VILNIAUS UNIVERSITETAS

Andrius Čiginas

POZICINIŲ STATISTIKŲ TIESINIŲ KOMBINACIJŲ SKIRSTINIŲ APROKSIMACIJOS BAIGTINĖSE POPULIACIJOSE

Daktaro disertacijos santrauka

Fiziniai mokslai, matematika (01P)

Vilnius, 2011

Disertacija rengta 2006-2011 metais Vilniaus universitete

Mokslinis vadovas:

Prof. habil. dr. Mindaugas Bloznelis (Vilniaus universitetas, fiziniai mokslai, matematika – 01P)

Disertacija ginama Vilniaus universiteto matematikos mokslo krypties taryboje:

Pirmininkas:

Prof. habil. dr. Alfredas Račkauskas (Vilniaus universitetas, fiziniai mokslai, matematika – 01P)

Nariai:

Prof. habil. dr. Algimantas Jonas Bikelis (Vytauto Didžiojo universitetas, fiziniai mokslai, matematika – 01P)

Prof. dr. Kęstutis Dučinskas (Klaipėdos universitetas, fiziniai mokslai, matematika – 01P)

Prof. dr. Marijus Radavičius (Vilniaus universitetas, fiziniai mokslai, matematika – 01P)

Prof. dr. Gediminas Stepanauskas (Vilniaus universitetas, fiziniai mokslai, matematika – 01P)

Oponentai:

Doc. dr. Danutė Krapavickaitė (Vilniaus Gedimino technikos universitetas, fiziniai mokslai, matematika – 01P)

Doc. dr. Marijus Vaičiulis (Vilniaus universitetas, fiziniai mokslai, matematika – 01P)

Disertacija bus ginama viešame Matematikos mokslo krypties tarybos posėdyje 2012 m. sausio mėn. 19 d. 15 val. Vilniaus universiteto Matematikos ir informatikos fakultete, Zigmo Žemaičio auditorijoje. Adresas: Naugarduko g. 24, LT-03225, Vilnius, Lietuva.

Disertacijos santrauka išsiųsta 2011 m. gruodžio mėn. ___ d.

Su disertacija galima susipažinti Vilniaus universiteto bibliotekoje.

VILNIUS UNIVERSITY

Andrius Čiginas

Approximations to distributions of linear combinations of order statistics in finite populations

Summary of Doctoral Dissertation

Physical sciences, Mathematics (01P)

Vilnius, 2011

The scientific work was carried out during 2006-2011 at Vilnius University

Scientific supervisor:

Prof. Dr. Habil. Mindaugas Bloznelis (Vilnius University, Physical Sciences, Mathematics – 01P)

The dissertation will be defended in the council of Mathematics of Vilnius University:

Chairman:
Prof. Dr. Habil. Alfredas Račkauskas (Vilnius University, Physical Sciences, Mathematics – 01P)
Members:
Prof. Dr. Habil. Algimantas Jonas Bikelis (Vytautas Magnus University, Physical Sciences, Mathematics – 01P)
Prof. Dr. Kęstutis Dučinskas (Klaipėda University, Physical Sciences, Mathematics – 01P)
Prof. Dr. Marijus Radavičius (Vilnius University, Physical Sciences, Mathematics – 01P)
Prof. Dr. Gediminas Stepanauskas (Vilnius University, Physical Sciences, Mathematics – 01P)
Oponents:
Assoc. Prof. Dr. Danutė Krapavickaitė (Vilnius Gediminas Technical University, Physical Sciences, Mathematics – 01P)
Assoc. Prof. Dr. Marijus Vaičiulis (Vilnius University, Physical Sciences, Mathematics – 01P)

The dissertation will be defended at the public meeting of the council on January 19, 2012, 3 p.m. in Vilnius University, Faculty of Mathematics and Informatics, Zigmas Žemaitis Hall (101). Address: Naugarduko str. 24, LT-03225, Vilnius, Lithuania.
The summary of dissertation was distributed on ___ of December, 2011.
The dissertation is available at the Library of Vilnius University.

1 Įvadas

Sprendžiant matematinės statistikos uždavinius susiduriama su sudėtingų atsitiktinių dydžių, vadinamų statistikomis, tikimybėmis. Tačiau tik labai atskirais atvejais statistikų tikimybinius skirstinius pavyksta tiksliai užrašyti. Todėl, daugeliu atvejų, skirstiniai aproksimuojami jų asimptotiniais skirstiniais. Geriausiai žinoma aproksimacija yra normalioji aproksimacija. Ji taikoma empiriniam vidurkiui ir tais atvejais, kai tiriama statistika yra sudėtingesnė nei imties stebėjimų suma, ir kai stebėjimai nebūtinai nepriklausomi ir vienodai pasiskirstę (n.v.p.). Dažnai pavyksta įrodyti, kad asimptotinis statistikos skirstinys (kai imties stebėjimų skaičius neaprėžtai auga) yra normalusis, tačiau praktikoje normaliosios aproksimacijos tikslumo gali nepakakti, ypač jei imties dydis nėra didelis. Vienas iš būdų patikslinti normaliąją aproksimaciją yra Edgeworth'o skleidinys, t.y., prie normaliosios aproksimacijos pridedamas vienas ar daugiau šio skleidinio narių, atspindinčių specifines statistikos ir imties savybes (žr. [16, 15]). Plačiai statistikų klasei Edgeworth'o aproksimacijos pateiktos [4] darbe.

Pagrindinis disertacijos tyrimo objektas yra *negrąžintinių* imčių pozicinių statistikų tiesinių kombinacijų (L-statistikų) trumpieji Edgeworth'o skleidiniai. Disertacijoje taip pat nagrinėjami ir kiti su skirstinių aproksimacijomis susiję klausimai: statistikų asimptotinis normalumas ir jų dispersijų vertinimas, Edgeworth'o skleidinio parametrų vertinimas, empiriniai Edgeworth'o skleidiniai, savirankos (bootstrap) aproksimacijos ir kiti.

1.1 Tikslai ir uždaviniai

Pagrindinis disertacijos uždavinys – L-statistikų skirstinių normaliosios aproksimacijos patikslinimas, kai imtys renkamos be grąžinimo. Disertacijoje keliami šie tikslai:

- Sukonstruoti trumpąjį L-statistikos Edgeworth'o skleidinį.
- Rasti išreikštines pagrindinių Hoeffding'o skleidinio komponenčių formules.
- Sukonstruoti pozicinių statistikų dispersijų viršutiniuosius įverčius.
- Nustatyti pakankamas ir paprastas normaliosios ir Edgeworth'o aproksimacijos sąlygas.
- Sukonstruoti L-statistikos dispersijos ir Edgeworth'o skleidinio parametrų įvertinius.

- Sukonstruoti ir ištirti Stjudentizuotų ir kartotinių imčių L-statistikų trumpuosius Edgeworth'o skleidinius.

1.2 Metodai

Statistikų savybėms tirti naudojame Hoeffding'o skleidinį. Teiginiams įrodyti taikome kombinatorinius ir tikimybinius metodus.

1.3 Darbo mokslinis naujumas

Disertacijoje gautos naujos Edgeworth'o aproksimacijos parametrų formulės, kurios itin patogios tiksliems (nereikalaujantiems apytikslių skaičiavimų) jų savirankos įvertiniams sudaryti. Nustatytos paprastos pakankamosios sąlygos, kurios užtikrina, kad trumpasis Edgeworth'o skleidinys yra tikslesnė statistikos skirstinio aproksimacija negu normalioji. Gautas tikslus L-statistikos dispersijos savirankos įvertinys. Sukonstruotas naujas negrąžintinės imties ekstremaliųjų reikšmių dispersijų viršutinysis įvertis.

1.4 Darbo struktūra

Disertacija parašyta anglų kalba. Ją sudaro įvadas, keturi skyriai ir literatūros sąrašas. Įvade paaiškinamos pagrindinių disertacijos uždavinių sąsajos su anksčiau išspręstais giminingais uždaviniais imčių be grąžinimo atveju ir su panašiais uždaviniais n.v.p. stebėjimų atveju. Pirmame disertacijos skyriuje pateikiame pagrindinius faktus apie Hoeffding'o skleidinį, L-statistikų atveju išvedame išreikštines pirmųjų trijų Hoeffding'o skleidinio narių ir skleidinio liekamųjų narių formules. Pirmojo skyriaus rezultatai taikomi antrajame bei sekančiuose skyriuose. Antrajame skyriuje pateikiame optimalų imties minimumo ir maksimumo dispersijų viršutinįjį įvertį. Panašius įverčius gauname ir kitoms pozicinėms statistikoms. Toliau skyriuje nustatome bendro pavidalo L-statistikų asimptotinį normalumą ir Edgeworth'o skleidinio tikslumą. Be to, asimptotinis normalumas ištiriamas ir specialiu – nupjautųjų vidurkių – atveju. Trečiasis disertacijos skyrius skirtas L-statistikos dispersijos ir parametrų, apibrėžiančių Edgeworth'o skleidinį, įvertiniams. Naudojame du įvertinių konstravimo metodus: klasikinį visrakčio (jackknife) metodą ir baigtinių populiacijų savirankos metodą. Minėtiems parametrams išvestos tikslios savirankos įvertinių formulės. Be to, keletui konkrečių L-statistikų, abu dispersijos įvertiniai palyginami kompiuterinio modeliavimo būdu. Ketvirtame disertacijos skyriuje sukonstruotos praktiniams taikymams skirtos Edgeworth'o aproksimacijos: Stjudentizuotos L-statistikos skirstinio Edgeworth'o skleidinys, empiriniai Edgeworth'o skleidiniai ir neparametrinės savirankos aproksimacijos [12]. Aptariami šių aproksimacijų tikslumai, kurie kompiuterinio modeliavimo būdu palyginami įvairioms konkrečioms L-statistikoms. Skyriaus pabaigoje pateikiamas trumpasis Edgeworth'o skleidinys sluoksninių imčių atveju, kai L-statistikos yra sluoksninės imties kvantiliai. Pateikiama išreikštinė kvantilio skirstinio aproksimacija bei jos empirinis savirankos analogas. Jų taikymas iliustruojamas naudojant realaus statistinio tyrimo duomenis.

2 Pagrindiniai rezultatai

Tegu $\mathcal{X} = \{x_1, \ldots, x_N\}$ žymi tyrimo kintamojo x matavimus populiacijoje $\mathcal{U} = \{u_1, \ldots, u_N\}$, t.y., realioji funkcija $f\colon \mathcal{U} \to \mathbb{R}$ kiekvienam populiacijos $\mathcal{U}$ elementui priskiria fiksuotą tyrimo kintamojo reikšmę. Tegu $\mathbb{X} = \{X_1, \ldots, X_n\}$ žymi dydžio $n < N$ paprastąją atsitiktinę imtį be grąžinimo iš aibės (populiacijos) $\mathcal{X}$. Tokiu atveju, stebėjimai $X_1, \ldots, X_n$ yra vienodai pasiskirstę, tačiau nėra nepriklausomi. Pozicines imties $\mathbb{X}$ statistikas pažymėkime $X_{1:n} \leq \cdots \leq X_{n:n}$, ir apibrėžkime L-statistiką:

$$L_n = L_n(\mathbb{X}) = \frac{1}{n} \sum_{j=1}^{n} c_j X_{j:n}. \tag{1}$$

Čia $c_1, \ldots, c_n$ yra duota realiųjų skaičių seka, kuri paprastai vadinama svoriais. L-statistikos apibendrina tokias gerai žinomas statistikas kaip imties vidurkis, nupjautieji vidurkiai, empiriniai kvantiliai ir Gini'o vidutinių skirtumų statistika (kiekviena iš šių statistikų gali būti užrašyta (1) pavidalu). Dažnai svorius $c_1, \ldots, c_n$ patogu nusakyti vadinamąja svorio funkcija $J\colon (0,1) \to \mathbb{R}$, su kuria

$$c_j = J\left(\frac{j}{n+1}\right), \quad 1 \leq j \leq n. \tag{2}$$

Keletą minėtų statistikų aptarkime išsamiau.

1 Pavyzdys. Fiksuotiems $0 < t_1 < t_2 < 1$ nupjautasis vidurkis apibrėžiamas taip:

$$M_{t_1;t_2} = ([t_2 n] - [t_1 n])^{-1} \sum_{j=[t_1 n]+1}^{[t_2 n]} X_{j:n},$$

kur $[\cdot]$ žymi skaičiaus sveikosios dalies funkciją. Aišku, kad tai yra (1) statistika ir ji nusakoma svorio funkcija $J(u) = (t_2 - t_1)^{-1}\mathbb{I}\{t_1 < u < t_2\}$. Čia $\mathbb{I}\{\cdot\}$ žymi indikatorinę funkciją. Pastebėkime, kad kraštutiniu atveju, kai $t_1 = 0$ ir $t_2 = 1$, statistika virsta paprasčiausiu imties vidurkiu. Tada $J \equiv 1$. Nupjautieji vidurkiai taikomi siekiant gauti stabilesnius populiacijos $\mathcal{X}$ centro įverčius, kai, pavyzdžiui, ekstremalios imties reikšmės nėra patikimos.

2 Pavyzdys. L-statistika, apibrėžiama svorio funkcija $J(u) = 6u(1-u)$, n.v.p. stebėjimų atveju taikoma logistinio skirstinio padėties parametrui vertinti, žr. [14]. Todėl, jeigu tariama, kad populiacijos $\mathcal{X}$ reikšmės yra nepriklausomi begalinės populiacijos (superpopuliacijos), turinčios logistinį skirstinį, stebėjimai, tai apibrėžtoji statistika gali būti naudinga vertinant populiacijos $\mathcal{X}$ centrą.

3 Pavyzdys. Gini'o vidutinių skirtumų statistika – vienas iš gerai žinomų sklaidos matų,

$$U_G = \binom{n}{2}^{-1} \sum_{1 \le i < j \le n} |X_i - X_j|$$

yra antrojo laipsnio U-statistika ir gali būti užrašyta (1) pavidalu (žr., pavyzdžiui, [2]), imant $c_j = (n+1)J(j/(n+1))/(n-1)$, $1 \le j \le n$, kur $J(u) = 2(2u-1)$.

Pažymėkime $\sigma_L^2 = \mathbf{Var}\, L_n$. Pagrindinis disertacijos tyrimo objektas yra pasiskirstymo funkcijos

$$F_n(x) = \mathbf{P}\left\{L_n - \mathbf{E}\, L_n \le x\sigma_L\right\} \tag{3}$$

aproksimacijos. Paprasčiausiu atveju, kai (1) yra imties vidurkis, (3) skirstinio normalioji aproksimacija (žr. [17]) ir Edgeworth'o skleidiniai (žr., pavyzdžiui, [5]) jau yra gerai ištirti. Bendro pavidalo L-statistikas galima laikyti tam tikru imties vidurkio apibendrinimu. Kita vertus, L-statistikos yra atskiras kur kas bendresnių simetrinių statistikų (simetrinių stebėjimų funkcijų) atvejis. Pastarųjų statistikų skirstinių asimptotinėms savybėms tirti [11] darbe panaudotas Hoeffding'o skleidinys, t.y., (1) statistikai galima užrašyti

$$L_n = \mathbf{E}\, L_n + \sum_{1 \le i \le n} g_1(X_i) + \sum_{1 \le i < j \le n} g_2(X_i, X_j) + \cdots . \tag{4}$$

Čia įtakos funkcijos $g_1(\cdot)$ ir $g_2(\cdot,\cdot)$ apibrėžia vadinamąsias tiesinę ir kvadratinę skleidinio dalis. Šis skleidinys yra alternatyvus Taylor'o skleidiniui, kai pastarasis negali būti pritaikytas. Tam, kad būtų įrodytas statistikos skirstinio asimptotinis normalumas, pakanka išskirti tiesinę (4) skleidinio dalį ir mokėti įvertinti likusiųjų skleidinio narių įtaką. Labai bendros sąlygos, pakankamos simetrinių statistikų skirstinių asimptotiniam normalumui, yra nustatytos [11]. Norint patikslinti normaliąją aproksimaciją trumpuoju Edgeworth'o skleidiniu, prireikia šiek tiek daugiau žinių apie (4) skleidinio narius. Iš tikrųjų, pagal [11], (3) pasiskirstymo funkcijos trumpasis Edgeworth'o skleidinys yra

$$G_n(x) = \Phi(x) - \frac{(q-p)\alpha + 3\kappa}{6\tau}(x^2-1)\Phi'(x), \tag{5}$$

kur $\Phi'(x)$ žymi standartinio normaliojo atsitiktinio dydžio pasiskirstymo funkcijos $\Phi(x)$ išvestinę ir $\tau^2 = Npq$, $p = n/N$, $q = 1-p$. Parametrams

$$\begin{aligned} \alpha &= \sigma_1^{-3}\frac{1}{N}\sum_{k=1}^{N} g_1^3(x_k), \\ \kappa &= \sigma_1^{-3}\tau^2\binom{N}{2}^{-1}\sum_{1\le k<l\le N} g_2(x_k,x_l)g_1(x_k)g_1(x_l), \end{aligned} \tag{6}$$

kur

$$\sigma_1^2 = \frac{1}{N}\sum_{k=1}^{N} g_1^2(x_k),$$

apibrėžti, kaip matome, pakanka (4) skleidinio funkcijų $g_1(\cdot)$ ir $g_2(\cdot,\cdot)$. Pažymėkime $n_* = \min\{n, N-n\}$. Kaip ir asimptotinio normalumo atveju, [11] darbe simetrinėms statistikoms gautos sąlygos, kurios užtikrina

$$\sup_{-\infty<x<+\infty} |F_n(x) - G_n(x)| = o(n_*^{-1/2}), \quad \text{kai} \quad n_* \to \infty \tag{7}$$

arba

$$\sup_{-\infty<x<+\infty} |F_n(x) - G_n(x)| = O(n_*^{-1}), \quad \text{kai} \quad n_* \to \infty, \tag{8}$$

yra labai bendros, todėl vienas iš disertacijos tikslų yra supaprastinti tas sąlygas, t.y., pakeisti jas sąlygomis L-statistikos svoriams $c_1, \ldots, c_n$ ir atsitiktinio dydžio X_1 momentams.

Imčių be grąžinimo atveju galima rasti vos keletą darbų, skirtų L-statistikų skirstinių aproksimacijoms. Pavyzdžiui, L-statistikų asimptotinis normalumas gana sudėtingiems imčių sudarymo modeliams nagrinėtas [28]. Žymiai švelnesnės sąlygos, pakankamos asimptotiniam imties kvantilio normalumui, paprastųjų atsitiktinių imčių be grąžinimo atveju gautos [13]. Tuo tarpu n.v.p. stebėjimų atveju asimptotinės L-statistikų savybės yra tirtos daugybės autorių. Pavyzdžiui, asimptotinis normalumas nagrinėtas [14, 30, 31], o Edgeworth'o skleidiniai tirti [20, 26, 18, 1, 22]. Palyginimui, disertacijoje gautos asimptotiniam normalumui pakankamos sąlygos (žr. 7 teoremą) yra panašios į sąlygas [31] darbe, o Edgeworth'o aproksimacijos atveju panašios sąlygos naudotos [20, 26].

2.1 Hoeffding'o skleidinys ir jo taikymai

Pirmajame disertacijos skyriuje, naudojant žinomas formules simetrinėms statistikoms (pagal [11]), gaunamos pirmųjų (1) statistikos Hoeffding'o skleidinio narių išraiškos. Čia pateiksime išreikštines ir paprastas įtakos funkcijų $g_1(\cdot)$ ir $g_2(\cdot,\cdot)$ (4) skleidinyje išraiškas. Neprarandant bendrumo galima laikyti, kad populiacijos $\mathcal{X}$ reikšmės yra išrikiuotos nemažėjančia tvarka, t.y., $x_1 \leq \cdots \leq x_N$. Pažymėkime $\triangle_i = x_{i+1} - x_i$, $1 \leq i \leq N-1$. Tegu

$$\mathcal{H}_{N,n,i}(j) = \binom{i}{j}\binom{N-i}{n-j} \Big/ \binom{N}{n}$$

žymi tikimybę, kad hipergeometrinis atsitiktinis dydis, nusakytas parametrais N, n ir i, įgyja reikšmę j.

4 Teorema.

(i) Visiems $1 \leq k \leq N$,

$$g_1(x_k) = -n^{-1} \sum_{j=1}^{n} c_j \sum_{i=1}^{N-1} \left(\mathbb{I}\{i \geq k\} - \frac{i}{N} \right) \mathcal{H}_{N-2,n-1,i-1}(j-1) \triangle_i \,. \tag{9}$$

(ii) Visiems $1 \leq k < l \leq N$,

$$g_2(x_k, x_l) = -n^{-1} \sum_{j=2}^{n} (c_j - c_{j-1}) \sum_{i=1}^{N-1} \phi_{k,l}(i) \mathcal{H}_{N-4,n-2,i-2}(j-2) \triangle_i, \tag{10}$$

kur

$$\phi_{k,l}(i) = \begin{cases} i(i-1)/A, & \textit{kai } 1 \leq i < k, \\ -(i-1)(N-i-1)/A, & \textit{kai } k \leq i < l, \\ (N-i-1)(N-i)/A, & \textit{kai } l \leq i < N, \end{cases} \tag{11}$$

kai čia $A = (N-1)(N-2)$.

Pirmajame disertacijos skyriuje taip pat gautos ir tam tikrų Hoeffding'o skleidinio liekamųjų narių komponenčių išraiškos, kurios, kaip ir skleidinio įtakos funkcijos, išreikštos per svorius $c_1, \ldots, c_n$ bei jų skirtumus.

Optimalus imties ekstremaliųjų reikšmių dispersijos įvertis. Siekdami išvengti trivialių atvejų, laikykime, kad $\mathbf{Var}\, X_1 > 0$.

5 Teorema. *Kai $m = 1$ ir $m = n$, tai turime*

$$\mathbf{Var}\, X_{m:n} \leq n \frac{N-n}{N-1} \mathbf{Var}\, X_1. \tag{12}$$

6 Pastaba. *Kai $m = 1$ ir $m = n$,* (12) *įvertis yra optimalus, t.y., galima rasti tokių populiacijų, kurioms* (12) *nelygybėje pasiekiama lygybė.*

Pozicinių statistikų dispersijų įverčiai buvo nagrinėti [23, 24, 25, 27]. Tiksliau, n.v.p. stebėjimų atveju, imties ekstremaliosioms reikšmėms yra gautas įvertis $\mathbf{Var}\, X_{m:n} \leq n\, \mathbf{Var}\, X_1$, žr. [24]. Toks pat įvertis yra optimalus ir tuo atveju, kai stebėjimai yra vienodai pasiskirstę, tačiau susieti bet kokia priklausomybe, žr. [27]. Kai imtys renkamos be grąžinimo, šiuos rezultatus baigtinės populiacijos pataisos daugikliu $(N - n)/(N - 1) < 1$ pagerina 5 teorema. Naująjį įvertį imties ekstremaliosioms reikšmėms įdomu palyginti ir su įverčiu $\mathbf{Var}\, X_{m:n} \leq (n/2)\, \mathbf{Var}\, X_1$, kuris gaunamas n.v.p. stebėjimų atveju, papildomai tariant, kad dydžio X_1 skirstinys yra simetrinis, žr. [23].

Darbuose [24, 27], o taip pat ir [25] (kur nagrinėjamas tas pats imties sudarymo modelis kaip ir [23]), optimalūs įverčiai gauti ir pozicinių statistikų $X_{m:n}$, $2 \leq m \leq n-1$ dispersijoms. Kai imtys renkamos be grąžinimo, naudodami 5 teoremos įrodymo metodus, gauname ir šių pozicinių statistikų dispersijų įverčius

$$\mathbf{Var}\, X_{m:n} \leq n \frac{N-n}{N-1} \max_{1 \leq i \leq N-1} \mathcal{H}_{N-2,n-1,i-1}(m-1)\, \mathbf{Var}\, X_1, \tag{13}$$

tačiau, bendru atveju, jie nėra optimalūs ta prasme, kad (13) nelygybė yra visuomet griežta visiems $n_* = \min\{n, N-n\} \geq 3$.

Aišku, kad imtys be grąžinimo yra atskiras [27] nagrinėto modelio atvejis. Todėl pastebėsime, kad (13) įverčiai pagerina atitinkamus [27] įverčius tuo atveju, kai baigtinės populiacijos pataisos daugiklis yra pakankamai mažas. Jeigu, imties ekstremaliųjų reikšmių atveju, fiksuotume imties dydį n ir leistume populiacijos dydžiui N neaprėžtai augti, tai imtis be grąžinimo virstų n.v.p. stebėjimų imtimi ir (12) įvertis taptų toks pat kaip ir [24].

Asimptotinis normalumas. Asimptotiniam L-statistikos elgesiui apibūdinti naudokime centruotą statistiką su $n^{1/2}$ normavimu:

$$S_n = n^{1/2}(L_n - \mathbf{E}\, L_n). \tag{14}$$

Pažymėkime $\tilde{\sigma}_n^2 = \mathbf{Var}\, S_n$. Sakykime, kad svoriai $c_1, \ldots, c_n$ yra apibrėžti (2) sąryšiu.

Būtina atkreipti dėmesį, kad, iš tikrųjų, žemiau esančių asimptotinių rezultatų formuluotėse turi būti kalbama apie baigtinių populiacijų seką $\mathcal{X}_r = \{x_{r,1}, \ldots, x_{r,N_r}\}$, kur $N_r \to \infty$ kai $r \to \infty$, ir apie L-statistikų seką $L_{n_r}(\mathbb{X}_r)$, kai čia $\mathbb{X}_r = \{X_{r,1}, \ldots, X_{r,n_r}\}$ yra imtis be grąžinimo iš $\mathcal{X}_r$. Norėdami supaprastinti žymėjimus indeksą r praleisime.

Pirmiausia pateiksime 3 teiginio [11] darbe išvadą, kurioje asimptotiniam normalumui pakankamos sąlygos formuluojamos svorio funkcijos $J(\cdot)$ ir dydžio X_1 momentų terminais. Nagrinėkime Lindeberg'o tipo Erdős–Rényi sąlygą: kiekvienam $\varepsilon > 0$,

$$\mathbf{E}\, g_1^2(X_1)\sigma_1^{-2}\mathbb{I}\{|g_1(X_1)| > \varepsilon\tau\sigma_1\} = o(1) \quad \text{kai} \quad n_* \to \infty. \tag{15}$$

Sakysime, kad funkcija $J(\cdot)$ intervale $(0,1)$ tenkina $\delta \geq 0$ eilės Hölder'io sąlygą, jeigu kažkokiai konstantai $B \geq 0$ galioja $|J(u) - J(v)| \leq B\,|u - v|^\delta$ visiems $u, v \in (0,1)$. Sakysime, kad atsitiktinis dydis $\tilde{\sigma}_n^{-1} S_n$ arba jo pasiskirstymo funkcija $F_n(x)$ yra asimptotiškai standartinė normalioji, jeigu kiekvienam $x \in (-\infty, +\infty)$ turime $\lim_{n\to\infty} F_n(x) = \Phi(x)$.

7 Teorema. *Tegu $n_* \to \infty$ ir tegu $\tilde{\sigma}_n$ išlieka teigiama visiems n_*. Sakykime, kad $\mathbf{E}\, X_1^2 < \infty$, ir $J(\cdot)$ yra aprėžta ir intervale $(0,1)$ tenkina $\delta > 1/2$ eilės Hölder'io sąlygą. Tegu galioja* (15)*. Tada $\tilde{\sigma}_n^{-1} S_n$ yra asimptotiškai standartinis normalusis.*

Antrasis 3 teiginio [11] taikymas skirtas specialiam L-statistikų atvejui – nupjautąjam vidurkiui. Pastebėkime, kad šiuo atveju (žr. 1 pavyzdį) svorio funkcija $J(u)$ nėra pakankamai glodi, t.y., ji aprėžta, tačiau netenkina Hölder'io sąlygos. Įveskime papildomą populiacijos $\mathcal{X}$ glodumo sąlygą. Laikykime, neprarasdami bendrumo, kad $x_1 \leq \cdots \leq x_N$. Sakykime, kad kažkokioms konstantoms $C > 0$ ir $1/2 < \delta \leq 1$ galioja

$$|x_m - x_l| \leq CN^{-\delta}\,|m - l| \tag{16}$$

visiems $1 \leq l < m \leq N$.

8 Teorema. *Tegu $n_* \to \infty$ ir tegu $\tilde{\sigma}_n$ išlieka teigiama visiems n_*. Sakykime, kad $\mathbf{E}\, X_1^2 < \infty$. Tarkime, kad kažkokiam $1/2 < \delta \leq 1$ yra patenkinta* (16) *sąlyga ir* $(1 - n/N)^{-2}\, nN^{2(\delta-1)} \to$

∞. *Tegu galioja* (15). *Tada, nupjautąjam vidurkiui,* $\tilde{\sigma}_n^{-1} S_n$ *yra asimptotiškai standartinis normalusis.*

Galima tokia (16) sąlygos interpretacija. Imkime $l = 1$ ir $m = N$. Jeigu populiacija $\mathcal{X}$ aprėžta, tai sąlyga galioja reikšmei $\delta = 1$. Kitu kraštutiniu atveju, kai $\delta = 1/2$, (16) sąlyga galioja bet kokiai baigtinei populiacijai, remiantis Nair–Thomson'o nelygybe $x_N - x_1 \leq \sigma\sqrt{2N}$, kur $\sigma^2 = \mathbf{Var}\, X_1$ (žr., pavyzdžiui, [3]). Taigi, (16) sąlyga atrodo labai švelni mažiems $\epsilon > 0$ iš $\delta = 1/2 + \epsilon$, t.y., ji galioja daugumai populiacijų. Aišku, tada 8 teoremoje n turi tolti į begalybę per daug neatsilikdamas nuo N, tuo tarpu, $\delta = 1$ atveju pakanka, kad $n \to \infty$ kiek norima lėtai lyginant su populiacijos dydžiu N.

Edgeworth'o skleidinys. Nagrinėkime (14) statistiką ir tegu jos svorius $c_1, \dots, c_n$ generuoja funkcija $J\colon (0,1) \to \mathbb{R}$. Funkcijai $g\colon \mathbb{R} \to \mathbb{C}$ žymėkime $\|g\|_{[a,b]} = \sup_{a<|t|<b} |g(t)|$, ir nagrinėkime tokią negardeliškumo sąlygą: kiekvienam $\varepsilon > 0$ ir kiekvienam $B > 0$ funkcija $\varphi(t) = \mathbf{E} \exp\{it\sigma_1^{-1} g_1(X_1)\}$ tegu tenkina

$$\liminf_{n_* \to \infty} \|\varphi\|_{[\varepsilon, B]} < 1. \tag{17}$$

Taip pat nagrinėkime Cramer'o-tipo sąlygą

$$\liminf_{n_* \to \infty} \|\varphi\|_{[\varepsilon, \tau]} < 1. \tag{18}$$

Sekančios teoremos apie statistikos $\tilde{\sigma}_n^{-1} S_n$ skirstinio aproksimavimo (5) Edgeworth'o skleidiniu tikslumą įrodyme remiamasi 2 teorema iš [11]. Nagrinėjame atvejį, kai svorio funkcija $J(\cdot)$ yra pakankamai glodi.

9 Teorema. *Tegu* $n_* \to \infty$ *ir tegu* $\tilde{\sigma}_n$ *išlieka teigiama visiems* n_*.

(i) Tegu galioja (17). *Sakykime, kad intervale* $(0,1)$ *funkcijos* $J(\cdot)$ *antroji išvestinė* $J''(\cdot)$ *yra aprėžta ir, kažkokiam* $\delta > 0$, $\mathbf{E}\,|X_1|^{3+\delta} < \infty$. *Tada galioja* (7).

(ii) Tegu galioja (18). *Sakykime, kad intervale* $(0,1)$ *funkcijos* $J(\cdot)$ *trečioji išvestinė* $J'''(\cdot)$ *yra aprėžta ir* $\mathbf{E}\,|X_1|^4 < \infty$. *Tada galioja* (8).

Asimptotinės sąlygos (17) ir (18), panaudotos baigtinėms populiacijoms [11] darbe, yra analogiškos negardeliškumo ir Cramer'o sąlygoms, naudojamoms tradiciniu n.v.p. stebėjimų atveju, žr., pavyzdžiui, [4]. Sąlyginis jų paprastumas glūdi tame, kad jos formuluojamos tik tiesinei statistikos daliai kaip ir paprasčiausiu imties vidurkio atveju, žr., pavyzdžiui, [5].

Pateikiame kompiuterinio modeliavimo pavyzdį, kuris rodo kaip (5) trumpasis Edgeworth'o skleidinys patikslina normaliąją aproksimaciją.

10 Modeliavimas. Dydžio $N = 100$ populiaciją $\mathcal{X}$ sudarykime generuodami logistinio skirstinio $\mathcal{L}(0,1)$ reikšmes. Sudarytos populiacijos $\mathcal{X}$ reikšmių vidurkis ir dispersija yra 0.004 ir 3.270. Nagrinėkime 2 pavyzdžio L-statistiką. Pastebėkime, kad jos svorio funkcijos glodumas yra pakankamas 9 teoremos taikymui. Imtims, kurių dydžiai $n = 5, 15, 30$, pateikiame keletą funkcijų $\tilde{F}_n$, G_n ir Φ q-kvantilių, žr. 1 lentelę. Čia $\tilde{F}_n$ žymi pasiskirstymo funkcijos F_n Monte–Karlo aproksimaciją, kuri gauta nepriklausomai generuojant 10^7 imčių be grąžinimo iš populiacijos $\mathcal{X}$.

1 lentelė: $\tilde{F}_n$, $n = 5, 15, 30$ aproksimacijos.

$q =$	0.01	0.05	0.10	0.25	0.50	0.75	0.90	0.95	0.99
$\tilde{F}_5^{-1}(q)$	-2.12	-1.53	-1.22	-0.69	-0.06	0.63	1.32	1.75	2.58
$G_5^{-1}(q)$	-2.08	-1.56	-1.25	-0.70	-0.06	0.64	1.32	1.75	2.56
$\tilde{F}_{15}^{-1}(q)$	-2.16	-1.58	-1.25	-0.69	-0.04	0.65	1.31	1.71	2.49
$G_{15}^{-1}(q)$	-2.16	-1.59	-1.26	-0.70	-0.04	0.65	1.31	1.71	2.49
$\tilde{F}_{30}^{-1}(q)$	-2.20	-1.60	-1.27	-0.69	-0.03	0.66	1.30	1.69	2.43
$G_{30}^{-1}(q)$	-2.22	-1.61	-1.27	-0.69	-0.03	0.66	1.30	1.69	2.43
$\Phi^{-1}(q)$	-2.33	-1.64	-1.28	-0.67	0.00	0.67	1.28	1.64	2.33

Matome, kad net labai mažo dydžio imtims Edgeworth'o aproksimacija yra efektyvi ir efektyvesnė už normaliąją.

2.2 Parametrų vertinimas

Trečiajame disertacijos skyriuje nagrinėjamas L-statistikos parametrų σ_L^2 bei α ir κ vertinimas. Naudojami klasikinis visrakčio ir [12] savirankos metodai.

Visrakčio metodas. Imčių be grąžinimo atveju, klasikinis dispersijos σ_L^2 visrakčio įvertinys nuo naudojamo n.v.p. stebėjimų atveju skiriasi tik baigtinės populiacijos pataisos daugikliu. Sakykime, kad L-statistikos svoriai $c_1, \ldots, c_n$ yra nusakyti (2) sąryšiu. Tuomet, imčiai be grąžinimo $\mathbb{X}$, visrakčio įvertinys yra apibrėžiamas taip:

$$S^2(L_n) = \Big(1 - \frac{n}{N}\Big) \frac{n-1}{n} \sum_{k=1}^{n} \big(L_{(k)} - \overline{L}\big)^2, \qquad \overline{L} = \frac{1}{n} \sum_{k=1}^{n} L_{(k)}, \tag{19}$$

kur $L_{(k)} = L_{n-1}(\mathbb{X} \backslash \{X_k\})$, $1 \le k \le n$ yra (1) L-statistikos su svoriais $c_j' = J(j/n)$, $1 \le j \le n-1$. Labai panašaus į šį įvertinio savybės yra tirtos, pavyzdžiui, [7, 11].

Parametrų α ir κ visrakčio įvertinius apibrėžiame panašiai kaip ir baigtinių populiacijų simetrinėms statistikoms, žr. [6]. Visiems $1 \le k \le n$, $1 \le i, j, r \le n$, $i \ne j$ pažymėkime

$$V_k = \overline{L} - L_{(k)}, \qquad \tilde{V}_r = \overline{\overline{L}} - \overline{L}_{(r)}, \qquad W_{ij} = \overline{\overline{L}} - \overline{L}_{(i)} - \overline{L}_{(j)} + L_{(i,j)},$$

kur

$$\overline{L}_{(r)} = \frac{1}{n-1} \sum_{1 \le j \le n,\ j \ne r} L_{(r,j)}, \qquad \overline{\overline{L}} = \binom{n}{2}^{-1} \sum_{1 \le i < j \le n} L_{(i,j)}.$$

Čia $L_{(i,j)} = L_{n-2}(\mathbb{X} \backslash \{X_i, X_j\})$ yra (1) pavidalo statistika su svoriais $c_l'' = J(l/(n-1))$, $1 \le l \le n-2$. Tada

$$\begin{aligned} \hat{\alpha}_J &= \hat{\sigma}_J^{-3} n^{1/2} \sum_{k=1}^{n} V_k^3, \\ \hat{\kappa}_J &= 2\hat{\sigma}_J^{-3} \left(1 - \frac{n}{N}\right) n^{1/2} \sum_{1 \le i < j \le n} W_{ij} \tilde{V}_i \tilde{V}_j, \end{aligned} \tag{20}$$

kur

$$\hat{\sigma}_J^2 = \sum_{k=1}^{n} V_k^2.$$

Panašių įvertinių savybės yra tirtos [6].

Saviranko metodas. Nagrinėjamiems parametrams sudarome išreikštinius [12] darbe pasiūlytos saviranko įvertinius. Taip pavyksta išvengti papildomos paklaidos, kuri atsiranda pačius saviranko įvertinius vertinant Monte–Karlo metodais. Panašus uždavinys L-statistikų dispersijai, kai stebėjimai yra n.v.p. ir naudojama klasikinė Efron'o saviranka, yra išspręstas [21] darbe. Įveskime skaičius

$$h_i(u) = \binom{N}{n}^{-1} \binom{i}{u} \binom{N-i}{n-u}, \qquad 0 \le u \le n-1, \quad 1 \le i \le N-1,$$

$$h_{ij}(u) = \binom{N}{n}^{-1} \binom{i}{u} \binom{N-j}{n-u}, \qquad 0 \le u \le n-1, \quad 1 \le i < j \le N-1,$$

ir

$$h_{ij}(u,v) = \binom{N}{n}^{-1} \binom{i}{u} \binom{j-i}{v-u} \binom{N-j}{n-v}, \qquad 0 \le u < v \le n-1, \quad 1 \le i < j \le N-1.$$

Užrašykime $N = mn + t$, kur $0 \le t < n$. Trumpumo dėlei rašykime $b_{ik} = \mathcal{H}_{n,t,i}(k)$, $0 \le k \le t$,

$1 \le i \le n-1$ ir pažymėkime

$$b_{ijkl} = \binom{n}{t}^{-1}\binom{i}{k}\binom{j-i}{l-k}\binom{n-j}{t-l}, \qquad 0 \le k \le l \le t, \quad 1 \le i < j \le n-1.$$

Tada dispersijos σ_L^2 savirankos įvertinys:

$$\hat{\sigma}_B^2 = n^{-2}\Bigg[\sum_{p=1}^{n} c_p^2 \widehat{\mathbf{Var}} X_{p:n} + 2\sum_{1\le p<r\le n} c_p c_r \Big\{\widehat{\mathbf{Var}} X_{p:n} + \sum_{u=0}^{p-1}\sum_{v=p}^{r-1}(\hat{\theta}_{uvB} - \hat{\theta}_{uB}\hat{\theta}_{vB})\Big\}\Bigg], \tag{21}$$

kur

$$\widehat{\mathbf{Var}} X_{p:n} = \sum_{u=0}^{p-1}(\hat{\theta}_{uuB} - \hat{\theta}_{uB}^2) + 2\sum_{0\le u<v\le p-1}(\hat{\theta}_{uvB} - \hat{\theta}_{uB}\hat{\theta}_{vB}), \qquad 1 \le p \le n.$$

Čia $\hat{\theta}_{uB}$, $\hat{\theta}_{uuB}$ ir $\hat{\theta}_{uvB}$ yra tikslūs populiacijos parametrų $\theta_u = \mathbf{E}\,\boldsymbol{\Delta}_{u:n}$, $\theta_{uu} = \mathbf{E}\,\boldsymbol{\Delta}_{u:n}^2$ ir $\theta_{uv} = \mathbf{E}\,\boldsymbol{\Delta}_{u:n}\boldsymbol{\Delta}_{v:n}$, kai čia pažymėta $\boldsymbol{\Delta}_{j:n} = X_{j+1:n} - X_{j:n}$, $0 \le j \le n-1$, savirankos įvertiniai, kuriuos pateikiame 11 teoremoje žemiau. Neapibrėžtumo $\boldsymbol{\Delta}_{0:n}$ išvengiame įvedę skaičių $x_0 := \min\{x_1, \ldots, x_N\}$ ir priskyrę $X_{0:n} := x_0$, ir laikydami, kad, beveik visur, $X_{0:n} \le X_{j:n}$ visiems $1 \le j \le n$.

11 Teorema. *Parametrų θ_u, θ_{uu}, $0 \le u \le n-1$ ir θ_{uv}, $0 \le u < v \le n-1$ savirankos įvertiniai yra*

$$\hat{\theta}_{uB} = \sum_{i=1}^{n-1}\sum_{k=0}^{t} h_{mi+k}(u) b_{ik}\boldsymbol{\Delta}_{i:n},$$

$$\hat{\theta}_{uuB} = \sum_{i=1}^{n-1}\sum_{k=0}^{t} h_{mi+k}(u) b_{ik}\boldsymbol{\Delta}_{i:n}^2 + 2\sum_{1\le i<j\le n-1}\sum_{0\le k\le l\le t} h_{mi+k;mj+l}(u) b_{ijkl}\boldsymbol{\Delta}_{i:n}\boldsymbol{\Delta}_{j:n},$$

$$\hat{\theta}_{uvB} = \sum_{1\le i<j\le n-1}\sum_{0\le k\le l\le t} h_{mi+k;mj+l}(u,v) b_{ijkl}\boldsymbol{\Delta}_{i:n}\boldsymbol{\Delta}_{j:n}.$$

Sudaryto (21) įvertinio savybės kol kas lieka nepakankamai gerai žinomos. Apie galimą nagrinėjamos savirankos įvertinių poslinkį, kai vertinami parametrai yra statistikų dispersijos, diskutuojama [12], taip pat žr. [9]. Disertacijoje apsiribojama (21) įvertinio palyginimu su (19) visrakčio įvertiniu kompiuterinio modeliavimo būdu įvairioms L-statistikoms.

Pateiksime parametrų α ir κ savirankos įvertinius. Juos sudarome panašiu principu kaip ir dispersijos σ_L^2 įvertinį. Pirmiausia, kiekvienam iš populiacijos $\mathcal{X}$ parametrų $g_1(k) := g_1(x_k)$, $1 \le k \le N$ ir $g_2(k,l) := g_2(x_k, x_l)$, $1 \le k < l \le N$, žr. (9) ir (10), randame tikslius savirankos įvertinius $\hat{g}_{1B}(k)$ ir $\hat{g}_{2B}(k,l)$.

12 Teorema.

(i) Visiems $1 \le k \le N$,

$$\hat{g}_{1B}(k) = -n^{-1} \sum_{j=1}^{n-1} \sum_{i=mj}^{mj+t} \left(\mathbb{I}\{i \ge k\} - \frac{i}{N} \right) \mathcal{H}_{n,t,j}(i-mj) \sum_{p=1}^{n} c_p \mathcal{H}_{N-2,n-1,i-1}(p-1) \mathbf{\Delta}_{j:n}.$$

(ii) Visiems $1 \le k < l \le N$,

$$\hat{g}_{2B}(k,l) = -n^{-1} \sum_{j=1}^{n-1} \sum_{i=mj}^{mj+t} \phi_{k,l}(i) \mathcal{H}_{n,t,j}(i-mj) \sum_{p=2}^{n} (c_p - c_{p-1}) \mathcal{H}_{N-4,n-2,i-2}(p-2) \mathbf{\Delta}_{j:n},$$

kai čia $\phi_{k,l}(i)$ *yra iš* (11).

Tada nežinomus (6) išraiškų parametrus pakeitę 12 teoremos įvertiniais gauname savirankos įvertinius

$$\begin{aligned} \hat{\alpha}_B &= \hat{\sigma}_{1B}^{-3} \frac{1}{N} \sum_{k=1}^{N} \hat{g}_{1B}^3(k), \\ \hat{\kappa}_B &= \hat{\sigma}_{1B}^{-3} \tau^2 \binom{N}{2}^{-1} \sum_{1 \le k < l \le N} \hat{g}_{2B}(k,l) \hat{g}_{1B}(k) \hat{g}_{1B}(l), \end{aligned} \tag{22}$$

kur

$$\hat{\sigma}_{1B}^2 = \frac{1}{N} \sum_{k=1}^{N} \hat{g}_{1B}^2(k).$$

2.3 Taikymai

Ketvirtajame disertacijos skyriuje aptariamas praktinis Edgeworth'o skleidinių taikymas, kai nežinomi jų parametrai pakeičiami įvertiniais. Keletos tokių aproksimacijų palyginimas kompiuterinio modeliavimo būdu parodo, kurios iš jų (kokioms statistikoms) tikslesnės.

Pirmiausia disertacijoje aptariame Stjudentizuotos L-statistikos skirstinio

$$F_{nS}(x) = \mathbf{P}\left\{ L_n - \mathbf{E}\, L_n \le x S(L_n) \right\}, \tag{23}$$

kur $S^2(L_n)$ yra (19) visrakčio įvertinys, aproksimaciją trumpuoju Edgeworth'o skleidiniu

$$G_{nS}(x) = \Phi(x) + \frac{(q-p+(q+1)x^2)\alpha + 3(x^2+1)\kappa}{6\tau} \Phi'(x). \tag{24}$$

Ši aproksimacija baigtinių populiacijų simetrinėms statistikoms sukonstruota [8]. Ją apibrėžiantys parametrai turi tas pačias reikšmes kaip ir apibrėžiantieji (5) Edgeworth'o aproksimaciją. Pažymėtina, kad (3) ir (23) skirstinių asimptotinės savybės yra panašios, tiksliau, apie (23) skirstinio normaliąją ir Edgeworth'o aproksimacijas gali būti formuojami labai panašūs tvirtinimai kaip ir 7–9 teoremose.

Paprastai populiacijos parametrai α ir κ, nusakantys (5) ir (24) Edgeworth'o aproksimacijas, nėra žinomi. Todėl praktikoje jie pakeičiami įvertiniais. Taip gaunami vadinamieji empiriniai Edgeworth'o skleidiniai. Disertacijoje lyginamos (3) ir (23) skirstinių dviejų rūšių empirinės Edgeworth'o aproksimacijos: kai parametrai α ir κ pakeičiami (20) visrakčio įvertiniais $\hat{\alpha}_J$ ir $\hat{\kappa}_J$, tada iš (5) ir (24) gauname atitinkamas empirines aproksimacijas $G_{nJ}(x)$ ir $G_{nSJ}(x)$ ir, kai parametrai pakeičiami (22) savirankos įvertiniais $\hat{\alpha}_B$ ir $\hat{\kappa}_B$, tada naujas aproksimacijas atitinkamai pažymime $G_{nB}(x)$ ir $G_{nSB}(x)$. Empirinės aproksimacijos $G_{nJ}(x)$ ir $G_{nSJ}(x)$ baigtinių populiacijų simetrinių statistikų atveju yra nagrinėjamos [6, 8]. Tuo tarpu $G_{nB}(x)$ ir $G_{nSB}(x)$ aproksimacijų savybės tebėra nepakankamai gerai žinomos, kol nėra įrodytas įvertinių $\hat{\alpha}_B$ ir $\hat{\kappa}_B$ suderintumas. Disertacijoje apsiribojama šių bei visų kitų aptartųjų L-statistikų skirstinių aproksimacijų palyginimu kompiuterinio modeliavimo būdu. Be to, disertacijoje jos lyginamos ir su tam tikromis neparametrinėmis (3) ir (23) skirstinių [12] savirankos aproksimacijomis.

Kompiuterinis modeliavimas. Stjudentizuotos L-statistikos pasiskirstymo funkcijos F_{nS} Edgeworth'o aproksimaciją G_{nS} ir empirines Edgeworth'o aproksimacijas G_{nSJ} ir G_{nSB} modeliuodami palyginsime tarpusavyje ir su normaliąja aproksimacija Φ. Tegu $\tilde{F}_{nS}$ žymi pasiskirstymo funkcijos F_{nS} Monte–Karlo aproksimaciją, kuri gauta nepriklausomai generuojant 10^6 imčių be grąžinimo iš populiacijos $\mathcal{X}$. Žemiau esančiose lentelėse pasiskirstymai $\tilde{F}_{nS}$, Φ, G_{nS}, G_{nSJ}, G_{nSB} palyginami imant jų q-kvantilius, $q = 0.01, 0.05, 0.10, 0.90, 0.95, 0.99$. Tiksliau, empirinėms funkcijoms G_{nSJ} ir G_{nSB} pateikiamos dvi kiekvieno empirinio kvantilio charakteristikos: jo vidurkio ir standartinės paklaidos (SP) įverčiai $\widehat{\mathbf{E}}(\cdot)$ ir $\widehat{\mathbf{S}}(\cdot)$, kurie suskaičiuojami naudojant 200 nepriklausomai sudarytų imčių be grąžinimo iš populiacijos $\mathcal{X}$. Be to, pateikiamos parametrų α ir κ reikšmės, bei jų įvertinių $\hat{\alpha}_J$, $\hat{\alpha}_B$ ir $\hat{\kappa}_J$, $\hat{\kappa}_B$ poslinkių (POSL) ir SP įverčiai.

13 Modeliavimas. Nagrinėkime Gini'o vidutinių skirtumų statistiką U_G, žr. 3 pavyzdį. Dydžio $N = 150$ populiaciją $\mathcal{X}$ sudarykime generuodami normaliojo skirstinio $\mathcal{N}(2, 4)$ reikšmes. Sudarytos populiacijos $\mathcal{X}$ reikšmių vidurkis ir dispersija yra 2.01 ir 4.03. Imties dydžiu imame $n = 45$.

2 lentelė: $\tilde{F}_{nS}$ aproksimacijos.

$q =$	0.01	0.05	0.10	0.90	0.95	0.99
$\tilde{F}_{nS}^{-1}(q)$	-2.92	-1.94	-1.47	1.18	1.49	2.05
$\Phi^{-1}(q)$	-2.33	-1.64	-1.28	1.28	1.64	2.33
$G_{nS}^{-1}(q)$	-2.68	-1.88	-1.43	1.16	1.45	1.89
$\widehat{\mathbf{E}}\, G_{nSJ}^{-1}(q)$	-2.65	-1.86	-1.42	1.17	1.46	1.93
$\widehat{\mathbf{E}}\, G_{nSB}^{-1}(q)$	-2.60	-1.81	-1.39	1.20	1.50	2.00
$\widehat{\mathbf{S}}\, G_{nSJ}^{-1}(q)$	0.13	0.10	0.08	0.06	0.09	0.18
$\widehat{\mathbf{S}}\, G_{nSB}^{-1}(q)$	0.13	0.10	0.07	0.05	0.08	0.17

3 lentelė: Parametrai α ir κ, ir jų įverčiai.

α	κ		$\hat{\alpha}_J$	$\hat{\alpha}_B$	$\hat{\kappa}_J$	$\hat{\kappa}_B$
2.04	-0.25	POSL	-0.25	-0.25	0.07	-0.07
		SP	0.51	0.50	0.09	0.08

Matosi, žr. 2 lentelę, kad G_{nS} aproksimuoja $\tilde{F}_{nS}$ žymiai tiksliau negu Φ. Pastebėkime, kad G_{nSB}, kaip G_{nS} įvertis, yra labiau paslinktas negu G_{nSJ}. Tai reiškia, kad parametro κ visrakčio įvertis yra labiau nusisekęs negu atitinkamas savirankos įvertis, žr. 3 lentelę, nors abiejų κ įverčių kokybė yra beveik tokia pati. Bet kuriuo atveju, abu G_{nSJ} ir G_{nSB} atrodo veiksmingesni už Φ.

14 Modeliavimas. Nagrinėkime nupjautąjį vidurkį $M_{0.2;0.8}$, žr. 1 pavyzdį. Dydžio $N = 150$ populiaciją $\mathcal{X}$ sudarykime generuodami eksponentinio skirstinio $\mathcal{E}(0.5)$ reikšmes. Sudarytos populiacijos $\mathcal{X}$ reikšmių vidurkis ir dispersija yra 1.99 ir 3.9. Imties dydžiu imame $n = 45$.

4 lentelė: $\tilde{F}_{nS}$ aproksimacijos.

$q =$	0.01	0.05	0.10	0.90	0.95	0.99
$\tilde{F}_{nS}^{-1}(q)$	-2.88	-1.98	-1.50	1.17	1.50	2.13
$\Phi^{-1}(q)$	-2.33	-1.64	-1.28	1.28	1.64	2.33
$G_{nS}^{-1}(q)$	-2.65	-1.87	-1.45	1.14	1.43	1.90
$\widehat{\mathbf{E}}\, G_{nSJ}^{-1}(q)$	-2.59	-1.84	-1.43	1.15	1.46	1.98
$\widehat{\mathbf{E}}\, G_{nSB}^{-1}(q)$	-2.65	-1.87	-1.45	1.13	1.43	1.90
$\widehat{\mathbf{S}}\, G_{nSJ}^{-1}(q)$	0.20	0.19	0.18	0.15	0.20	0.32
$\widehat{\mathbf{S}}\, G_{nSB}^{-1}(q)$	0.11	0.09	0.07	0.06	0.08	0.17

5 lentelė: Parametrai α ir κ, ir jų įverčiai.

α	κ		$\hat{\alpha}_J$	$\hat{\alpha}_B$	$\hat{\kappa}_J$	$\hat{\kappa}_B$
0.34	0.53	POSL	-0.04	0.03	0.01	0.01
		SP	0.17	0.16	1.04	0.24

Matome, kad G_{nS} patikslina Φ, žr. 4 lentelę. Aproksimacija G_{nSJ}, kitaip nei G_{nSB}, yra paslinktasis G_{nS} įvertis. Be to, jos SP yra didesnė. Tai paaiškinama sąlyginai didele įverčio $\hat{\kappa}_J$ SP, žr. 5 lentelę. Taigi, atsižvelgiant į tai, kiek priklausomai nuo imties gali kisti empirinės aproksimacijos, G_{nSB} yra labiau naudinga negu Φ ir G_{nSJ}.

L-statistikos, nagrinėjamos 13 ir 14 modeliavimuose, pasižymi skirtingomis glodumo savybėmis, t.y., skiriasi jų svorio funkcijų glodumas. Glodžiai 13 modeliavimo L-statistikai turime, kad Edgeworth'o skleidinys ir empirinis Edgeworth'o skleidinys su parametrų visrakčio įvertiniais yra veiksmingi. 14 modeliavimo statistika nėra glodi. Šiuo atveju Edgeworth'o skleidinys taip pat naudingesnis už normaliąją aproksimaciją. Tačiau empirinių Edgeworth'o skleidinių kokybė čia jau kitokia. Iš tikrųjų, parametro κ visrakčio įvertis nėra itin kokybiškas. Tuo tarpu empirinė Edgeworth'o aproksimacija su parametrų saviranko̧s įvertiniais, turėjusi poslinkį 13 modeliavime, dabar yra nepaslinkta ir stabili, t.y., efektyvi 14 modeliavime.

Galima paaiškinti, kodėl nupjautojo vidurkio atveju empirinė Edgeworth'o aproksimacija su parametrų visrakčio įvertiniais nėra tokia sėkminga. Tą gali lemti mažiausiai dvi priežastys, kurios yra gana gerai žinomos n.v.p. stebėjimų atveju, tačiau nėra pakankamai gerai ištirtos negrąžintinėms imtims. Pirmoji priežastis: tais atvejais, kai L-statistikos nėra glodžios (arba yra nepakankamai glodžios), Edgeworth'o aproksimacijų tikslumui užtikrinti prireikia papildomų populiacijos pasiskirstymo funkcijos glodumo sąlygų, žr., pavyzdžiui, [18, 1]. Net ir laikant, kad baigtinė populiacija yra gauta iš kažkokios glodžios superpopuliacijos, ji visai nebūtinai yra pakankamai glodi, jei jos dydis N yra nelabai didelis. Antroji priežastis: klasikinis statistikos dispersijos visrakčio įvertinys gali būti nesuderintas, jeigu L-statistika nėra pakankamai glodi, žr., pavyzdžiui, [29].

Disertacijoje pateikiama ir daugiau kompiuterinio modeliavimo pavyzdžių, kurie dar tiksliau atskleidžia ypatingai empirinių aproksimacijų privalumų ir trūkumų priklausomybę nuo L-statistikos svorio funkcijos $J\colon (0,1) \to \mathbb{R}$ glodumo.

Kvantilio skirstinio aproksimacijos sluoksninėms imtims. Sakykime, kad populiacija $\mathcal{X} = \{x_1, \ldots, x_N\}$ yra suskaidyta į $h \geq 1$ bendrų elementų neturinčių sluoksnių: $\mathcal{X} = \mathcal{X}_1 \cup \cdots \cup \mathcal{X}_h$, kur $\mathcal{X}_k = \{x_{k,1}, \ldots, x_{k,N_k}\}$, $1 \leq k \leq h$. Aišku, kad $N = N_1 + \cdots + N_h$. Tegu $\mathbb{X}_k = \{X_{k,1}, \ldots, X_{k,n_k}\}$ žymi dydžio $n_k < N_k$ paprastąją atsitiktinę imtį be grąžinimo iš sluoksnio $\mathcal{X}_k$. Laikome, kad imtys $\mathbb{X}_1, \ldots, \mathbb{X}_h$ yra nepriklausomos. Pažymėkime

$\mathbb{X}' = \mathbb{X}_1 \cup \cdots \cup \mathbb{X}_h$ ir $n = n_1 + \cdots + n_h$. Tegu

$$F_{N,k}(x) = \frac{1}{N_k} \sum_{i=1}^{N_k} \mathbb{I}\{x_{k,i} \le x\} \quad \text{ir} \quad \widehat{F}_{n,k}(x) = \frac{1}{n_k} \sum_{i=1}^{n_k} \mathbb{I}\{X_{k,i} \le x\}$$

žymi k-ojo sluoksnio pasiskirstymo funkciją ir jos empirinį atitikmenį. Tada populiacijos $\mathcal{X}$ pasiskirstymo funkcija ir imtimi $\mathbb{X}'$ paremtas jos įvertinys yra

$$F(x) = \sum_{k=1}^{h} \frac{N_k}{N} F_{N,k}(x) \quad \text{ir} \quad \widehat{F}_n(x) = \sum_{k=1}^{h} \frac{N_k}{N} \widehat{F}_{n,k}(x).$$

Nagrinėkime populiacijos β-kvantilį, $0 < \beta < 1$, kuris apibrėžiamas taip: $F^{-1}(\beta) = \inf\{x : F(x) \ge \beta\}$. Apibrėžkime jo įvertinį

$$X_\beta = \widehat{F}_n^{-1}(\beta) = \inf\{x : \widehat{F}_n(x) \ge \beta\}.$$

Pažymėkime $\sigma_\beta^2 = \mathbf{Var}\, X_\beta$. Disertacijoje konstruojame pasiskirstymo funkcijos

$$F_\beta(x) = \mathbf{P}\{X_\beta - \mathbf{E}\, X_\beta \le x\sigma_\beta\} \tag{25}$$

Edgeworth'o tipo aproksimacijas, siekdami patikslinti normaliąją aproksimaciją, kuri sluoksninių imčių be grąžinimo atveju yra nagrinėta [28, 19] darbuose. Panašiai kaip ir vienasluoksnės imties atveju, pagrindinė priemonė tokių aproksimacijų sudarymui yra simetrinių statistikų Hoeffding'o skleidinys $X_\beta = \mathbf{E}\, X_\beta + L + Q + R$, nagrinėtas [7] darbe sluoksninėms imtims be grąžinimo. Čia L ir Q vadinamos tiesine ir kvadratine skleidinio dalimis, o R yra liekamasis narys. U-statistikų atveju, kai $R \equiv 0$, [10] darbe sukonstruotas trumpasis Edgeworth'o skleidinys

$$H_\beta(x) = \Phi(x) - \frac{\alpha_\beta + 3\kappa_\beta}{6\sigma_\beta^3} \Phi'(x)(x^2 - 1). \tag{26}$$

Laikydami, kad R dispersija sąlyginai maža, šią aproksimaciją naudojame ir sluoksninės imties kvantilio atveju. Disertacijoje pateikiame išreikštines parametrų α_β, κ_β ir σ_β^2 formules. Kadangi šie parametrai priklauso nuo visos populiacijos reikšmių, tai disertacijoje taip pat pateikiame jų [12] saviraankos įvertinius, vertinamus Monte–Karlo metodu. Taip gauname (25) pasiskirstymo funkcijos (26) aproksimacijos empirinį atitikmenį. Abi šias $F_\beta(x)$ aproksimacijas palyginame su normaliąja aproksimacija kompiuterinio modeliavimo būdu sluoksninės imties medianos atveju.

Rezultatų apie Edgeworth'o skleidinį apibendrinimai sluoksninėms imtims nėra paprasti. Pavyzdžiui, (26) aproksimacijos parametras κ_β atspindi ir tam tikras tarpsluoksnines sąveikas, žr. [10].

3 Apibendrinimas

3.1 Ginamos išvados

- Sukonstruotas trumpasis Edgeworth'o skleidinys.
- Gautas optimalus imties ekstremaliųjų reikšmių dispersijų viršutinysis įvertis.
- L-statistikų asimptotinį normalumą ir tam tikrus Edgeworth'o aproksimacijos tikslumus užtikrinančios sąlygos nusakytos statistiką apibrėžiančių svorių glodumu ir populiacijos momentų aprėžtumu. Asimptotinis normalumas ištirtas ir neglodžių svorių (nupjautojo vidurkio) atveju.
- Sudarytos tikslios L-statistikos dispersijos ir Edgeworth'o skleidinio parametrų savirankos įvertinių formulės. Tokiu būdu pašalinta papildoma paklaida, kuri atsiranda vertinant savirankos įvertinius.
- Kompiuterinio modeliavimo pavyzdžiais parodyta, kaip Edgeworth'o aproksimacijos tikslumas priklauso nuo statistikos glodumo savybių. Taip pat parodyta, kad empirinis Edgeworth'o skleidinys su parametrų savirankos įvertiniais yra pranašesnis už skleidinį su parametrų visrakčio įvertiniais tuo atveju, kai L-statistikos svoriai nėra pakankamai glodūs.

3.2 Rezultatų pristatymas

Disertacijos rezultatai buvo pristatyti šiose konferencijose:

1. Tarptautinėje konferencijoje, surengtoje Norrfällsviken'e, Švedijoje pavadinimu "The Third Baltic-Nordic Conference on Survey Statistics", 2011 m. birželio 13–17 dienomis. Pranešimo tema: On an optimal bound for the variance of sample maximum.

2. LII Lietuvos matematikų draugijos konferencijoje, surengtoje Vilniuje 2011 m. birželio 16–17 dienomis. Pranešimo tema: Approximations to distribution of median in stratified samples. (Su T. Rudžiu)

3. Tarptautiniame seminare, surengtame Vilniuje pavadinimu "Workshop on Survey Sampling Theory and Methodology", 2010 m. rugpjūčio 23–27 dienomis. Pranešimo tema: Bootstrap for variance of finite population L-statistic.

4. Tarptautinėje konferencijoje, surengtoje Vilniuje pavadinimu "10th International Vilnius Conference on Probability and Mathematical Statistics", 2010 m. birželio 28 – liepos 2 dienomis. Pranešimo tema: An Edgeworth expansion for finite population L-statistics.

5. LI Lietuvos matematikų draugijos konferencijoje, surengtoje Šiauliuose 2010 m. birželio 17–18 dienomis. Pranešimo tema: Bootstrap, jackknife and Edgeworth approximations for finite population L-statistics.

6. Tarptautinėje vasaros mokykloje, surengtoje Kyiv'e, Ukrainoje pavadinimu "The Baltic-Nordic-Ukrainian Summer School on Survey Statistics", 2009 m. rugpjūčio 23–27 dienomis. Pranešimo tema: Orthogonal decomposition of finite population L-statistics.

7. L Lietuvos matematikų draugijos konferencijoje, surengtoje Vilniuje 2009 m. birželio 18–19 dienomis. Pranešimo tema: Orthogonal decomposition of finite population L-statistics.

3.3 Publikacijų sąrašas

Disertacijos rezultatai yra publikuoti šiuose leidiniuose:

1. A. Čiginas. Second-order approximations of finite population L-statistics. *Statistics*, 2011. (įteiktas)

2. A. Čiginas. An Edgeworth expansion for finite population L-statistics. *Lith. Math. J.*, 2011. (priimtas); taip pat žr. arXiv:1103.4220v2 [math.ST].

3. A. Čiginas. An exact bootstrap for variance of finite-population L-statistic. *Lith. Math. J.*, 51:322–329, 2011.

4. A. Čiginas and T. Rudys. Approximations to distribution of median in stratified samples. *Liet. Mat. Rink. LMD darbai*, 52, 2011. (priimtas)

5. A. Čiginas. Bootstrap, jackknife and Edgeworth approximations for finite population L-statistics. *Liet. Mat. Rink. LMD darbai*, 51:391–396, 2010.

6. A. Čiginas. Orthogonal decomposition of finite population L-statistics. *Liet. Mat. Rink. LMD darbai*, 50:287–292, 2009.

3.4 Summary

Approximations to distributions of L-statistics, where samples are drawn without replacement, are considered in the thesis. The asymptotic properties of distributions are explored using Hoeffding's decomposition of the statistics. In the first chapter of the thesis, explicit expressions of the first three terms and remainder terms of the Hoeffding decomposition are obtained. The main applications of the decomposition are given in the second chapter. First, a new upper bound for the variance of the sample minimum and maximum is presented, which is optimal in the form provided. Similar bounds are shown for the other order statistics. Next, the asymptotic normality and validity of one-term Edgeworth expansion are considered. In addition to the asymptotic normality of L-statistics of more general form, the case of the trimmed means is also analyzed. In the third chapter, estimation of the variance and parameters that define the Edgeworth expansion of an L-statistic is considered. Two competitive methods are used: the classical jackknife and certain finite population bootstrap. In the case of bootstrap, exact formulas of the estimators of all the parameters of interest are obtained. In the fourth chapter, several practical variants of the Edgeworth expansion are considered: an Edgeworth expansion for a Studentized L-statistic, empirical Edgeworth expansions, and (in a sense related) non-parametric bootstrap approximations. Their second-order correctness is discussed, and their efficiencies for various L-statistics in a simulation study are compared. A generalization of one-term Edgeworth expansions to the case of stratified simple random samples drawn without replacement, where the L-statistics are quantiles of the stratified sample, is also considered. An explicit expression of the approximation to distribution of the quantile as well as its empirical version based on bootstrap are presented.

3.5 Trumpos žinios apie autorių

Gimimo data ir vieta

1982 m. vasario 4 d., Vilnius.

Išsilavinimas

2000 - 2011 Vilniaus universitetas, Matematikos ir informatikos fakultetas:

- 2006 - 2011 Doktorantūros studijos,
- 2006 Matematikos magistro diplomas,
- 2004 Matematikos bakalauro diplomas.

Darbo patirtis

- 2004 - iki dabar, Metodologijos ir kokybės skyriaus vyriausiasis specialistas, Lietuvos statistikos departamentas.
- 2006 - 2010, Matematinės informatikos katedros asistentas, Vilniaus universitetas.

Literatūra

[1] I. Alberink, G. Pap, and M. van Zuijlen. Edgeworth expansions for L-statistics. *Prob. Math. Statist.*, 21:277–302, 2001.

[2] B. Arnold, N. Balakrishnan, and H. Nagaraja. *A First Course in Order Statistics*. SIAM, Philadelphia, 2008.

[3] N. Balakrishnan, C. Charalambides, and N. Papadatos. Bounds on expectation of order statistics from a finite population. *J. Statist. Plann. Inference*, 113:569–588, 2003.

[4] V. Bentkus, F. Götze, and W. van Zwet. An Edgeworth expansion for symmetric statistics. *Ann. Statist.*, 25:851–896, 1997.

[5] M. Bloznelis. One- and two-term Edgeworth expansions for a finite population sample mean. Exact results. I,II. *Lith. Math. J.*, 40:213–227, 329–340, 2000.

[6] M. Bloznelis. Empirical Edgeworth expansion for finite population statistics I,II. *Lith. Math. J.*, 41:120–134, 207–218, 2001.

[7] M. Bloznelis. A note on the bias and consistency of the jackknife variance estimator in stratified samples. *Statistics*, 37:489–504, 2003.

[8] M. Bloznelis. Edgeworth expansions for Studentized versions of symmetric finite population statistics. *Lith. Math. J.*, 43:221–240, 2003.

[9] M. Bloznelis. Bootstrap approximation to distributions of finite population U-statistics. *Acta Appl. Math.*, 96:71–86, 2007.

[10] M. Bloznelis. Second-order and resampling approximation of finite population U-statistics based on stratified samples. *Statistics*, 41:321–332, 2007.

[11] M. Bloznelis and F. Götze. Orthogonal decomposition of finite population statistics and its applications to distributional asymptotics. *Ann. Statist.*, 29:899–917, 2001.

[12] J. Booth, R. Butler, and P. Hall. Bootstrap methods for finite populations. *J. Amer. Statist. Assoc.*, 89:1282–1289, 1994.

[13] A. Chatterjee. Asymptotic properties of sample quantiles from a finite population. *Ann. Inst. Statist. Math.*, 63:157–179, 2011.

[14] H. Chernoff, J. Gastwirth, and M. J. Jr. Asymptotic distribution of linear combinations of functions of order statistics with applications to estimation. *Ann. Math. Statist.*, 38:52–72, 1967.

[15] H. Cramer. *Mathematical Methods of Statistics.* Princeton University Press, Princeton, NJ, 1946.

[16] F. Edgeworth. The law of error. *Trans. Cambridge Phil. Soc.*, 20:36–65, 113–141, 1905.

[17] P. Erdős and A. Rényi. On the central limit theorem for samples from a finite population. *Publ. Math. Inst. Hungar. Acad. Sci.*, 4:49–61, 1959.

[18] N. Gribkova and R. Helmers. The empirical Edgeworth expansion for a Studentized trimmed mean. *Math. Methods Statist.*, 15:61–87, 2006.

[19] S. Gross. Median estimation in sample surveys. In *Proceedings of the Section on Survey Research Methods*, pages 181–184. American Statistical Association, 1980.

[20] R. Helmers. Edgeworth expansions for linear combinations of order statistics with smooth weight functions. *Ann. Statist.*, 8:1361–1374, 1980.

[21] A. Hutson and M. Ernst. The exact bootstrap mean and variance of an L-estimator. *J. Roy. Statist. Soc. B*, 62:89–94, 2000.

[22] Y. Maesono. An Edgeworth expansion and a normalizing transformation for L-statistics. *Bull. Inform. Cybernet.*, 39:25–43, 2007.

[23] S. Moriguti. Extremal properties of extreme value distributions. *Ann. Math. Statist.*, 22:523–536, 1951.

[24] N. Papadatos. Maximum variance of order statistics. *Ann. Inst. Statist. Math.*, 47:185–193, 1995.

[25] N. Papadatos. A note on maximum variance of order statistics from symmetric populations. *Ann. Inst. Statist. Math.*, 49:117–121, 1997.

[26] H. Putter and W. van Zwet. Empirical Edgeworth expansions for symmetric statistics. *Ann. Statist.*, 26:1540–1569, 1998.

[27] T. Rychlik. Extreme variances of order statistics in dependent samples. *Stat. Probab. Lett.*, 78:1577–1582, 2008.

[28] J. Shao. L-statistics in complex survey problems. *Ann. Statist.*, 22:946–967, 1994.

[29] J. Shao and C. Wu. A general theory for jackknife variance estimation. *Ann. Statist.*, 17:1176–1197, 1989.

[30] G. Shorack. Functions of order statistics. *Ann. Math. Statist.*, 43:412–427, 1972.

[31] S. Stigler. Linear functions of order statistics with smooth weight functions. *Ann. Statist.*, 2:676–693, 1974.