Liet. matem. rink., T. 42, spec. nr., 529–536

Pasikliautinųjų intervalų skaičiavimo būdų palyginimas diskretiems biologiniams duomenims

Alesia KANOPAITĖ (VGTU), Eduardas MICKEVIČIUS (Ekologijos institutas), Marijus RADAVIČIUS (VGTU)
el. paštas: mrad@ktl.mii.lt

1. Įvadas

Taikymuose dažnai tenka skaičiuoti pasikliautinuosius intervalus (PI). Skirtingai nuo paprasto stebėjimų aritmetinio vidurkio, aprašančio „tipinę" tiriamojo kintamojo reikšmę, PI atspindi ir tos reikšmės tikslumą bei turimos informacijos patikimumą. Standartiniuose statistiniuse paketuose realizuotos PI skaičiavimo procedūros skaičiuoja taip vadinamą Stjudento pasikliautinąjį intervalą (SPI). Šis metodas remiasi prielaida, kad tiriamosios populiacijos skirstinys yra normalusis, be šios prielaidos SPI yra tik tikrojo PI aproksimacija. Šiuo metu yra žinoma daug kitų, daug tikslesnių PI aproksimavimo metodų, kurie remiasi Edžvorto (Edgeworth) skleidiniais ar didžiųjų nuokrypių tikimybių asimptotika, tačiau realiuose statistiniuose tyrimuose jie nėra plačiai taikomi. To priežastis yra sunkiai praktiškai patikrinamos prielaidos, kuriomis jie remiasi, bei asimptotinis (kaip ir SPI metodo) pobūdis, neleidžiantis daryti apie gautą PI pagrįstų išvadų realybėje visada baigtinio, o dažnai ir gana mažo dydžio imtimi.

Ryšium su vis platesniu statistinių metodų taikymu vis aktualesnis darosi jų patikimumas ir universalumas. Todėl pastaruoju metu vis daugiau dėmesio skiriama praktiniams PI apskaičiavimo klausimams ir įvairių PI skaičiavimo metodų palyginimui baigtinio dydžio imtim analitiškai ir modeliavimo būdu [1, 3].

Šiame darbe nagrinėjami PI sudaryti, naudojant pakartotinų imčių (bootstrap, resampling) metodus. Modeliavimo būdu atliktas jų palyginimas tarpusavyje ir su SPI. Tyrimas remiasi realiais biologiniais duomenimis (pateiktais antrojo iš autorių) apie urvinių plėšrūnų (barsukų, lapių ir mangutų) urvų pasiskirstymą 1 km^2 ploto kvadratuose tolygiai išdėstytuose visoje Lietuvos teritorijoje. Pasiskirstymai ypatingi tuo, kad yra diskretūs, įgyja tik keletą reikšmių (iki 9), yra labai asimetriški su santykinai dideliu nulinės reikšmės dažnumu ir nereguliariomis „uodegomis" (žiūr. 2 skyrelį). Žodžiu, turi pakankamai nepalankių ypatybių, kad keltų abejones, jog SPI šiuo atveju, kai imčių dydžiai svyruoja tarp 3 ir 187, yra geriausias PI skaičiavimo metodas.

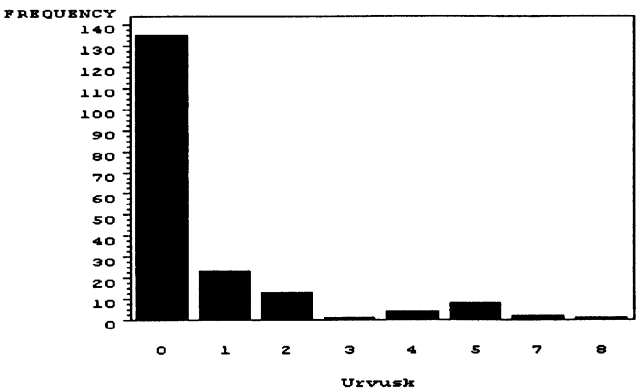
2-ame skyrelyje aptariami duomenys ir konkretus PI radimo uždavinys. Toliau trumpai supažindinama su pakartotinų (pseudo)imčių metodologija, paprasčiausiais ja pagrįstais PI konstravimo būdais bei papildomos informacijos panaudojimo galimybėmis. Paskutinis skyrelis skirtas tyrimo metodikos aprašymui, rezultatams ir išvadoms.

2. Duomenų aptarimas

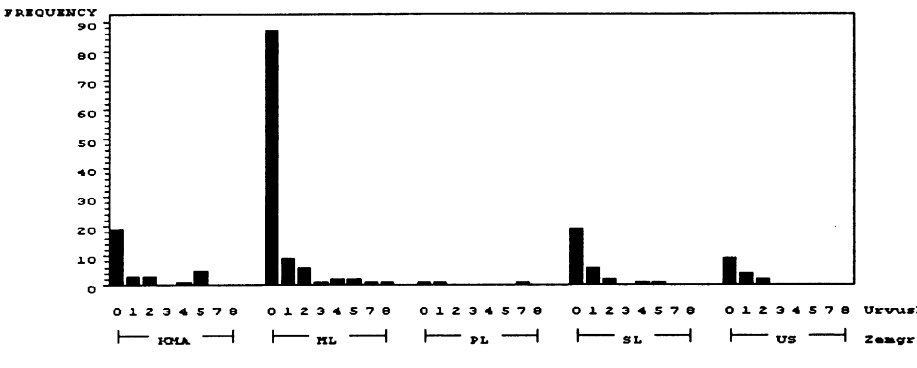
Naudojant kartografinius žemėlapius (1:25 000) buvo išrinkti 371 1 km^2 ploto kvadratai. Jie skirstomi į dvi grupes: „miško“ (jų 184) ir „lauko“ (jų 187). Šiame darbe naudojami tik „lauko“ kvadratai, nes jie tolygiai šachmatine tvarka išsidėstę Lietuvos teritorijoje (sisteminė imtis) ir todėl leidžia be poslinkio prognozuoti urvų skaičių visoje Lietuvos teritorijoje ir įvairiuose sluoksniuose. Kadangi kvadratų numeracija yra atsitiktinė atžvilgiu jų gamtinių sąlygų ir imtis sudaro tik 0,25% viso ploto, galima laikyti, kad duomenys apie kvadratus yra paprastoji grąžintinė imtis. Duomenyse užregistruotas urvų skaičius ir gamtinės sąlygos kiekviename kvadrate.

Urvinių plėšrūnų (barsukų, lapių, mangutų) urvų vidutinio skaičiaus kvadrate pasikliautinieji intervalai leistų daryti tam tikras išvadas apie jų skaitlingumą ir paplitimą. Biologams svarbios ir jų reikšmės atskirose žemėvaizdžio grupėse: KMA (kalvotosios moreninės aukštumos), ML (moreninės lygumos), PL (plynaukštės), SL (smėlėtosios lygumos), US (upių slėniai).

1 ir 2 pav. pateiktos urvų dažnių diagramos rodo, kad skirstiniai yra labai asimetriški dažnai su santykinai dideliu nulinės reikšmės dažnumu ir nereguliariomis „uodegomis“.



1 pav. Urvų pasiskirstymo visoje teritorijoje dažniai.



2 pav. Urvų pasiskirstymo įvairiuose žemėvaizdžiuose dažniai.

1 lentelė ·
Stjudento pasikliautinieji intervalai

Zemgr	*N* Obs	Mean	Lower 95% CL for Mean	Upper 95% CL for Mean
KMA	31	1,2258065	0,5252885	1,9263244
ML	109	0,5229358	0,2599673	0,7859043
PL	3	2,6666667	-6,7381269	12,0714603
SL	29	0,6551724	0,1862545	1,1240904
US	15	0,5333333	0,1217500	0,9449167

Jiems suskaičiuoti SPI pateikti 1-oje lentelėje. Koks jų tikslumas ir patikimumas?

Įvairiems atvejams, tame tarpe ir atvaizduotiems 1 ir 2 pav., Monte Karlo metodu buvo generuota tūkstantis imčių ir joms paskaičiuoti vidurkiai. Visais atvejais Shapiro-Wilk kriterijus atmetė hipotezę apie tų vidurkių skirstinio normališkumą (rezultatai čia nepateikiami, juos galima rasti [4]).

3. Pasikliautinųjų intervalų apskaičiavimas pakartotinų imčių metodu

Tegu $Y_1, \ldots, Y_n$ yra dydžio n paprasčiausia imtis su pasiskirstymo funkcija (p.f.) F ir tegu $T(Y_1, \ldots, Y_n)$ yra statistika, nepriklausanti nuo imties elementų tvarkos. Ją patogu užrašyti tokiu pavidalu $T(\widehat{F})$, kur $\widehat{F}$ yra empirinė p.f. (e.p.f.). Konkrečiai imčiai $y_1, \ldots, y_n$ galima suskaičiuoti statistikos T reikšmę, bet be papildomų (parametrinių) prielaidų apie F negalima daryti jokių išvadų apie galimas šios statistikos reikšmes pakartotinoms dydžio n imtims, t.y., apie jos tikimybinį skirstinį. Tam reiktų daug kart kartoti imties rinkimo procedūrą. Pakartotinų imčių (bootstrap) metodologija siūlo tokią išeitį iš šios situacijos.

Jeigu p.f. G mažai skiriasi nuo p.f. F_0, tai natūralu tikėtis, kad ir statistikos $T(\widehat{F})$ skirstinys tuo atveju, kai $F = G$, tam tikra prasme mažai skirsis nuo atvejo, kai $F = F_0$. Jeigu n pakankamai didelis, tai $\widehat{F} \approx F$. Paėmę $F = \widehat{F}$ turime galimybę kiek norima kartų kartoti imties rinkimo procedūrą realaus rinkimo neatliekant, o jį tik imituojant, pvz., kompiuterio pagalba (Monte Karlo metodas). Tam reikia iš aibės $\{Y_1, \ldots, Y_n\}$ atsitiktinai su lygiomis kiekvieno elemento išrinkimo tikimybėmis išrinkti n elementų. Tokias imtis vadinsime (neparametrinėmis) pakartotinomis (pseudo)imtimis (bootstrap imtimis) ir žymėsime $Y_1^*, \ldots, Y_n^*$, o jų e.p.f. $\widehat{F}^*$. Turint R pakartotinų pseudoimčių, o tuo pačiu ir R statistikos $T^* = T(\widehat{F}^*)$ reikšmių $T_1^*, \ldots, T_R^*$, galima įvertinti $T(\widehat{F})$ skirstinį, vidurkį, dispersiją, kvantilius ir kitas tikimybines charakteristikas. Tegu $c = c(T|F)$ žymi mus dominančią charakteristiką.

Pakartotinų imčių metodo [2] paklaidą sudaro *statistinė* paklaida ir *modeliavimo* paklaida. Pirmoji atsiranda dėl skirtumo tarp p.f. F ir $\widehat{F}$, kuris sąlygoja ir skirtumą tarp $c(T|F)$ ir $c(T|\widehat{F})$. Paprastai yra tam tikra laisvė pasirenkant c ir statistiką T. Tuomet siekiama jas parinkti taip, kad $c(T|F)$ kuo mažiau kistų atžvilgiu F, t.y., turėtų kuo mažesnę absoliutiniu dydžiu įtakos funkciją (influence function) tikrosios p.f. F aplinkoje.

Tam naudojama standartizacija (studetization), transformacijos [2] bei papildomos statistikos [3].

Modeliavimo paklaida atsiranda dėl to, kad vietoje $c(T|\widehat{F})$ imamas jo įvertinys gautas iš imties $T_1^*, \ldots, T_R^*$. Su pakankamai dideliu R ši paklaida gali būti kaip norint maža, tačiau beprasmiška siekti žymiai didesnio tikslumo už statistinę paklaidą. Monografijoje [2], pavyzdžiui, 0,025 kvantilio vertinimui, rekomenduojama imti $R = 40n$.

Sudarant PI, pakartotinų (pseudo)imčių metodologija taikoma reikalingų kvantilių įvertinimui. Duotam p, $0 < p < 1$, p-kvantilis įvertinamas k-ąja imties $T_1^*, \ldots, T_R^*$ pozicine statistika $T_{(k)}^*$, $k = k(p) = [(R+1)p]$. Tegu $\alpha_0 = 2\alpha$ yra duotas reikšmingumo lygis, $\widehat{\theta}_\alpha$ ir $\widehat{\theta}_{1-\alpha}$ žymi statistikos T atitinkamai apatinį ir viršutinį PI rėžius, $l = k(\alpha)$, $u = k(1-\alpha)$. Bazinis (basic bootstrap) PI nusakomas formulėmis

$$\widehat{\theta}_\alpha = 2t - t_{(u)}^*, \quad \widehat{\theta}_{1-\alpha} = 2t - t_{(l)}^*.$$

Standartizuoto bazinio (studentized basic bootstrap) PI rėžiai yra

$$\widehat{\theta}_\alpha = t - sZ_{(u)}^*, \quad \widehat{\theta}_{1-\alpha} = t - sZ_{(l)}^*,$$

kur s^2 yra T dispersijos įvertinys iš pradinės imties, $Z_{(k)}^*$ žymi imties

$$Z_r^* = (T_r^* - T)/s_r^*, \quad r = 1, \ldots, R,$$

k-ąją pozicinę statistiką, $(s_r^*)^2$ žymi statistikos T dispersijos įvertinį iš r-tos pakartotinos pseudoimties $Y_1^*, \ldots, Y_n^*$. Bazinis PI su (simetrizuojančia) transformacija h ($h = h$ (basic bootstrap)) užsirašo tokiu būdu

$$\widehat{\theta}_\alpha = h^{-1}\Big\{2h(t) - h\big(t_{(u)}^*\big)\Big\}, \quad \widehat{\theta}_{1-\alpha} = h^{-1}\Big\{2h(t) - h\big(t_{(l)}^*\big)\Big\}.$$

Nors transformacija gali žymiai pagerinti bazinį pasikliautinąjį intervalą, po jos kartais gali būti naudinga panaudoti dar ir standartizaciją.

Procentilių (procentile bootstrap) PI sudarymo metodas, kaip ir aukščiau aprašytasis, remiasi prielaida, kad egzistuoja simetrizuojanti transformacija. Tačiau šiuo atveju galutinėse formulėse ji nefigūruoja, ir todėl jos nereikia žinoti. Procentilių PI rėžiai tokie:

$$\widehat{\theta}_\alpha = t_{(l)}^*, \quad \widehat{\theta}_{1-\alpha} = t_{(u)}^*.$$

Juos taip pat rekomenduojama standartizuoti.

Šiame darbe mus dominanti statistika $T = T(\widehat{F})$ yra tiesiog aritmetinis vidurkis $\bar{Y}$.

Papildomos informacijos panaudojimas sudarant pasikliautinuosius intervalus. Kai stebėjimų skirstiniams aprašyti taikomi parametriniai modeliai, gana dažnai tų parametrų būna ne vienas, o keli. Trumpai aptarsime, kaip galima būtų konstruoti pasikliautinuosius intervalus, naudojant šią papildomą informaciją. Analogiškas metodas aptariamas [3].

Tegu modelio parametrai yra (a, b), kur a yra skaliaras. Tarkime, kad $(\hat{a}, \hat{b})$ yra koks nors parametrų (a, b) įvertinys, pvz., didžiausio tikėtinumo, ir sąlyginė statistikos $\hat{a}$ pasiskirstymo funkcija, kai $\hat{b}$ įgyja fiksuotą reikšmę b, $F_{\hat{a}|\hat{b}}(t|b)$, yra nežinomo vidurkio μ ir parametro b funkcija, griežtai monotoniška ir tolydi atžvilgiu μ ir t kiekvienai $\hat{b}$ reikšmei b. Nemažinant bendrumo galima teigti, kad ji monotoniškai mažėja. Raide γ žymėsime pasikliovimo lygmenį, $\alpha_1 = (1-\gamma)/2$, $\alpha_2 = (1+\gamma)/2$. Tuomet egzistuoja tokia monotoniškai didėjanti atžvilgiu pirmojo argumento funkcija $g(a, \alpha|b)$, kad $\Pr\{g(\hat{a}, \alpha|b) \leqslant \mu|\hat{b} = b\} = \alpha$, $0 < \alpha < 1$. Remiantis šia tapatybe vidurkio γ-pasikliautinąjį intervalą galima apibrėžti lygybėmis

$$\mu_1 = g\left(\hat{a}, \alpha_2|\hat{b}\right), \quad \mu_2 = g\left(\hat{a}, \alpha_1|\hat{b}\right). \tag{1}$$

Pritaikysime šį PI sudarymo metodą duotai konkrečiai situacijai. Iš 1 ir 2 pav. matosi, kad paprastą parametrinį modelį parinkti nepavyks, kadangi pastebimas neproporcingai didelis kvadratų be urvų skaičius. Todėl natūralu pabandyti kvadratų be urvų skaičių modeliuoti atskirai. Viena iš galimų alternatyvų būtų tokia. Tegu X žymi urvų skaičių kvadrate. Tarkime, kad jį galima išreikšti kaip dviejų nepriklausomų atsitiktinių dydžių Z ir Y sandaugą $X = Z \cdot Y$, kur Z turi binominį skirstinį $B(1, p_0)$, o Y – kokį nors parametrinį skirstinį, pvz. Puasono ar geometrinį. Pastaraisiais dviem atvejais visos aukščiau aprašytos sąlygos išpildytos, ir $(n_0, \bar{Y})$, kur n_0 yra kvadartų be urvų skaičius, sudaro pakankamą statistiką. Šio modelio atitikimas duomenims buvo patikrintas įvairiems atvejams, naudojant tikėtinumo kriterijų χ^2. Aišku, jis nevisada tiko. Tačiau ir be parametrinių prielaidų apie Y skirstinį galima tikėtis, kad $\bar{Y}$ skirstinys bus reguliaresnis už $\bar{X}$. Pritaikius (1), vidurkio $\mu = \mathbf{E}X$ PI randamas iš parametro $a = \mathbf{E}Y$ paskliautinojo intervalo (a_1, a_2) pagal formules:

$$\mu_1 = a_1\,(1 - n_0/n)\,, \quad \mu_2 = a_2\,(1 - n_0/n)\,. \tag{2}$$

4. Pasikliautinųjų intervalų palyginimas

Lyginami metodai. Modeliavimo būdu buvo palyginti tarpusavyje tokie PI metodai: SPI, optimalusis (o), basic bootstrap (bb), percentile bootstrap (pb), studentized basic bootstrap (sbb), studentized percentile bootstrap (spb), h=log (basic bootstrap) (log), h=sqrt (basic bootstrap) (sqrt), basic bootstrap v0 (bb0), studentized basic bootstrap v0 (sbb0), scaled basic bootstrap v0 (scbb0), percentile bootstrap v0 (pb0), studentized percentile bootstrap v0 (spb0), scaled percentile bootstrap v0 (scpb0). PI pagrindinės charakteristikos yra (įvertintas) realus jų pasikliovimo lygmuo ir ilgis.

Optimalusis PI yra Monte Karlo metodu naudojant pakartotinas imtis randamas minimalaus ilgio PI. Jis parenkamas kaip trumpiausias intervalas, apimantis ne mažiau kaip 95% generuotų reikšmių.

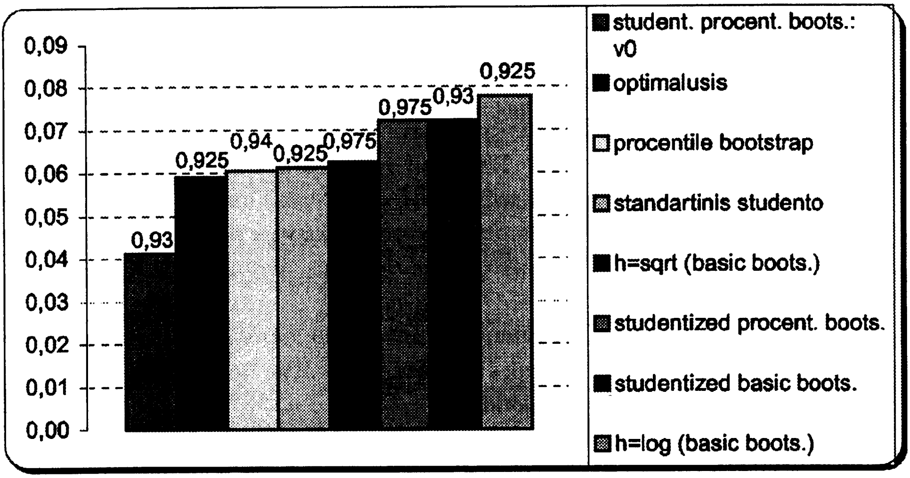
PI (log) ir (sqrt) gauti naudojant "basic bootstrap" metodą transformuotai statistikai, paėmus atitinkamai natūrinį logaritmą ar ištraukus kvadratinę šaknį iš generuotų duomenų vidurkio.

PI (bb0)–(scpb0) yra analogiški aprašytiems anksčiau, bet remiasi ne bendru urvų skaičiaus vidurkiu, o tik kvadratų su urvais vidurkiu v0. Pastarajam atitinkamu metodu apskaičiuoti PI transformuojami į reikiamus naudojant (2). Dviejuose PI su 'sc' (scaled) taikomas specialus standartizacijos būdas. Standartinis nuokrypis įvertinamas $\bar{Y}$ pagalba, remiantis geometrinio skirstinio dispersijos išraiška per vidurkį.

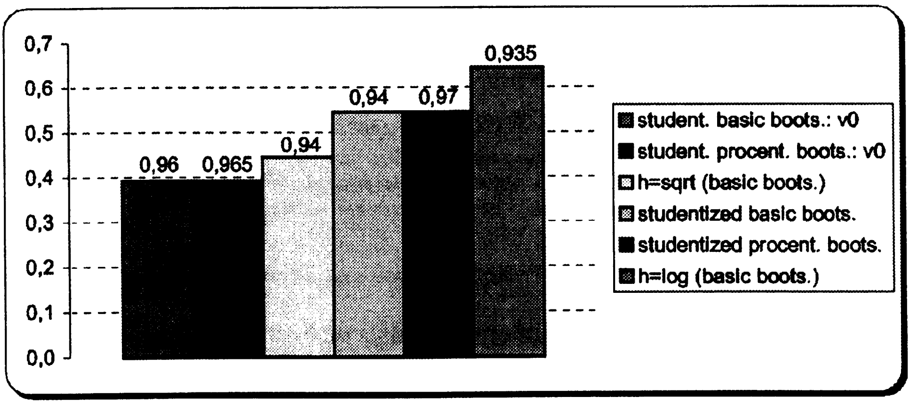
Palyginimo metodika ir rezultatai. PI metodų palyginimui buvo apskaičiuoti urvų skaičiaus tirtuose kvadratuose santykiniai dažniai lapėms, barsukams ir mangutams visoje teritorijoje bei atskirai pagal žemėvaizdžių grupes. Atmetus atvejus su labai mažu imties dydžiu, gavosi 12 gana skirtingų variantų, kuriems visais lyginamaisiais metodais buvo apskaičiuoti PI. Kvantilių, reikalingų PI sudaryti, apskaičiavimui buvo generuojama $R = 1000$ pakartotinų imčių. Tokios yra praktinės rekomendacijos [2]. Siekiant įvertinti PI kokybę ir „realų" pasikliovimo lygmenį, šie pakartotinų imčių eksperimentai buvo atliekami $N = 200$ kartų. Iš šių pakartotinų eksperimentų buvo įvertinti vidutiniai PI ilgiai ir vidutinis PI kraštų atstumas iki generuotos imties vidurkio. Metodai, kurių įvertintas (realus) pasikliautinumo lygmuo mažesnis už 0,95 ne daugiau, kaip leidžia 2σ taisyklė, buvo surūšiuoti vidutinio (per 200 modeliavimo realizacijų) PI ilgio didėjimo tvarka. 1-ajam suteikiami 6 balai, 2-ajam – 5 ir t.t. Balus gaudavo tik pirmasis šešetukas. Šie balai

2 lentelė
Metodų palyginimas

Nr.	Metodas	Įvertinimas												Suma
1	$h = sqrt$ (basic bootstrap)	2		4	2			5	3	2	6		6	30
2	Percentile bootstrap v0		5		6	5	6					6		28
3	Studentized percentile bootstrap v0	6	3	5	3		5			6				28
4	Studentized percentile bootstrap	1		2		2		3	1	1	4	5	4	23
5	Optimalusis	5	1				3		6	5				20
6	Studentized basic bootstrap			3	1			4	2		5		5	20
7	Standartinis Stjudento	3				3		6	4	3				19
8	Percentile bootstrap	4				4	1		5	4				18
9	Basic bootstrap v0		6			6								12
10	Studentized basic bootstrap v0		2	6	4									12
11	Scaled percentile bootstrap v0				5		4							9
12	h=log (basic bootstrap)			1							3		3	7
13	Scaled basic bootstrap v0		4											4
14	Basic bootstrap						2							2



3 pav. Bendro urvų skaičiaus vidurkio pasikliautinieji intervalai.



4 pav. Lapių urvų skaičiaus KMA-ose vidurkio pasikliautinieji intervalai.

ir pateikti 2 lentelėje, stulpelyje „Įvertinimas". Stulpelyje „Suma" yra jų suma per visus 12 variantų.

Iliustravimui 3 ir 4 pav. grafiškai pateikti 2-jų iš 12 atvejų, būtent bendro (visų rūšių visoje teritorijoje) urvų skaičiaus ir lapių urvų skaičiaus kalvotose moreninėse aukštumose, vidurkio PI-ų vidutiniai ilgiai ir įvertinti pasikliovimo lygmenys (jie nurodyti virš atitinkamų stulpelių).

Išvados

1. Nors normalioji aproksimacija beveik visais atvejais buvo netiksli (nulinė hipotezė atmetama), tačiau standartinis Stjudento metodas nagrinėjamoje situacijoje pateikė pakankamai gerus rezultatus. Bendrai pagal vidutinį ilgį jis yra 7-as iš 14, jo įvertintas

pasikliovimo lygmuo buvo ne mažesnis už nominalų 95% pasikliovimo lygmenį. Matyt, tai galima paaiškinti tuo, kad tiriamoje situacijoje atsitiktinius dydžius galima laikyti (beveik) aprėžtais ar net binominiais.

2. Geriausi yra šie metodai: $h = sqrt$ (basic bootstrap), Percentile bootstrap v0 ir Studentized percentile bootstrap v0. Tiek $h = sqrt$ (basic bootstrap), tiek Percentile bootstrap v0 metodų esmę (pastarųjų netiesiogiai) sudaro pasikliautinųjų intervalų konstravimas transformuotai statistikai. Tai reiškia, kad šiuo atveju pasikliautinųjų intervalų kokybė priklauso nuo tinkamai parinktos transformacijos. Vadinasi, tranformacija pasiteisina ir yra reikalinga.

3. Literatūroje [2] rekomenduojama, sudarant PI, naudoti standartizavimą. Nors v0 paremtiems metodams standartizavimas esminės įtakos neturėjo, tačiau klasikinius pakartotinų imčių metodus ženkliai pagerino.

4. Standartizacija, kai vietoj standartinio nuokrypio naudojamas parametriniu modeliu paremtas įvertis (PI (scbb0) ir (scpb0)), nepasiteisino.

Literatūra

[1] L.D. Brown, T.T. Cao, A. DasGupta, Confidence intervals for a binomial proportion and asymptotic expansions, *Ann. Statist.*, **30**(1), 160–201.

[2] A.C. Davison, D.V. Hinkley, *Bootstrap Methods and their Application*, Cambridge University Press (1998).

[3] P. Kabaila, A large sample approximation to the profile plug-in upper confidence limit, *8th Vilnius Conference Theory and Mathematical Statistics. Abstracts of Communications*, TEV, Vilnius (2002).

[4] A. Kanopaitė, *Urvinių žinduolių teritorinio pasiskirstymo statistinė analizė*, Magistrinis darbas, Vilniaus Gedimino technikos universitetas (2002).

Comparison of confidence interval construction methods for discrete biological data

A. Kanopaitė, E. Mickevičius, M. Radavičius

The bootstrap (resampling) methods of construction of confidence intervals (c.i.) are considered. The c.i. obtained are compared with each other and with the standard student c.i. by simulation. The study is based on a real biological data about distribution of badger, fox, and raccoon dog borrow number in 1 km^2 square areas uniformly allocated throughout Lithuania territory. The distributions take only a few nonzero values with relatively large proportion of zeros, are very skewed and have irregular tails. In this case the normal approximation, the student c.i. is based on, of the distribution of the sample mean is not sufficiently accurate for small samples.