každé t, vyhovujúce podmienke plynúcej z vlastností kombinačného čísla:*

$$Max(0, n + N(F_N(y_{(l)}) - 1)) \leq t \leq Min(n, NF_N(y_{(l)})),$$

platí:

$$P(T(l) = t) = \frac{\binom{NF_N(y_{(l)})}{t}\binom{N(1-F_N(y_{(l)}))}{n-t}}{\binom{N}{n}}.$$

Pridaním podmienky pre $t : wt < 0.5$ dostávame pre každé prípustné t:

$$p_l = P(wT(l) < 0.5) = \sum_{(t)} \frac{\binom{NF_N(y_{(l)})}{t}\binom{N(1-F_N(y_{(l)}))}{n-t}}{\binom{N}{n}}$$

□

Podľa Lemy4.4 a Lemy4.5 : $p_l = P(Y_{(med)} > y_{(l)})$.
Preto vieme vyjadriť rovnosť:

$$P(Y_{(med)} = y_{(l)}) = p_{l-1} - p_l.$$

Pri výbere rozsahu n objektov zo základného súboru rozsahu N potom pre r-tý moment náhodnej veličiny $Y_{(med)}$ dostávame vzťah:

$$E(Y^r_{(med)}) = \sum_{l=1}^{N} (p_{l-1} - p_l) y^r_{(l)}.$$

Nakoniec získavame vzťah pre varianciu mediánu:

$$V(Y_{(med)}) = E(Y^2_{(med)}) - E^2(Y_{(med)}) = \sum_{l=1}^{N}(p_{l-1} - p_l)y^2_{(l)} - (\sum_{l=1}^{N}(p_{l-1} - p_l)y_{(l)})^2 \quad (60)$$

a jej odhad :

$$\hat{V}(Y_{(med)}) = E(Y^2_{(med)}) - E^2(Y_{(med)}) = \sum_{l=1}^{n}(\widehat{p}_{l-1} - \widehat{p}_l)Y^2_{(l)} - (\sum_{l=1}^{n}(\widehat{p}_{l-1} - \widehat{p}_l)Y_{(l)})^2, \quad (61)$$

kde:

$$\widehat{p}_l = \sum_{(t)} \frac{\binom{NF_n(y_{(l)})}{t}\binom{N(1-F_n(y_{(l)}))}{n-t}}{\binom{N}{n}}. \quad (62)$$

Pri určení$(1-\alpha)100\%$-ného intervalu spoľahlivosti pre medián vychádzame z normálneho rozdelenia:

$$Pr\left(u_{\frac{\alpha}{2}} \leq \frac{Y_{(med)} - y_{(med)}}{\hat{V}(Y_{(med)})^{1/2}} \leq -u_{\frac{\alpha}{2}}\right) = 1 - \alpha.$$

Preto interval spoľahlivosti:

$$[Y_{(med)} + u_{\frac{\alpha}{2}}(\hat{V}(Y_{(med)}))^{1/2}, Y_{(med)} - u_{\frac{\alpha}{2}}(\hat{V}(Y_{(med)}))^{1/2}] \quad (63)$$

4.3.2 Oblastný jednoduchý náhodný výber bez návratu

Pri jednoduchom náhodnom výbere z oblasti $i, i \in \{1, 2, \dots K\}$ platí, že pravdepodobnosť zahrnutia ľubovoľného objektu $y_k, k \in \mathbf{U}$ z danej oblasti do výberového súboru π_k predstavuje podiel $\pi_k = \frac{n_i}{N_i}$. Váha w_k každého objektu vo výberovom súbore potom závisí na oblasti i do ktorej je objekt zahrnutý, $w_k = \frac{\frac{1}{\pi_k}}{N} = \frac{N_i}{n_i N}$.

Náhodnou premennou $T_i(l)$ označme počet objektov vo výbere z i-tej oblasti, ktoré sú menšie alebo rovné ako hodnota objektu $y_{(l)}$ základného súboru s K oblasťami.

Lema 4.6. *Pre základný súbor rozsahu N, s prvkami $y_{(l)}$, $l \in [1, N]$, rozdelený do K oblastí rozsahu $N_1, N_2, \dots, N_K$ a výberový súbor rozsahu n, s prvkami $Y_{(l)}$, $l \in [1, n]$ rozdelený do K oblastí s rozsahom $n_1, n_2, \dots, n_K$, mediánom $Y_{(med)}$ a váhami objektov w_i platí:*

$$(Y_{(med)} > y_{(l)}) \Longleftrightarrow \left(\sum_{i=1}^{K} T_i(l).w_i < 0.5\right) \quad (64)$$

Dôkaz 4.1.

$$(\sum_{i=1}^{K} T_i(l)w_i < 0.5) = \left(\sum_{i=1}^{K} n_i F_{n_i}(y_{(l)})\frac{N_i}{n_i N} < 0.5\right)$$
$$= (F_n(y_{(l)}) < 0.5) = (y_{(l)} < F_n^{-1}(0.5)) = (y_{(l)} < Y_{(med)})$$

□

Lema 4.7. *Náhodná premenná $T_i(l)$ má pre pevne dané l, $l \in [1, N]$ Hypergeometrické rozdelenie s parametrami $N_i F_{N_i}(y_{(l)})$, $(1-N_i F_{N_i}(y_{(l)}))$, n_i, kde $F_{N_i}(y_{(l)})$ označuje distribučnú funkciu základného súboru pre oblasť i v bode základného súboru $y_{(l)}$, $N_i F_{N_i}(y_{(l)})$ počet jednotiek v $i-tej$ oblasti, ktoré nadobúdajú hodnotu menšiu alebo rovnú ako $y_{(l)}$ a n_i rozsah výberu z i-tej oblasti.*
Pre také K-tice $(t_1, t_2, \ldots, t_K)$ spĺňajúce podmienky:

- $\sum_{i=1}^{K} t_i w_i < 0.5$,
- $Max(0, n_i + N_i(F_{N_i}(y_{(l)}) - 1)) \leq t_i \leq Min(n_i, N_i F_{N_i}(y_{(l)}))$ *pre* $\forall i \in \{1, 2, \ldots, K\}$,

a pre $l \in [1, N]$ platí:

$$p_l = P\left(\sum_{i=1}^{K} T_i(l)w_i < 0.5\right) = \sum_{(t_1,t_2,\ldots,t_k)} \prod_{i=1}^{K} \frac{\binom{N_i F_{N_i}(y_{(l)})}{t_i}\binom{N_i(1-F_{N_i}(y_{(l)}))}{n_i - t_i}}{\binom{N_i}{n_i}}, \qquad (65)$$

kde $p_0 = 1$.

Dôkaz 4.2. *Z Hypergeometrického rozdelenia náhodnej premennej $T_i(l)$ pre t spĺňajúce vlastnosti kombinačného čísla:*

$$Max(0, n_i + N_i(F_{N_i}(y_{(l)}) - 1)) \leq t \leq Min(n_i, N_i F_{N_i}(y_{(l)})), \qquad (66)$$

platí:

$$P(T_i(l) = t) = \frac{\binom{N_i F_{N_i}(y_{(l)})}{t}\binom{N_i(1-F_{N_i}(y_{(l)}))}{n_i - t}}{\binom{N_i}{n_i}}.$$

Označme $\mathrm{T(l)} = (T_1(l), T_2(l), \ldots, T_K(l))$. *Pre danú K-ticu* $\mathrm{t} = (t_1, t_2, \ldots, t_K)$ *spĺňajúcu* (66) *pre* $\forall t_i, i \in \{1, 2, \ldots, K\}$ *platí:*

$$P(\mathrm{T(l)} = \mathrm{t}) = \prod_{i=1}^{K} \frac{\binom{N_i F_{N_i}(y_{(l)})}{t_i}\binom{N_i(1-F_{N_i}(y_{(l)}))}{n_i - t_i}}{\binom{N_i}{n_i}}.$$

Pre všetky prípustné K-tice $(t_1, t_2, \ldots, t_K)$ *spĺňajúce navyše podmienku* $\sum_{i=1}^{K} t_i w_i < 0.5$*, potom platí:*

$$P(\sum_{i=1}^{K} T_i(l) w_i < 0.5) = \sum_{(t_1,t_2,\ldots,t_k)} \prod_{i=1}^{K} \frac{\binom{N_i F_{N_i}(y_{(l)})}{t_i} \binom{N_i(1-F_{N_i}(y_{(l)}))}{n_i - t_i}}{\binom{N_i}{n_i}}.$$

□

Podľa Lemy 4.6 a Lemy 4.7, $p_l = P(Y_{(med)} > y_{(l)})$. Z toho vieme vyjadriť rovnosť:

$$P(Y_{(med)} = y_{(l)}) = p_{l-1} - p_l.$$

Pri výbere rozsahu $n = n_1 + n_2 + \cdots + n_K$ zo základného súboru rozsahu $N = N_1 + N_2 + \cdots + N_K$ potom pre r-tý moment náhodnej premennej $Y_{(med)}$ dostávame vzťah:

$$E(Y_{(med)}^r) = \sum_{l=1}^{N} (p_{l-1} - p_l) y_{(l)}^r.$$

Nakoniec získavame vzťah pre varianciu madiánu pri malom výbere:

$$V(Y_{(med)}) = E(Y_{(med)}^2) - E^2(Y_{(med)}) = \sum_{l=1}^{N} (p_{l-1} - p_l) y_{(l)}^2 - (\sum_{l=1}^{N} (p_{l-1} - p_l) y_{(l)})^2 \quad (67)$$

a jej odhad:

$$\hat{V}(Y_{(med)}) = E(Y_{(med)}^2) - E^2(Y_{(med)}) = \sum_{l=1}^{n} (\widehat{p}_{l-1} - \widehat{p}_l) Y_{(l)}^2 - (\sum_{l=1}^{n} (\widehat{p}_{l-1} - \widehat{p}_l) Y_{(l)})^2, \quad (68)$$

kde:

$$\widehat{p}_l = \sum_{(t_1,t_2,\ldots,t_k)} \prod_{i=1}^{K} \frac{\binom{N_i F_{n_i}(y_{(l)})}{t_i} \binom{N_i(1-F_{n_i}(y_{(l)}))}{n_i - t_i}}{\binom{N_i}{n_i}}. \quad (69)$$

5 Pokrytie mediánu pri simuláciách určeným intervalom spoľahlivosti

V tejto časti overíme spoľahlivosť odvodených intervalov spoľahlivosti pre medián pomocou simulácii. Na začiatku skonštruujeme ZS rovnakým postupom ako v časti (4.1.5). Jednoduchým náhodným výberom bez návratu(JNV) realizujeme najskôr výber n prvkov zo základného súboru s jednou oblasťou rozsahu N a oblastný výber rozsahu $n_1, n_2, \ldots, n_K$ z K oblastí, v našom prípade ide navyše o proporcionálny oblastný výber (OBL_p) s rozsahmi výberu spĺňajúcimi (41).

Pre ďalšie výpočty odhadneme medián výberového súboru s jednou oblasťou pomocou postupnosti ako uvádzame v časti 3.1 a pri oblastnom výbere postupnosťou váh objektov podľa časti 3.2. Budeme zachytávať pokrytie skutočného mediánu základného súboru odhadnutým intervalom spoľahlivosti pomocou:

1. Distribučnej funkcie (DF):

   Vychádzame zo vzťahu pre interval spoľahlivosti (21) s použitím vzťahu (36) pre výpočet variancie pri výbere z jednej oblasti a vzťahu (39) pri oblastnom výbere. Medián predstavuje polovičný kvantil, preto dosadzujeme $p = 0.5$. Empirické distribučné funkcie odhadujeme pre jednotlivé oblasti v už získanom odhade mediánu pri výbere z jednej aj K oblastí.

2. Hustoty (HST):

   Isterval spoľahlivosti určíme z odvodeného výrazu (54) použitím (55) pri výbere z jednej oblasti a použitím (57) pri výbere z K oblastí pričom používame odhad empirických distribučných funkcií z predchádzajúcich výpočtov. Hodnotu hustoty v mediáne odhadneme pomocou empirickej distribučnej funkcie, ako uvádzame v (53).

3. Výpočtu pre malé výbery (MV):

   Podľa vzťahu (63) vieme určiť príslušný 95%-ný interval spoľahlivosti pre jednoduchý náhodný výber bez návratu z jednej aj viacerých oblastí. Pri výpočte variancií sme najskôr určili $\hat{p}_{(l)}$ pre $l \in [1, n]$ podľa vzťahov (62), (69). Vo výpočtoch vystupuje empirická distribučná funkcia výberového súboru s jednou oblasťou a

empirické distribučné funkcie jednotlivých oblastí pri oblastnom výbere, ktoré sme už použili pri predchádzajúcich spôsoboch výpočtu variancie. Kvôli časovej náročnosti sme obmedzili výpočet $\hat{p}_{(l)}$ na hodnoty $\hat{p}_{(l)} > 0.0001$. Ďalej do odhadu variancie (61) a (68) vstupujú objekty zrealizovaného výberu.

Pri odhade 95%-ného intervalu spoľahlivosti volíme príslušné kritické hodnoty normálneho rozdelenia $u_{0.025}$, $-u_{0.025}$ a určujeme pravdepodobnosť pokrytia mediánu základného súboru odhadnutým intervalom spoľahlivosti pri opakovaní 10000 krát. Zmeníme rozsah základného súboru a rozsah výberového súboru a volíme počet oblastí $K = 2$. Pravdepodobnosti pokrytia sú v Tabuľke 6. Pri väčšom rozsahu výberu dosahovala vo väčšine prípadov pravdepodobnosť pokrytia mediánu odvodeným intervalom spoľahlivosti vyššie hodnoty. Z výpočtov vidieť, že kvalita určených intervalov spoľahlivosti závisí od rozsahu základného a výberového súboru ale výsledky môžu byť ovplyvnené aj tvorbou základného súboru.

		$N = 1000$		$N = 100$	
	n	JNV	OBL_p	JNV	OBL_p
DF	30%	0.965	0.971	0.949	0.944
	10%	0.954	0.960	0.908	0.914
	5%	0.942	0.959	0.939	0.745
HST	30%	0.941	0.954	0.914	0.912
	10%	0.934	0.939	0.963	0.944
	5%	0.943	0.957	0.895	0.764
MV	30%	1.000	1.000	1.000	1.000
	10%	1.000	1.000	0.988	0.985
	5%	1.000	1.000	0.949	0.855

Tabuľka 6

6 Optimálna alokácia pri oblastnom výbere

Ako sme videli pri oblastnom výbere, v časti 4.1.5, rôzne spôsoby rozmiestnenia výberu z oblastí, proporcionálny a neproporcionálny výber, vedú k rôznym hodnotám disperzii odhadov. Dôvodom je rozdielnosť sledovaného znaku v oblastiach, teda veľké rozdiely v disperiách oblastí. Sformulujeme presnejšie úlohu, v ktorej budeme minimalizovať disperzie (38) a (56), aby sme tak získali optimálne rozsahy výberov z jenotlivých oblastí pre minimalnu disperziu výberov pri pevne daných, jednotkových, nákladoch a pevne danom rozsahu výberu $n = \sum_{i=1}^{K} n_i$.

Označme celkové náklady na realizáciu oblastného výberu C:

$$C = \sum_{i=1}^{K} c_i n_i,$$

kde c_i sú náklady plynúce zo zistenia znaku u vzorky z i-tej oblasti. To vedie na úlohu minimalizácie funkcie pri oblastnom výbere:

$$\min_{n_1,\ldots,n_K} \left\{ \sum_{i=1}^{K} p_i^2 \frac{1 - \frac{n_i}{N_i}}{n_i} F_{N_i}(r_{y,p})\{1 - F_{N_i}(r_{y,p})\} \right\}, \tag{70}$$

pri väzbe:

$$\sum_{i=1}^{K} n_i = n$$

plynúcej z pevne daného rozsahu výberu.

Lagrangeova funkcia má potom tvar :

$$L(n_1, \ldots, n_K, \lambda) = \min_{n_1,\ldots,n_K} \left\{ \sum_{i=1}^{K} p_i^2 \frac{1 - \frac{n_i}{N_i}}{n_i} F_{N_i}(r_{y,p})\{1 - F_{N_i}(r_{y,p})\} + \lambda \left(\sum_{i=1}^{K} n_i - n \right) \right\}.$$

Pomocou parciálnych derivácií hľadáme lokálne extrémy Lagrangeovej funkcie, ktoré môže nadobúdať iba v stacionárnych bodoch, preto:

$$\frac{dL(n_1, \ldots, n_K, \lambda)}{dn_1} = -\frac{N_1^2}{N^2} \frac{1}{n_1^2} F_{N_1}(r_{y,p})\{1 - F_{N_1}(r_{y,p})\} + \lambda = 0,$$

$$\vdots$$

$$\frac{dL(n_1, \ldots, n_K, \lambda)}{dn_K} = -\frac{N_K^2}{N^2} \frac{1}{n_K^2} F_{N_K}(r_{y,p})\{1 - F_{N_K}(r_{y,p})\} + \lambda = 0,$$

$$\frac{dL(n_1,\ldots,n_K,\lambda)}{d\lambda} = \sum_{i=1}^{K} n_i - n = 0.$$

Z matice druhých derivácií Lagrangeovej funkcie vidieť, že v každom stacionárnom bode $n_i > 0$ je matica kladne definitná, Lagrangeova funkcia je tak v danom bode konvexná a preto získaný stacionárny bod predstatvuje lokálne minimum.

Z prvých K derivácií platí pre každé $i \in \{1, 2, \ldots, K\}$:

$$\sqrt{\lambda} n_i = \frac{N_i}{N} \sqrt{F_{N_i}(r_{y,p})\{1 - F_{N_i}(r_{y,p})\}}, \tag{71}$$

odkiaľ po sčítaní a využití poslednej derivácie dostávame:

$$\sqrt{\lambda} n = \sum_{i=1}^{K} \frac{N_i}{N} \sqrt{F_{N_i}(r_{y,p})\{1 - F_{N_i}(r_{y,p})\}}. \tag{72}$$

Výsledkom podielu (71) a (72) je optimálny podiel rozsahu z i-tej oblasti, k pevne danému rozsahu výberového súboru:

$$\frac{n_i}{n} = \frac{N_i \sqrt{F_{N_i}(r_{y,p})\{1 - F_{N_i}(r_{y,p})\}}}{\sum_{i=1}^{K} N_i \sqrt{F_{N_i}(r_{y,p})\{1 - F_{N_i}(r_{y,p})\}}}. \tag{73}$$

7 Ďalšie metódy odvodenia intervalu spoľahlivosti pre medián pri oblastnom výbere

V tejto časti uvedieme a zhodnotíme dve metódy pre odhad mediánu pri oblastnom výbere. Uvedené metódy nie sú univerzálne ale dajú sa použiť pri niektorých predpokladoch. Vychádzať budeme z článku [4], kde výber predstavuje postupnosť nezávislých, rovnako rozdelených pozorovaní z oblasti s príslušnou spojitou distribučnou funkciou. My sa pokúsime na získané metódy pozrieť aj zo strany základného súboru s konečným rozsahom, teda výberom z nespojitých ditribučných funkcií.

7.1 Prístup cez kontingenčnú tabuľku

Prvý postup vychádza zo základného súboru rozdeleného na dve oblasti s rovnakým rozsahom oblastí $v_1 = \frac{N_1}{N} = v_2 = \frac{N_2}{N} = \frac{1}{2}$ a využíva Fisherov exaktný test, vhodný na testovanie závislosti a homogenity v kontingenčnej tabuľke pri malom rozsahu pozorovaní.

Základný súbor je charakterizovaný distribučnou funkciou $F_N(y)$, pre ktorú platí:

$$F_N(y) = \frac{1}{2}F_{N_1}(y) + \frac{1}{2}F_{N_2}(y).$$

Ak $y = r_{y,0.5}$ potom z vlastností mediánu vyplýva:

$$F_N(r_{y,0.5}) = \frac{1}{2}F_{N_1}(r_{y,0.5}) + \frac{1}{2}F_{N_2}(r_{y,0.5}) = \frac{1}{2},$$

z čoho:

$$F_{N_1}(r_{y,0.5}) + F_{N_2}(r_{y,0.5}) = 1. \tag{74}$$

Oblastným výberom zo spojitých distribučných funkcií oblastí získame výberový súbor rozsahu n, $n = n_1 + n_2$. Ďalej označme $n_i(y_{k,i} < r_{y,0.5})$ počet prvkov vo výbere z i-tej oblasti, $i \in \{1, 2\}$, ktoré sú menšie ako medián a $n_i(y_{k,i} > r_{y,0.5})$ počet prvkov vo výbere z i-tej oblasti väčších ako medián. Náhodná veličina $n_1(y_{k,1} < r_{y,0.5})$ má binomické rozdelenie s parametrom n_1 a $F_{N_1}(r_{y,0.5})$ a náhodná veličina $n_2(y_{k,2} > r_{y,0.5})$ má binomické rozdelenie s parametrom n_2 a $(1 - F_{N_2}(r_{y,0.5})) = F_{N_1}(r_{y,0.5})$ podľa predpokladu (74).

Oblasť 1	$n_1(y_{k,1} < \Theta)$	$n_1(y_{k,1} > \Theta)$
Oblasť 2	$n_2(y_{k,2} > \Theta)$	$n_2(y_{k,2} < \Theta)$

Tabuľka 7: Kontingenčná tabuľka.

Oblasť 1	$n_1 F_{n_1}(\Theta)$	$n_1(1 - F_{n_1}(\Theta))$
Oblasť 2	$n_2(1 - F_{n_2}(\Theta))$	$n_2 F_{n_2}(\Theta)$

Tabuľka 8: Kontingenčná tabuľka cez ditribučné funkcie.

Pre ľubovoľné Θ, vieme pre výberový súbor zapísať výsledky do kontingenčnej tabuľky, Tabuľka 7.

Ďalej vieme pomocou distribučných funkcií Tabuľku 7 previesť na Tabuľku 8.

V Tabuľke 8, pre Θ, ktoré nepatrí do výberového súboru, ostáva vždy súčet hodnôt v každom riadku konštantný. Pri zavedení testu na kontingenčnú tabuľku, Tabuľka 8, testujeme či rozdelenie náhodnej premennej $n_1(y_{k,1} < \Theta)$ má spoločný parameter s rozdelením $n_2(y_{k,2} > \Theta)$ a rozdelenie $n_1(y_{k,1} > \Theta)$ s rozdelením $n_2(y_{k,2} < \Theta)$. Ide o nulovú hypotézu, v ktorej vystupuje ODDS ration, pomer šancí. V kontingenčnej tabuľke $2x2$: $O_{11}, O_{12}, O_{21}, O_{22}$ ide o podiel pomerov $\frac{O_{11}:O_{12}}{O_{21}:O_{22}}$. Hypotéza o homogenite má potom tvar:

$$H_0 : ODDS = 1.$$

Všetky také Θ, pre ktoré homogenitu nezamietame na úrovni $(1 - \alpha)100\%$, predstavujú $(1-\alpha)100\%$-ný interval spoľahlivosti pre odhadovaný medián $r_{y,0.5}$ kvôli splneniu predpokladu (74).

Pri výbere rozsahu $n = n_1 + n_2$ z oblastí zo spojitými distribučnými funkciami, z exponenciálneho rozdelenia s rôznym parametrom λ pre každú z oblastí, sa pozrieme na pokrytie mediánu základného súboru nami určeným 95%-ným intervalom spoľahlivosti. Získané výsledky pravdepodobností pokrytia pri rôznych rozsahoch výberu zobrazuje Tabuľka 9, ide o dolnú hranicu 95%-ného intervalu spoľahlivosti pravdepodobnosti pokrytia, ktorú sme získali z vygenerovaných hodnôt pravdepodobnosti pokrytia.

n	pravdepodobnosť pokrytia mediánu	n	pravdepodobnosť pokrytia mediánu
300	0.943653	30	0.9674957
100	0.9421689	10	0.9439626
50	0.9365658	5	0.8526653

Tabuľka 9: Dolná hranica pravdepodobnosti pokrytia 95%-ným IS určeným Fisherovým testom pri výbere z oblastí so spojitými distribučnými funkciami.

Pri výberových schémach v predchádzajúcich kapitolách sme uvažovali výber z konečnej populácie. Vytvoríme takúto situáciu a pozrieme sa na pravdepodobnosti pokrytia mediánu pri jednoduchom náhodnom výbere s návratom a bez návratu. Realizujeme teda výber z oblastí s nespojitými distribučnými funkciami. Dolnú hranicu 95%-ného intervalu spoľahlivosti pre pravdepodobnosť pokrytia pri rôznych rozsahoch a typoch výberu zobrazuje Tabuľka 10.

$N = 1000$			$N = 100$		
n	pravdepodobnosť pokrytia mediánu		n	pravdepodobnosť pokrytia mediánu	
	s návratom	bez návratu		s návratom	bez návratu
30%	0.9258193	0.9482221	30%	0.9478286	0.9721223
10%	0.9144292	0.9420678	10%	0.9464919	0.947755
5%	0.8973252	0.9355419	5%	0.9139355	0.9082801

Tabuľka 10: Dolná hranica pravdepodobnosti pokrytia 95%-ným IS určeným Fisherovým testom pri výbere z konečného základného súboru, jednoduchý náhodný výber s návratom a bez návratu.

V tomto prípade, výberom z konečného základného súboru, sú v niektorých prípadoch dosiahnuté výsledky horšie v porovnaní s výsledkami dosiahnutými pri nezávislom výbere pozorovaní s príslušnou spojitou distribučnou funkciou oblasti. Pri výbere s návratom, môžu vznikať opakovania, pri výbere bez návratu nie je zachovaná nezávislosť, ktorá bola pri výbere zo spojitých distribučných funkcií.

7.2 Prístup cez poradové štatistiky

Poradové štatistiky sme už spomínali pri bodovom odhade pre medián usporiadaním objektov výberového súboru do neklesajúcej postupnosti. Priamo sa dajú využiť aj pri intervalovom odhade pre medián. Opíšeme danú problematiku a pozrieme sa na vlastnosti pravdepodobnosti pokrytia určenými intervalmi spoľahlivosti.

Náhodná veličina $nF_n(r_{y,p})$, predstavujúca počet prvkov výberového súboru, ktoré sú menšie alebo rovné ako $r_{y,p}$, má pri nezávislom výbere binomické rozdelenie s parametrami n, $F_N(r_{y,p}) = p$. Potom pre strednú hodnotu platí:

$$E[nF_n(r_{y,p})] = nF_N(r_{y,p}) = np.$$

Usporiadajme prvky výberového súboru $Y_1, Y_2 \ldots, Y_n$ do neklesajúcej postupnosti:

$$Y_{(1)} \leq Y_{(2)} \leq \cdots \leq Y_{(r)} \leq \cdots \leq Y_{(n-r)} \leq \cdots \leq Y_{(n)}.$$

Potom pravdepodobnosť, že interval určený r-tou a $(n-r)$-tou poradovou štatistikou pokrýva prvok $r_{y,p}$ vieme vyjadriť ako súčet pravdepodobností:

$$Pr[Y_{(r)} \leq r_{y,p} \leq Y_{(n-r)}] = Pr(r \leq nF_n(r_{y,p}) \leq (n-r)) = \sum_{i=r}^{n-r} Pr(nF_n(r_{y,p}) = i). \quad (75)$$

Lema 7.1. *Ak S predstavuje počet nastatia udalosti pri d nezávislých opakovaniach, pričom p_i označuje pravdepodobnosť nastatia udalosti pri i-tom opakovaní, $E(S) = dp$ je pevne daná a existuje také $b, c \in Z$, že:*

$$0 \leq b \leq dp \leq c \leq d,$$

potom

$$\sum_{j=b}^{c} \binom{d}{j} p^j (1-p)^{d-j} \leq Pr(b \leq S \leq c) \leq 1.$$

Dolná hranica je dosiahnutá pre $p_1 = p_2 = \cdots = p_n = p$.

Aplikáciou Lemy 7.1 na $S = nF_n(r_{y,p})$, pre $0 \leq r \leq np \leq (n-r)$ pri nezávislom náhodnom výbere, t.j. pri jednoduchom náhodnom výbere s návratom, zo základného súboru tvoreného jednou oblasťou, dochádza k rovnosti pravdepodobností nastatia udalosti pri n opakovaniach $p_1 = p_2 = \cdots = p_n = p$. Potom podľa Lemy 7.1 nadobúda pravdepodobnosť pokrytia $r_{y,p}$ dolnú hranicu, t.j. $\sum_{j=r}^{n-r} \binom{n}{j} p^j (1-p)^{n-j}$.

Pre medián, $r_{y,0.5}$, je pravdepodobnosť pokrytia:

$$Pr(r \leq nF_n(r_{y,p}) \leq (n-r)) = \sum_{j=r}^{n-r} \binom{n}{j} 0.5^n. \tag{76}$$

Podobne môžme Lemu 7.1 aplikovať pri nezávislom oblastnom výbere. Podľa binomického rozdelenia počtu prvkov v jednotlivých oblastiach $n_i F_{n_i}(r_{y,p})$:

$$E\left(n_i F_{n_i}(r_{y,p})\right) = n_i F_{N_i}(r_{y,p}).$$

Ak zavedieme podmienku proporcionálneho oblastného výberu, (41), tak vieme pre takýto oblastný výber odvodiť rovnakú hodnotu dolnej hranice pravdepodobnosti pokrytia ako pri výbere zo základného súboru nedeleného na oblasti. Použitím vlastnosti proporcionálneho oblastného výberu a vlastnosti distribučnej funkcie (19):

$$\sum_{i=1}^{K} n_i F_{N_i}(r_{y,p}) = \sum_{i=1}^{K} \frac{n_i}{n} n F_{N_i}(r_{y,p}) = \sum_{i=1}^{K} \frac{N_i}{N} n F_{N_i}(r_{y,p}) = n F_N(r_{y,p}) = np. \tag{77}$$

Stredná hodnota počtu prvkov menších alebo rovných ako $r_{y,p}$ pri oblastnom výbere v celom výberovom súbore sa dá použitím vzťahu (77) napísať :

$$E\left(\sum_{i=1}^{K} n_i F_{n_i}(r_{y,p})\right) = \sum_{i=1}^{K} E(n_i F_{n_i}(r_{y,p})) = \sum_{i=1}^{K} n_i F_{N_i}(r_{y,p}) = np. \tag{78}$$

Aplikáciou Lemy 7.1 na $S = \sum_{i=1}^{K} n_i F_{n_i}(r_{y,p})$, pre $0 \leq r \leq np \leq n-r$ dostávame pre oblastný výber dolné ohraničenie pravdepodobnosti pokrytia p kvantilu základného súboru, $r_{y,p}$:

$$Pr[Y_{(r)} \leq r_{y,p} \leq Y_{(n-r)}] = Pr\left(r \leq \sum_{i=1}^{K} n_i F_{n_i}(r_{y,p}) \leq n-r\right) \geq \sum_{j=r}^{n-r} \binom{n}{j} p^j (1-p)^{n-j}. \tag{79}$$

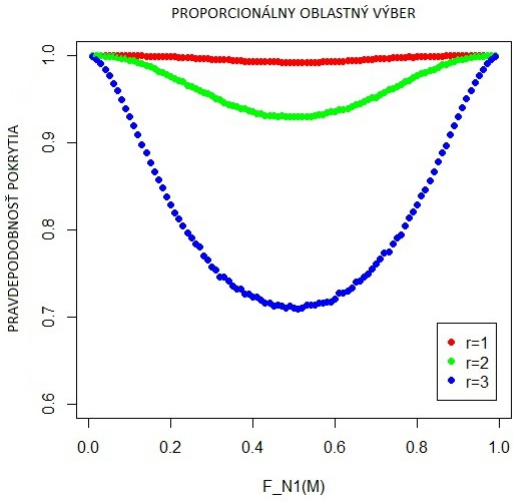
Pre medián je dolná hranica pravdepodobnosti pokrytia vyjadrená nerovnosťou:

$$Pr\left(r \leq \sum_{i=1}^{K} n_i F_{n_i}(r_{y,0.5}) \leq n-r\right) \geq \sum_{j=r}^{n-r} \binom{n}{j} 0.5^n.$$

Rovnosť nastáva iba v prípade, keď pre každé $i \in \{1, 2 \ldots, K\}$ platí $F_{N_i}(r_{y,0.5}) = 0.5$. V danom prípade je pravdepodobnosť pokrytia taká istá ako by sme dostali pri výbere z celého základného súboru nedeleného na oblasti. V prípade dvoch oblastí, ak $F_{N_1}(r_{y,0.5}) = 0.5$ tak $F_{N_2}(r_{y,0.5}) = 0.5$ za predpokladu rovnakých rozsahov oblastí základného súboru podľa (74), čo využijeme pri grafickom zobrazení.

Zo získaných vlastností pravdepodobnosti pokrytia pri týchto typoch výberov potom musí platiť, že pravdepodobnosť pokrytia p kvantilu, určená poradovou štatistikou $Y_{(r)}$ a $Y_{(n-r)}$, pri proporcionálnom výbere z dvoch oblastí, neklesne pod pravdepodobnosť pokrytia pri výbere zo základného súboru nedeleného na oblasti. Túto vlastnosť ilustruje Obr.2. Ide o podobný ilustratívny príklad ako v [4]. Obr.2 sme získali opakovanou realizáciou jednoduchého náhodného výberu z exponenciálneho rozdelenia s rôznym parametrom λ v každej oblasti (t.j. výberom zo spojitých distribučných funkcií) a rozsahmi $n_1 = 4$, $n_2 = 4$. Hodnotu mediánu základného súboru sme určili výpočtom, pomocou predpisu distribučných funkcií oblastí.

Na Obr.2 je vidieť pravdepodobnosti pokrytia intervalmi spoľahlivosti určenými poradovou štatistikou $Y_{(r)}$ a $Y_{(n-r)}$ pre $r = \{1, 2, 3\}$ pri rôznych hodnotách $F_{N_1}(r_{y,0.5})$. V každom z troch prípadov je pravdepodobnosť pokrytia zodpovedajúca výberu zo základného súboru nedeleného na oblasti (zodpovedá hodnota $F_{N_1}(r_{y,0.5}) = 0.5$) pod pravdepodobnosťou pokrytia pri proporcionálnom oblastnom výbere.



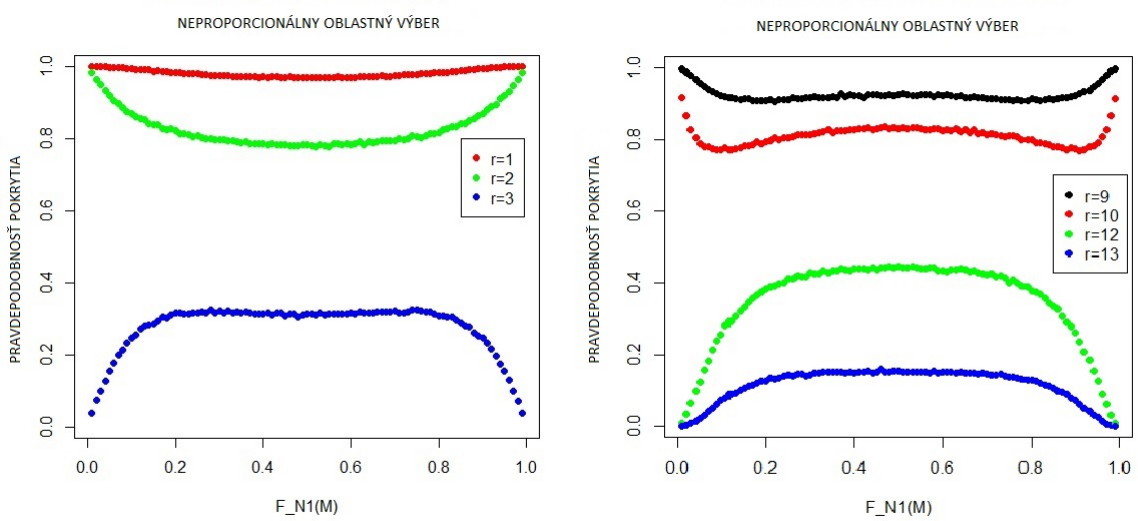
Obr. 2: Pravdepodobnosť pokrytia pri výbere z oblastí so spojitými distribučnými funkciami, proporcionálny oblastný výber.

Nutnosť podmienky proporcionálneho oblastného výberu zachytáva Obr.3 závislosti distribučnej funkcie prvej oblasti a pravdepodobnosti pokrytia mediánu intervalom spoľahlivosti, ktorý je daný poradovými štatistikami výberového súboru. Ide o neproporcionálny oblastný výber rozsahu $n_1 = 2$ a $n_2 = 4$ z dvoch oblastí v grafe naľavo kde je interval spoľahlivosti daný poradovou štatistikou $Y_{(r)}$, $r \in \{1, 2, 3\}$ a pri rozsahoch

výberov $n_1 = 10, n_2 = 16$ v grafe vpravo pre $r \in \{9, 10, 12, 13\}$.

Z oboch grafov Obr.3 je vidieť, že dochádza k prípadu, keď pravdepodobnosť pokrytia pri oblastnom, neproporcionálnom, výbere klesla pod pravdepodobnosť pokrytia pri výbere z jednej oblasti.

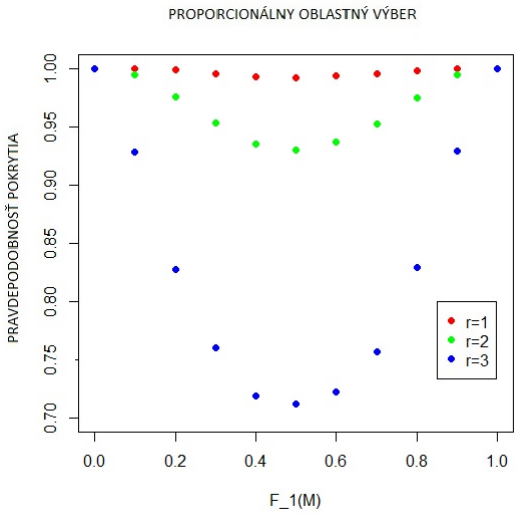
Obr.3 sme získali opäť simuláciami pravdepodobností pokrytia v prípade predpokladu exponenciálneho rozdelenia s rôznym parametrom pre každú z dvoch oblastí.



Obr. 3: Pravdepodobnosť pokrytia pri výbere z oblastí so spojitými distribučnými funkciami, neproporcionálny oblastný výber.

Aj v tomto prípade sa ďalej pozrieme na výber z konečnej populácie. Určíme pravdepodobnosti pokrytia mediánu pri jednoduchom náhodnom výbere s návratom a bez návratu vždy z oblastí s konečným rozsahom. Realizujeme teda výber z oblastí s nespojitými distribučnými funkciami. Diskrétne hodnoty predstavujúce závislosť distribučnej funkcie prvej oblasti v mediáne základného súboru a pravdepodobnosti pokrytia určeným intervalom spoľahlivosti, sú zobrazené na Obr.4 a Obr.5 . Pri výpočte pravdepodobností pokrytia najskôr nastavíme dve oblasti základného súboru ($N_1 = N_2 = 10$) a určíme medián základného súboru, $r_{y,0.5}$. Oblasti nastavíme tak (pomocou parametra λ pri dátach z exponenciálneho rozdelenia), aby postupne $F_1(r_{y,0.5}) \in \{0, 0.1, 0.2, \ldots, 1\}$. Za každým realizujeme ďalej proporcionálny oblastný výber s návratom rozsahu $n_1 = n_2 = 4$, Obr.4, a neproporcionálny oblastný výber s návratom rozsahu $n_1 = 2, n_2 = 4$, Obr.5. Hranice intervalu spoľahlivosti určujeme poradovými štatistikami $Y_{(r)}$ a $Y_{(n-r)}$, pre $r = \{1, 2, 3\}$. Pravdepodobnosť pokrytia potom zrátame opakovaním výberu (10000

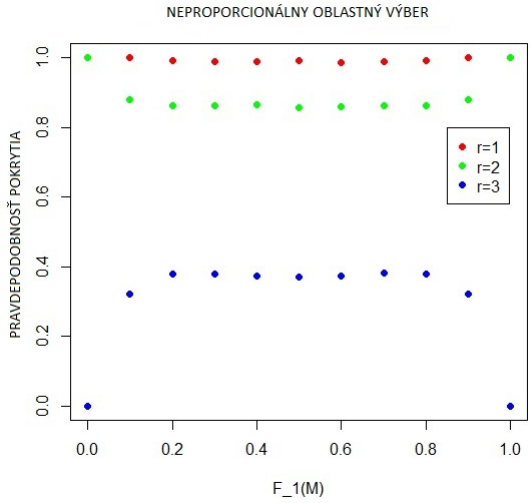
krát) z oblastí základného súboru a získaním informácie o pokrytí mediánu základného súboru daným intervalom spoľahlivosti.



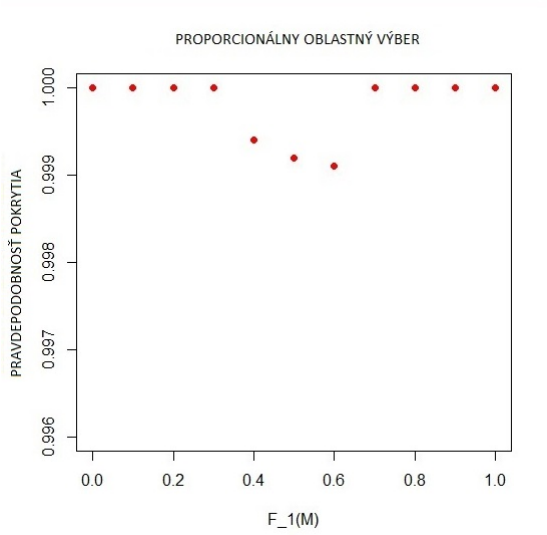
Obr. 4: Pravdepodobnosť pokrytia pri výbere z oblastí s nespojitými distribučnými funkciami, proporcionálny oblastný výber s návratom.

Z našich simulácii pri jednoduchom náhodnom výbere s návratom z konečnej množiny, spomedzi prvkov z exponenciálneho rozdelenia s daným parametrom, vidieť, že sa zachováva vlastnosť o neprekročení pravdepodobnosti pokrytia pri oblastnom proporcionálnom výbere pod pravdepodobnosť pokrytia pri výbere z jednej oblasti. Podobne z Obr.5 vidieť opodstatnenosť podmienky proporcionality oblastného výberu pri každom zvolenom r.

Ak porušíme nezávislosť pri výbere z konečnej množiny, získané hodnoty pravdepodobnosti pokrytia a nezachovanie vlastnosti o neprekročení pravdepodobnosti pokrytia ilustruje Obr.6.



Obr. 5: Pravdepodobnosť pokrytia pri výbere z oblastí s nespojitými distribučnými funkciami, neproporcionálny oblastný výber s návratom.



Obr. 6: Pravdepodobnosť pokrytia pri výbere z oblastí s nespojitými distribučnými funkciami, proporcionálny oblastný výber bez návratu.

Záver

Cieľom našej práce bolo spracovať už známe výsledky pre bodový a intervalový odhad kvantilov používané v teórii náhodného výberu a analyzovať situácie pre rôzne výberové schémy pravdepodobnostného náhodného výberu.

Na začiatku sme predstavili základnú problematiku náhodného výberu cez rôzne výberové schémy, základné pojmy a náhodné veličiny používané pri pravdepodobnostnom výbere pre ich zadefinovanie. Predstavili sme spôsob bodového odhadu kvantilu pri náhodnom výbere z jednej aj viacerých oblastí. Širší prístup k odhadu prinieslo viacero spôsobov odvodenia intervalového odhadu p kvantilu základného súboru.

V prístupe cez empirickú distribučnú funkciu sme odvodili odhady variancií pri viacerých výberových schémach a simulačne overili získané vzťahy. Pri tvorbe základného súboru s konečným rozsahom sme použili výber z exponenciálneho rozdelenia.

Pri prístupe cez asymptotické rozdelenie mediánu, vhodné pri veľkom rozsahu súboru, sme odvodili vzťah pre varianciu mediánu cez hustotu normálneho rozdelenia. Naopak, pri súbore malého rozsahu sme predstavili odvodenie použitím hypergeometrického rozdelenia náhodnej veličiny a to pri výbere bez návratu z jednej aj K oblastí.

Bližšie sme sa pozreli na oblastný náhodný výber pri určovaní intervalu spoľahlivosti pri dvoch oblastiach s určitými vlastnosťami cez kontingenčnú tabuľku a poradové štatistiky. Pri poradových štatistikách sme poukázali na vyššiu kvalitu intervalového odhadu pri proporcionálnom oblastnom výbere v porovnaní s výberom z jednej oblasti a to aj pri výbere z oblastí s diskrétnou distribučnou funkciou za podmienky nezávislého výberu.

Z našich nadobudnutých poznatkov pri písaní tejto práce, odvádzaní vzťahov a overovaní simuláciami nie je určenie najpresnejšieho spôsobu odhadu p kvantilu, resp. najčastejšie používaného mediánu, jednoznačné. Na voľbu vhodného prístupu má vplyv rozdelenie základného súboru do oblastí, rozdelenie pozorovaných znakov ale aj rozsah výberového súboru, na ktorý sme sa pozreli cez minimalizáciu disperzie.

Podarilo sa nám tak preniknúť do danej problematiky odhadov cez odvodenie vzťahov ich overením simuláciami pri exponenciálnom rozdelení a opísaním vlastností výberov za určitých podmienok. Získané postupy možno aplikovať aj na iné situácie v závislosti od riešeného problému.

Literatúra

[1] Casella, G., Berger, R. L.: *Statistical Inference*, Duxbury, USA, 2002,

[2] Gross, S. T., College, B.: *Median estimation in sample surveys*, dostupné na internete (13.04.2014)
http://www.amstat.org/sections/srms/Proceedings/papers/1980_037.pdf,

[3] Kalas, J.: *Vybrané kapitoly z teórie náhodného výberu*,skriptá, Fakulta matematiky, fyziky a informatiky Univerzity Komenského, Polygrafické stredisko UK v Bratislave, 1996,

[4] McCarthy, P. J.: *Sampling and Distribution- Free Confidence Intervals for Median*,American Statisical Association, USA, 2013,

[5] Motoyama, H.: *Note on a simple derivation of the asymptotic normality of sample quantities from a finite population*, Behaviormetrika 39(2012), No1, 1-8,

[6] Renyi, A.: *Teorie pravděpodobnosti*, Academia, nakladatelství Československé akademie věd, Praha, 1972,

[7] Särndal, K-E., Swensson, B., Wretman, J. : *Model Assisted Survey Sampling*, Springer- Verlag New York, USA, 2003

[8] Sittera, R. R, Wub Ch.: *A note on Woodruff confidence intervals for quantiles* , Department of Statistics and Actuarial Science, Simon Fraser University, Burnaby, BC, Canada, 1999.