

### III SKYRIUS

## NEŽINOMŲ SKIRSTINIO PARAMETRŲ VERTINIMAS IR PASIKLIAUTINŲJŲ INTERVALŲ SKAIČIAVIMAS

**1. Nežinomų skirstinio parametrų vertinimas.** Apibendrinus II skyriuje pateiktą informaciją galima teigti, kad aplinkotyros duomenų statistinis modelis – turimos kintamųjų reikšmės yra atsitiktiniai dydžiai, nusakomi tam tikru skirstiniu. Šį teiginį formalizuosime: konkreti imtis  $(x_1, x_2, \dots, x_n)$  yra generuota atsitiktinio vektoriaus  $\mathbf{X} = (X_1, X_2, \dots, X_n)$  su skirstiniu, priklausančiu skirstinių šeimai  $P = P(\theta_1, \theta_2, \dots, \theta_k)$ . Šeima  $P = P(\theta_1, \theta_2, \dots, \theta_k)$  yra nusakoma žinomos analizinės išraiškos (formulės) skirstiniu, priklausančiu nuo  $k$  parametrų  $(\theta_1, \theta_2, \dots, \theta_k)$ . Žinodami šių parametrų reikšmes, galime visiškai nusakyti skirstinį. Parametrinių skirstinių šeimų pavyzdžiai: normalių skirstinių šeima  $P = P(m, \sigma)$ , susidedanti iš visų galimų normalių skirstinių; Puasono skirstinių šeima  $P(\lambda)$  (susideda iš visų Puasono skirstinių), Bernulio skirstinių šeima  $P(p)$ . Normalių skirstinių šeimos atveju turime du nežinomus parametrus:  $\theta_1$  yra vidurkis  $m$ , o  $\theta_2$  -  $\sigma$ . Žinodami  $m$  ir  $\sigma$ , galime visiškai nusakyti normaliojo skirstinio tankį funkciją. Puasono ir Bernulio skirstinių atveju turime vieną nežinomą parametą:  $\theta_1$  yra  $\lambda$  arba  $p$  (3.1.1 pav.).

Pažymėkime  $p(\mathbf{y}, \boldsymbol{\theta})$  (arba  $p_X(\mathbf{y}, \boldsymbol{\theta})$ ),  $\boldsymbol{\theta} = (\theta_1, \theta_2, \dots, \theta_k)$ , - vektoriaus  $\mathbf{X}$  tankio funkcija, jei  $\mathbf{X}$  koordinatės tolydūs ats. d., arba tikimybė, kad  $\mathbf{X}$  įgis reikšmę  $\mathbf{y}$ , jei  $\mathbf{X}$  koordinatės diskretūs ats. dydžiai. Funkcija  $p(\mathbf{X}, \boldsymbol{\theta}) = p(X_1, X_2, \dots, X_n, \boldsymbol{\theta})$  vadinama imties tikėtinumo funkcija (3.1.2 pav.).  $p(\mathbf{x}, \boldsymbol{\theta})$  funkcinė išraiška žinoma; ji visiškai charakterizuoja ats. vektoriaus  $\mathbf{X}$  skirstinį. Kadangi  $\mathbf{X}$  koordinatės nepriklausomi ats. dydžiai, tai imties tikėtinumo funkcija lygi šių koordinatinių tankių ar tikimybių sandaugai (3.1.3 pav.).

Analizuojant ekologų sukauptus duomenis dažniausiai laikoma, kad imtis yra atsitiktinis  $n$ -matis vektorius su turinčiomis tą patį skirstinį koordinatėmis. Tuomet skirstinių klasė  $P$  apibūdinama kaip vienmačio ats. d. skirstinių klase. Imties, turinčios tą patį normalių skirstinių tikėtinumo funkcija pateikta 3.1.3 pav. Jei naudojamas regresinis (*faktorius*  $\rightarrow$  *atsakas*) modelis, tai laikoma, kad atsako reikšmės yra atsitiktiniai dydžiai, kurių skirstinys priklauso parametrinei skirstinių klasei  $P(\mathbf{b})$ , o  $X_i$  skirstinio parametras  $b_i$  priklauso nuo individo faktoriaus reikšmės  $z_i$ :  $\mathbf{b}_i = \mathbf{b}(\boldsymbol{\theta}, z_i)$ , tuomet tikėtinumo funkcijos išraiška pateikta (3.1.4 – 3.1.5 pav.).

Norėdami nustatyti konkretų kintamojo skirstinį, reikia įvertinti nežinomus skirstinio parametrus  $\theta_1, \theta_2, \dots, \theta_k$  (mūsų minėtuose pavyzdžiuose  $k$  lygus vienam arba dviem). Nežinomų parametrų įverčiai dažniausiai gaunami didžiausio tikėtinumo metodu. Kartais naudojamas Bajeso ar pakartotinos atrankos metodas (3.1.6 pav.).

**Didžiausio tikėtinumo metodas.** Parametrų  $\theta_1, \theta_2, \dots, \theta_k$  įverčiais parenkame tokias konkrečios imties funkcijas, kurioms esant tikėtinumo funkcija taške  $(x_1, x_2, \dots, x_n)$  būtų didžiausia (t.y. gauta imtis  $(x_1, x_2, \dots, x_n)$  yra labiausiai tikėtina). Ieškant tikėtinumo funkcijos maksimumo, ji logaritmuojama, surandamos logaritmuotos tikėtinumo funkcijos išvestinės parametrų atžvilgiu ir prilyginamos nuliui (3.1.7 pav.). Gauta lygčių sistema išsprendžiama  $\theta_1, \theta_2, \dots, \theta_k$  atžvilgiu. Šios sistemos sprendiniai (tikslūs ar nustatyti artutiniais metodais)  $\hat{\theta}_1, \hat{\theta}_2, \dots, \hat{\theta}_k$  laikomi ieškomais nežinomų parametrų įverčiais. Didžiausio tikėtinumo metodu gauti parametrų įverčiai pakankamai geri: kuo imtis didesnė, tuo labiau tikėtina, kad parametro įvertis labai mažai skirsis nuo tikrosios parametro reikšmės.

Imties skirstinio parametrų įverčiai, gauti didžiausio tikėtinumo metodu, pateikti 3.1.8 pav.

Analizuojant kelių kintamųjų (faktorius ir atsako į jį) tarpusavio ryšį, naudojami sudėtingesni modeliai, kaip antai netiesinė, logistinė regresija. Juose nežinomi parametrai vertinami artutiniais metodais, todėl ne visuomet galima pateikti tikslias parametrų įverčių formules.

**Bajeso metodas.** Naudojant šį metodą daroma prielaida, kad parametro  $\theta$  reikšmės yra atsitiktinės su tam tikru žinomu skirstiniu. Taikant didžiausio tikėtinumo metodą, nežinomų parametrų reikšmės nustatomos, remiantis tik imties reikšmėmis, o Bajeso metodas naudoja papildomai ir informaciją apie  $\theta$ .

Norint rasti  $\theta$  įvertį reikia nustatyti jo aposteriorinį skirstinį – skirstinį esant fiksuotai imčiai  $\mathbf{x} = (x_1, x_2, \dots, x_n)$   $p(\theta|\mathbf{x})$ . Jis nustatomas pagal Bajeso formulę (3.1.9 pav.). Jei įmanoma,  $p(\theta|\mathbf{x})$  nustatomas analiziškai. Jei  $p(\theta|\mathbf{x})$  analiziškai neįmanoma, tai šis skirstinys modeliuojamas, naudojant kompiuteriu generuotus reikiamus atsitiktinius dydžius.

**Pakartotinės atrankos (*resampling*) metodai.** Naudodami imties reikšmes, generuojame naujas imtis. Kiekvienai sugeneruotai imčiai įvertiname nežinomus parametrus. Šių parametrų reikšmių vidurkis ir yra nežinomų parametrų įvertis. Iš  $n$  dydžio imties reikšmių galima sugeneruoti  $n^n$  skirtingų  $n$  dydžio imčių. Ši procedūra vadinama visa pakartotine atranka (*complete resampling*). Turime imtį: 1, 6, 9. Šios imties vidurkis  $\bar{x} = 5,33$ ,  $s = 4,04$ . Iš 1, 6, 9 reikšmių galima sudaryti  $3^3 = 27$  skirtingų imčių. 3.1.10 pav. pateikti kiekvienos šių imčių vidurkis ir standartinis nuokrypis. Šių 27 dydžių vidurkiai yra populiacijos vidurkis ir standartinis nuokrypis, gauti pakartotinės atrankos metodu. Jie atitinkamai lygūs 5,33 ir 2,96. Priklausomai nuo naujų imčių generavimo taisyklių, naudojami plėtros (*bootstrap*) bei atmetos reikšmės (*jackknife*) metodai (3.1.11 pav.).

**Parametrų įverčių kintamumo charakteristika.** Nežinomų parametrų, kaip antai vidurkio, dispersijos, tikimybės, įverčiai, gauti tiek didžiausio tikėtinumo, tiek Bajeso metodu, yra

imties funkcijos. Kadangi šie įverčiai yra atsitiktinių dydžių funkcijos, t.y. atsitiktiniai dydžiai, todėl būtina įvertinti jų kintamumą. Pastebėsime, kad visi 3.1.8 pav. pateikti parametru įverčiai yra nepriklausomų ats. d. sumų funkcijų reikšmės, todėl imtis pakankamai didelė, šių įverčių skirstinius galima laikyti normaliaisiais (remdamiesi Centrine ribine teorema). Normaliojo ats.d. kintamumą charakterizuoja dispersija arba standartinis nuokrypis. Todėl pateikiant įvertintus kintamojo (populiacijos) parametrus pateikiami ir jų standartinių nuokrypių įverčiai. Standartinių nuokrypių įverčiai dar vadinami standartinėmis paklaidomis (*standard error*), žymimi SE(...). 3.1.12 lentelėje pateikti parametru įverčių standartiniai nuokrypiai ir standartinės paklaidos.

Sakykime, kiekybinis kintamasis turi unimodalų skirstinį su vidurkiu  $m$  ir dispersija  $\sigma^2$ . Teorinio nežinomo vidurkio  $m$  įvertis yra imties vidurkis  $\bar{x}$ . Teorinė  $\bar{x}$  dispersija yra  $\sigma^2/n$ , standartinis nuokrypis –  $\sigma/\sqrt{n}$ , o standartinis nuokrypio įvertis yra  $s/\sqrt{n}$ , žymimas  $s_x$ . Įvertis  $s_x$  vadinamas standartine vidurkio paklaida (*standard error of mean*, sutrumpintai SE). Kiekybinio rodiklio vidurkis paprastai pateikiamas kartu su savo standartine paklaida: pavyzdžiui,  $\bar{x} \pm s_x$ .

Analizuojant sudėtingesnius modelius, pavyzdžiui, regresinius, vertinamas ne tik vidurkio, bet ir viso daugiamačio parametro  $\theta = (\theta_1, \theta_2, \dots, \theta_k)$  įverčio  $\hat{\theta} = (\hat{\theta}_1, \hat{\theta}_2, \dots, \hat{\theta}_k)$  kintamumas. Jis vertinamas kovariacijų matrica  $cov(\hat{\theta})$  (3.1.13 pav.). Jei  $\theta$  vertinamas Bajeso metodu, tai  $\theta$  kintamumas nustatomas modeliuojant  $p(\theta|\mathbf{x})$  skirstinį arba naudojantis tikslia  $p(\theta|\mathbf{x})$  išraiška. Jei  $\theta$  vertinamas pakartotinos atrankos metodu, tai jo standartinis nuokrypis lygus kiekvienoje imtyje nustatytų standartinių nuokrypių vidurkiui.

**Skirstinio parametru pasikliautinieji intervalai.** Sakykime,  $X$  – tolydusis ats. d., turintis vidurkį  $m$  ir dispersiją  $\sigma^2$ ,  $x_1, x_2, \dots, x_n$  - šio ats. d. imtis ( $x_i$  – normalusis ats. d.). Vidurkio  $m$  įvertis yra imties vidurkis  $\bar{x}$ . Jis taip pat yra atsitiktinis: skirtingoms konkrečioms imtims  $\bar{x}$  reikšmės, o tuo pačiu ir  $m$  įverčiai bus skirtingi. Todėl yra aktualu įvertinti vidurkio įverčio patikimumą – nustatyti, kaip jis skiriasi nuo tikrosios parametro reikšmės.

Nežinomo parametro įverčio patikimumui vertinti įvedama pasiklovimo lygmens sąvoka. Parenkamas taip vadinamas pasiklovimo lygmuo (*confidence level*) arba patikimumas  $P$  – skaičius, artimas vienetui, pvz.:  $P = 0,9; 0,95; 0,99$ . Pasiklovimo lygmuo  $P$  yra tikimybė, jog skirstinio parametras  $\theta$  randasi intervale  $[\theta_{ap}(\mathbf{x}), \theta_{virš}(\mathbf{x})]$ , čia  $\mathbf{x} = (x_1, x_2, \dots, x_n)$  – atsitiktinė imtis – nepriklausomi ats. d.

Intervalas  $[\theta_{ap}(\mathbf{x}), \theta_{virš}(\mathbf{x})]$  vadinamas parametro  $\theta$  pasikliautinu intervalu (*confidence interval*); ats. dydžiai  $\theta_{ap}(\mathbf{x})$  ir  $\theta_{virš}(\mathbf{x})$  vadinami pasikliautino intervalo apatine ir viršutine riba.  $P$  kartais pateikiamas procentais. Nuo  $P$  parinkimo priklauso pasiklovimo intervalų ilgis - kuo  $P$  artimesnis vienetui, tuo pasikliautinas intervalas platesnis (3.2.2 pav.). Pasiklovimo intervalas

$[\theta_{ap}(\mathbf{x}), \theta_{vir}(\mathbf{x})]$  nusako nežinomo parametro  $\theta$  įvertinimo tikslumą – tai intervalas, kuriame su artima 1 tikimybe yra tikroji parametro reikšmė.

Dažniausiai naudojamų parametru pasiklovimo intervalai sudaromi taip: nustatoma imties funkcija, vadinama statistika,  $T(\theta, \mathbf{x})$ , į kurią dažniausiai įeina ir parametro įvertis  $\hat{\theta}(\mathbf{x})$ , turinti žinomą skirstinį (standartinį normalųjį, Studento,  $\chi^2$  ar kitokį). Pasiklovimo intervalo ribos priklauso nuo parametro  $\theta$  įverčio, statistikos  $T(\theta, \mathbf{x})$  skirstinio kvantilių ir nuo pasiklovimo lygmens  $P$ . Biomedicinos duomenų analizėje standartinė patikimumo reikšmė  $P = 0,95$  arba jei  $P$  pateikiamas procentais -  $P = 95\%$ . Konkrečiai imčiai nustatyto parametro pasiklovimo intervalo interpretacija pateikta 3.2.4 pav.

Aplinkos duomenų analizėje dažniausiai skaičiuojami normaliojo skirstinio vidurkio ir dvinario kintamojo tikimybės pasiklovimas intervalas.

**Normaliojo skirstinio vidurkio pasiklovimas intervalas** pateiktas 3.2.5 pav. Šiame pasiklovime intervale su statistine tikimybe  $P$  yra skirstinio vidurkio reikšmė  $m$ . Tai reiškia, kad tam tikram skaičiui konkrečių imčių apskaičiavus vidurkio PI,  $P$  procentų atvejų skirstinio vidurkio reikšmė bus apskaičiuotame pasiklovime intervale. Dydis  $s/\sqrt{n}$ , naudojamas vidurkio pasiklovimo intervalo formulėje, yra imties vidurkio standartinė paklaida; taigi žinant imties vidurkį ir standartinę paklaidą, galima paskaičiuoti vidurkio pasiklovimą intervalą bet kuriam patikimumui. Jei imtis didelė ( $n > 100$ ), tuomet vidurkio pasiklovimo intervalo formulėje vietoje Studento skirstinio kvantilio galima naudoti standartinio normaliojo skirstinio atitinkamo lygio kvantilį – kai  $P = 0,95$ ,  $(1 + P)/2 = 0,975$ ,  $z_{0,975} = 1,96$ .

Vidurkio pasiklovimo intervalo skaičiavimas 3.2.5 pav. pateikta formule ir plėtos metodu bei rezultatų palyginimas pateiktas 3.2.6 pav.

**Dvinario kintamojo tikimybės pasiklovimas intervalas.** Sakykime, dvinario kintamojo imtis - ats. dydis  $X$ , įgyjantis dvi reikšmes – 1 ir 0 su tikimybėmis  $P\{X = 1\} = p$ ,  $P\{X = 0\} = 1 - p$ , čia  $p$  – nežinomas parametras. Tikimybės  $p$  įvertis yra:  $k/n$ , čia  $k$  – vienetų skaičius imtyje. Tikslus tikimybės  $p$  pasiklovimas intervalas pateiktas 3.2.7 pav.

Naudojantis šiomis formulėmis, tikslų tikimybės pasiklovimą intervalą paskaičiuoti problematiška. Todėl, kai  $n$  pakankamai didelis, vietoje tikslios intervalo apatinės ir viršutinės reikšmės naudojamos artutinės (3.2.8 pav.)

**Pasiklovimų intervalų grafinis vaizdavimas.** Pasiklovimieji intervalai gali būti grafiškai pateikti stačiakampe diagrama arba stulpeliu ar tašku su ūseliais (*whiskers*). Standartinė stačiakampė diagrama brėžiama taip: stačiakampio viduryje kvadratėliu pažymimas vidurkis; stačiakampis braižomas nuo vidurkio minus standartinė paklaida  $(\bar{x} - s/\sqrt{n})$  iki vidurkis plus

standartinė paklaida ( $\bar{x} + s / \sqrt{n}$ ). Nuo stačiakampio apačios brėžiamas apatinis “ūsas” tęsiasi iki ( $\bar{x} - 1,96s / \sqrt{n}$ ), o viršutinis “ūsas” prasideda nuo stačiakampio viršaus ir brėžiamas iki reikšmės ( $\bar{x} + 1,96s / \sqrt{n}$ ). 1,96 yra normaliojo skirstinio 0,975 lygio kvantilis, taigi stačiakampės diagramos “ūsai” nurodo vidurkio pasikliautiną intervalą, skaičiuotą standartiniam patikimumui  $P = 0,95$  ir pakankamai didelei imčiai (3.2.9 pav.). Vidurkio pasikliautinas intervalas stulpeliu ar tašku su ūseliais vaizduojamas taip: taško Y koordinatė ar stulpelio aukštis atitinka vidurkį, o ūselių dydis -  $t_{(1+P)/2}SE$ .

Pateikus kelių imčių (ar kelių parametru) vidurkio pasikliautiną intervalą grafiškai, nesunku vidurkius palyginti. Jei pasikliautini intervalai nepersikerta, galima daryti išvadą, kad imčių skirstinių (populiacijų) vidurkiai yra skirtingi. Iš 3.2.9 – 3.2.10 pav. matome, kad balandžio mėn.  $SO_4^{-2}$  jonų koncentracija krituliuose patikimai didesnė nei gegužės – spalio mėn.