VILNIAUS GEDIMINO TECHNIKOS UNIVERSITETAS
MATEMATIKOS IR INFORMATIKOS INSTITUTAS

Aleksej BAKŠAJEV

STATISTINIŲ HIPOTEZIŲ TIKRINIMAS, NAUDOJANT N-METRIKAS

DAKTARO DISERTACIJOS SANTRAUKA

FIZINIAI MOKSLAI, MATEMATIKA (01P)

Vilnius LEIDYKLA TECHNIKA 2010

Disertacija rengta 2004–2009 metais Matematikos ir informatikos institute.
Mokslinis vadovas
prof. habil. dr. Rimantas RUDZKIS (Matematikos ir informatikos institutas, fiziniai mokslai, matematika – 01P).
Mokslinis konsultantas
prof. habil. dr. Yurij TYURIN (Maskvos valstybinis M. V. Lomonosovo universitetas, fiziniai mokslai, matematika – 01P).
Disertacija ginama Vilniaus Gedimino technikos universiteto Matematikos mokslo krypties taryboje:
Pirmininkas
prof. habil. dr. Kęstutis KUBILIUS (Matematikos ir informatikos institutas, fiziniai mokslai, matematika – 01P).
Nariai:
prof. habil. dr. Vygantas PAULAUSKAS (Vilniaus universitetas, fiziniai mokslai, matematika – 01P),
prof. habil. dr. Alfredas RAČKAUSKAS (Vilniaus universitetas, fiziniai mokslai, matematika – 01P),
prof. habil. dr. Leonas SAULIS (Vilniaus Gedimino technikos universitetas, fiziniai mokslai, matematika – 01P),
prof. habil. dr. Jonas Kazys SUNKLODAS (Vilniaus Gedimino technikos universitetas, fiziniai mokslai, matematika – 01P).
Oponentai:
prof. habil. dr. Vilijandas BAGDONAVIČIUS (Vilniaus universitetas, fiziniai mokslai, matematika – 01P),
doc. dr. Marijus RADAVIČIUS (Matematikos ir informatikos institutas, fiziniai mokslai, matematika – 01P).

Disertacija bus ginama viešame Matematikos mokslo krypties tarybos posėdyje 2010 m. vasario 26 d. 10 val. Matematikos ir informatikos institute, 203 auditorijoje.
Adresas: Akademijos g. 4, LT-08663 Vilnius, Lietuva.
Tel.: (8 5) 274 4952; faksas (8 5) 270 0112; el. paštas doktor@vgtu.lt
Disertacijos santrauka išsiuntinėta 2010 m. sausio 25 d.
Disertaciją galima peržiūrėti Vilniaus Gedimino technikos universiteto (Saulėtekio al. 14, LT-10223 Vilnius, Lietuva) ir Matematikos ir informatikos instituto (Akademijos g. 4, LT-08663 Vilnius, Lietuva) bibliotekose.
VGTU leidyklos „Technika“ 1716-M mokslo literatūros knyga.

VILNIUS GEDIMINAS TECHNICAL UNIVERSITY
INSTITUTE OF MATHEMATICS AND INFORMATICS

Aleksej BAKŠAJEV

STATISTICAL TESTS BASED ON N-DISTANCES

SUMMARY OF DOCTORAL DISSERTATION

PHYSICAL SCIENCES, MATHEMATICS (01P)

Vilnius LEIDYKLA TECHNIKA 2010

The scientific work was prepared at Institute of Mathematics and Informatics in 2005–2009.
Scientific Supervisor
Prof Dr Habil Rimantas RUDZKIS (Institute of Mathematics and Informatics, Physical Sciences, Mathematics – 01P).
Scientific Consultant
Prof Dr Habil Yurij TYURIN (Lomonosov Moscow State University, Physical Sciences, Mathematics – 01P).
The dissertation is being defended at the Council of Scientific Field of Mathematics at Vilnius Gediminas Technical University:
Chairman
Prof Dr Habil Kęstutis KUBILIUS (Institute of Mathematics and Informatics, Physical Sciences, Mathematics – 01P).
Members:
Prof Dr Habil Vygantas PAULAUSKAS (Vilnius University, Physical Sciences, Mathematics – 01P),
Prof Dr Habil Alfredas RAČKAUSKAS (Vilnius University, Physical Sciences, Mathematics – 01P),
Prof Dr Habil Leonas SAULIS (Vilnius Gediminas Technical University, Physical Sciences, Mathematics – 01P),
Prof Dr Habil Jonas Kazys SUNKLODAS (Vilnius Gediminas Technical University, Physical Sciences, Mathematics – 01P).
Opponents:
Prof Dr Habil Vilijandas BAGDONAVIČIUS (Vilnius University, Physical Sciences, Mathematics – 01P),
Assoc Prof Dr Marijus RADAVIČIUS (Institute of Mathematics and Informatics, Physical Sciences, Mathematics – 01P).

The dissertation will be defended at the public meeting of the Council of Scientific Field of Mathematics at the Institute of Mathematics and Informatics, Room 203, at 10 a. m. on 26 February 2010.
Address: Akademijos g. 4, LT-08663 Vilnius, Lithuania.
Tel. +370 5 274 4952; fax +370 5 270 0112; e-mail: doktor@vgtu.lt
The summary of the doctoral dissertation was distributed on 25 January 2010.
A copy of the doctoral dissertation is available for review at the Libraries of Vilnius Gediminas Technical University (Saulėtekio al. 14, LT-10223 Vilnius, Lithuania) and the Institute of Mathematics and Informatics (Akademijos 4, LT-08663 Vilnius, Lithuania).

Įvadas

Mokslinė problema

Šiuolaikinėje statistikoje sudėtingesnių sprendimų priėmimas būtinai apima ir hipotezių tikrinimą. Disertacijoje nagrinėjama klasikinių statistinių hipotezių tikrinimo problema, naudojant N-metrikų teoriją.

Darbo aktualumas

Duomenų analizėje tyrėjai stengiasi visapusiškai panaudoti ne tik turimą statistinę, bet ir apriorinę informaciją, todėl dažnai pradeda savo tyrimus nuo prielaidų apie stebėjimų skirstinius. Tai yra daroma dėl kelių priežasčių:

- Esamų duomenų skirstinys gali paaiškinti duomenis teikiantį procesą. Jei siūlomas nagrinėjamo objekto modelis yra teisingas, imties skirstinys turi tenkinti atitinkamas prielaidas.
- Pasiskirstymo charakteristikos gali būti susijusios su svarbiais bazinio modelio parametrais.
- Žinios apie duomenų pasiskirstymą leidžia efektyviai taikyti standartines modelio identifikavimo procedūras ir minimizuoti vidutinius statistinių sprendimų nuostolius.

Kartais tokios prielaidos apie nagrinėjamų duomenų skirstinį ar jo tipą yra padaromos remiantis metodika, kurią taikant duomenys buvo gauti ar surinkti. Bet, kaip taisyklė, prireikia patikrinti, ar pasirinktas pasiskirstymas yra adekvatus. Tyrėjas gali domėtis, ar imties skirstinys sutampa su aprioriškai parinktu (paprastoji hipotezė), ar priklauso kokiai nors skirstinių šeimai (sudėtinė hipotezė). Daugiamačių duomenų atveju kartu su suderinamumo hipotezės tikrinimu dažnai prireikia patikrinti prielaidą apie stebinių komponenčių nepriklausomumą. Kita uždavinių grupė apima dviejų ar keleto imčių skirstinių palyginimą. Tam taikomi homogeniškumo hipotezės statistiniai kriterijai.

Sprendžiant šias problemas buvo sukurta daugybė homogeniškumo, suderinamumo ir nepriklausomumo statistinių kriterijų, iš kurių parinkti tinkamiausią praktikoje nėra lengva. Galingiausio testo parinkimas iš keleto galimų yra viena svarbiausių statistikos problemų. Tačiau daugumai paplitusių teorinių modelių galingiausi pagal visas alternatyvas statistiniai kriterijai nėra žinomi arba neegzistuoja. Todėl naujų konstruktyvių statistinių kriterijų, jautrių įvairiems hipotezių tipams, sukūrimas išlieka aktualus ir mūsų dienomis.

L. B. Klebanovas savo darbuose (Zinger, Klebanov ir Kakosyan, 1989; Klebanov, 2005) įvedė naują tikimybinių matų klasę, vadinamąsias N-metrikas. Ši klasė pasižymi daugeliu gerų savybių ir gali būti sėkmingai taikoma naujiems

galingiems ir patogiems statistiniams kriterijams sudaryti. Tokių statistikų konstravimas ir jų savybių tyrimas tapo itin aktualus po minėtų Klebanovo darbų.

Tyrimo objektas

Disertacija skirta suderinamumo (parametriniu ir neparametriniu atveju), homogeniškumo, simetriškumo ir nepriklausomumo hipotezių tikrinimui, naudojant N-metrikas. Sudaromi atitinkami statistiniai kriterijai, tiriamos pasiūlytų kriterijų asimptotinės savybės bei įvertinamas jų galingumas.

Tikslas ir uždaviniai

Pagrindinis disertacinio darbo tikslas yra N-metrikų teorijos pritaikymas klasikinėms statistinėms suderinamumo, homogeniškumo, simetriškumo bei nepriklausomumo hipotezėms tikrinti. Siekiant tikslo buvo sprendžiami šie uždaviniai:

- minėtų hipotezių testinių statistikų konstravimas, naudojant N-metrikas;
- pasiūlytų kriterijų kritinės srities nustatymas bei testinių statistikų asimptotinių skirstinių gavimas;
- pasiūlytų N-metrikos tipo kriterijų bei klasikinių testų palyginimas, naudojant Bahaduro asimptotinį santykinį efektyvumą (Bahadur, 1960; Nikitin, 1995).

Kartu su teoriniais rezultatais pasiūlytų N-metrikos tipo testų galingumas ištirtas, naudojant Monte-Karlo metodą.

Tyrimų metodika

Disertacijoje taikomi įvairūs bendros tikimybių teorijos, atsitiktinių procesų ir matematinės statistikos metodai. Pasiūlytų kriterijaus statistikų asimptotinis skirstinys įrodytas naudojant U-statistikų teoriją (Lee, 1990; Koroljuk ir Borovskich, 1994) bei stochastinių procesų silpnojo konvergavimo savybes (Bulinskii ir Shiryaev, 2005). Empirinėje dalyje pateikti rezultatai gauti taikant Monte Karlo metodą, naudojantis statistiniu paketu R.

Mokslinis naujumas

Mokslinio darbo originalumas ir naujumas yra susijęs su formuluojamomis užduotimis. Siūlomi metodai pratęsia, apibendrina ir papildo Klebanovo (Klebanov, 2005), Baringhauso ir Franzo (Baringhaus ir Franz, 2004) bei Szekely ir Rizzo (Szekely ir Rizzo, 2005) rezultatus. Siūlomi suderinamumo, simetriškumo, nepriklausomumo ir tolygumo hipersferoje kriterijai bei nustatyti atitinkamų testinių statistikų asimptotiniai skirstiniai nebuvo anksčiau nagrinėti statistikos literatūroje.

Darbo rezultatų praktikinė vertė

Darbe sukurti statistiniai kriterijai gali būti pritaikyti realiuose duomenų analizės hipotezių tikrinimo uždaviniuose.

Ginamieji teiginiai

- Teiginiai apie suderinamumo kriterijaus N-metrikos statistikų asimptotinį skirstinį.
- N-metrikos testinės statistikos sudarymo bei asimptotinio skirstinio nustatymo metodai tikrinant tolygumo sferoje S^{p-1} hipotezę.
- Teiginiai apie homogeniškumo kriterijaus testinių statistikų asimptotinį skirstinį; pasiūlytų testų kritinės aibės nustatymo algoritmai taikant butstrepo (angl. bootstrap) bei perstatinių (angl. permutation) metodus; nepriklausančio nuo tiriamo skirstinio homogeniškumo testo sudarymo būdas bei teiginiai apie šio kriterijaus asimptotinį elgesį.
- N-metrikos tipo kriterijaus konstrukcija bei asimptotinio skirstinio nustatymas simetriškumo ir nepriklausomumo hipotezėms tikrinti.
- Teiginiai apie N-metrikų tipo testinių statistikų skaičiavimo išraiškas su įvairiais stipriai neigiamai apibrėžtais branduoliais.
- Pasiūlytų N-atstumų tipo bei klasikinių suderinamuno testų palyginimas vienmačiu atveju, naudojant Bahaduro asimptotinį santykinį efektyvumą.

Disertacijos struktūra

Disertacinis darbas sudarytas iš įvado, kurio paskirtis yra supažindinti skaitytojus su nagrinėjama problematika, ir keturių skyrių, kuriuose pateikiami pagrindiniai rezultatai. Darbo pabaigoje pateikiamos bendrosios išvados ir literatūros sąrašas. Disertacija parašyta anglų kalba. Bendra darbo apimtis – 148 puslapių.

1. Pagalbiniai rezultatai

Disertacinis darbas yra skirtas N-metrikų teorijos pritaikymui klasikinėms statistinėms suderinamumo, homogeniškumo, simetriškumo bei nepriklausomumo hipotezėms tikrinti.

Pirmas skyrius susideda iš dviejų dalių, kuriose pateikta pagalbinių rezultatų apžvalga. Pirmame poskyryje yra trumpai aprašyti N-metrikų teorijos bendri aspektai ir sąvokos, reikalingos darbe pasiūlytų statistinių testų konstravimui.

Tegul $(\mathfrak{X}, \mathfrak{U})$ yra mačioji erdvė ir B yra tikimybinių matų μ, apibrėžtų šioje erdvėje, aibė. Tarkime, kad L yra reali ir tolydi simetrinė funkcija, B_L yra aibė tikimybinių matų μ iš B, tenkinančių salygą

$$\int_{\mathfrak{X}}\int_{\mathfrak{X}} L(x,y)\,d\mu(x)\,d\mu(y) < \infty.$$

Viena iš pagrindinių N-metrikų teorijos sąvokų yra stipriai neigiamai apibrėžtas branduolys.

1 apibrėžimas. *Jeigu* $\forall n \in \mathbb{N}$, $\forall x_1, \ldots, x_n \in \mathfrak{X}$ *ir* $\forall c_1, \ldots, c_n \in \mathbb{C}$ *su sąlyga, kad* $\sum_{i=1}^{n} c_i = 0$, *funkcija* $L(x,y)$ *tenkina nelygybę*

$$\sum_{i=1}^{n}\sum_{j=1}^{n} L(x_i,x_j)c_i\bar{c_j} \leq 0,$$

tai L *yra neigiamai apibrėžtas branduolys.*

2 apibrėžimas. *Tegul* P *yra matas, apibrėžtas erdvėje* $(\mathfrak{X},\mathfrak{U})$, *o* $h(x)$ *– funkcija, tenkinanti sąlygą, kad* $\int_{\mathfrak{X}} h(x)\,dP(x) = 0$. *Tada branduolys* $L(x,y)$ *yra stipriai neigiamai apibrėžtas, jeigu* $L(x,y)$ *yra neigiamai apibrėžtas ir kiekvienam* P *iš lygybės*

$$\int_{\mathfrak{X}}\int_{\mathfrak{X}} L(x,y)h(x)h(y)\,dP(x)\,dP(y) = 0$$

išplaukia, kad $P\{x : h(x) \neq 0\} = 0$.

Pažymėkime

$$\begin{aligned} N(\mu,\nu) \quad &:= 2\int_{\mathfrak{X}}\int_{\mathfrak{X}} L(x,y)d\mu(x)d\nu(y) - \int_{\mathfrak{X}}\int_{\mathfrak{X}} L(x,y)d\mu(x)d\mu(y) - \\ &- \int_{\mathfrak{X}}\int_{\mathfrak{X}} L(x,y)d\nu(x)d\nu(y), \end{aligned}$$

čia $\mu, \nu \in B_L$.

Klebanov parodė (Klebanov, 2005), kad jeigu $L(x,y) = L(y,x)$ ir $L(x,x) = 0\ \forall x, y \in \mathfrak{X}$, tai nelygybė

$$\begin{cases} N(\mu,\nu) > 0, & \mu \neq \nu, \\ N(\mu,\nu) = 0, & \mu = \nu, \end{cases}$$

yra teisinga visiems matams $\mu, \nu \in B_L$ tada ir tik tada, kai $L(x,y)$ yra stipriai neigiamai apibrėžtas branduolys. Šitas teiginys leidžia konstruoti suderintuosius kriterijus prieš visas alternatyvas.

Nustatytas disertacijoje pasiūlytų testinių statistikų asimptotinis skirstinys nulinės hipotezės atveju sutampa su tam tikros Gauso atsitiktinių dydžių ζ_i, $i \in \mathbb{N}$,

kvadratinės formos $Q = \sum_{i=1}^{\infty} \lambda_i \zeta_i^2$ skirstiniu. Todėl **1.2 poskyryje** yra trumpai aprašyti Gauso atsitiktinių dydžių kvadratinės formos pasiskirstymo funkcijos skaičiavimo metodai, remiantis formos Q charakteristinės funkcijos apgręžimu (Imhof, 1961; Sukhatme, 1972; Martynov, 1975).

Natūralus minėtų statistinių hipotezių tikrinimo būdas yra N-atstumo tarp empirinio ir hipotetinio (suderinamumo hipotezė) arba tarp dviejų empirinių skirstinių (homogeniškumo, simetrijos ir nepriklausomybės hipotezės) nagrinėjimas. Sekantys du skyriai yra skirti nurodytų hipotezių kriterijaus statistikų sudarymui.

2. Suderinamumo testas

Antrame skyriuje išsamiau analizuojami suderinamumo testai. Šis skyrius susideda iš dviejų dalių, skirtų parametriniam ir neparametriniam atvejams.

Paprastosios suderinamumo hipotezės tikrinimas

Tegul $X_1, \ldots, X_n$ yra p-mačio atsitiktinio dydžio X su nežinoma tolydžia pasiskirstymo funkcija $F(x)$ nepriklausomi stebėjimai. Neparametrinė nulinė suderinamumo hipotezė turi pavidalą $H_0 : F(x) = G(x)$, čia $G(x)$, $x \in \mathbb{R}^p$, yra žinoma tolydi pasiskirstymo funkcija. Kriterijaus statistikos

$$T_n = -n \int_{\mathbb{R}^{2p}} L(x,y)\, d(F_n(x) - G(x))\, d(F_n(y) - G(y)) \tag{1}$$

didelių reikšmių atveju hipotezė H_0 turi būti atmesta, čia $F_n(x)$ yra empirinė pasiskirstymo funkcija, sukonstruota naudojant $X_1, \ldots, X_n$.

Testo kritinės aibės nustatymui darbe yra nagrinėjamas statstikos T_n ribinis skirstinys, kuris yra nustatytas taikant du metodus: V-statistikų asimptotinę teoriją (Lee, 1990; Koroljuk ir Borovskich, 1994) ir empirinių procesų bei jų funkcionalų silpnojo konvergavimo teoremas.

Pirmas metodas nusako bendrą pasiūlytos statistikos (1) ribinio dėsnio nustatymo būdą. Statistika T_n galima perrašyti von Miseso funkcionalu su simetriniu branduoliu $H(x,y)$:

$$T_n = \frac{1}{n} \sum_{i=i}^{n} \sum_{j=1}^{n} H(X_i, X_j), \tag{2}$$

čia

$$H(x,y) := \mathbf{E}L(x,X) + \mathbf{E}L(X,y) - L(x,y) - \mathbf{E}L(X,X^{'}),$$

X, $X^{'}$ yra nepriklausomi atsitiktiniai dydžiai su pasiskirstymo funkcija $G(x)$. Asimptotinis T_n skirstinys yra nustatytas 1 ir 2 teoremose:

1 teorema. *Jeigu $G(x)$ tenkina sąlygą $\mathbf{E}H^2(X, X^{'}) < \infty$, tai, esant teisingai nulinei hypotezei, asimptotinis statistikos T_n skirstinys sutampa su kvadratinės formos*

$$Q = \mathbf{E}[L(X, X^{'}) - L(X, X)] + \sum_{j=1}^{\infty} \lambda_j(\zeta_j^2 - 1) \tag{3}$$

skirstiniu, čia ζ_j, $j = 1, 2, ...$, yra nepriklausomi standartiniai Gauso atsitiktiniai dydžiai, λ_j – integralinio operatoriaus A tikrinės reikšmės,

$$Af(y) = \mathbf{E}H(X, y)f(X). \tag{4}$$

Toliau nagrinėjamas statistikos T_n ribinis skirstinys alternatyvos atveju. Pažymėkime

$$a := \mathbf{E}H(Y, Y^{'}) = 2\mathbf{E}L(X, Y) - \mathbf{E}L(X, X^{'}) - \mathbf{E}L(Y, Y^{'})$$

ir $T_n^* = \frac{T_n}{\sqrt{n}} - a$, čia X, $X^{'}$ bei Y, $Y^{'}$ yra nepriklausomi atsitiktiniai dydžiai su pasiskirstymo funkcijomis X,$X^{'} \sim F(x)$ ir Y,$Y^{'} \sim G(x)$.

2 teorema. *Tegul $H^*(x) := [\mathbf{E}H(x, Y) - a]^2$, $\mathbf{E}H^*(Y^{'}) > 0$ ir $\mathbf{E}H(Y, Y^{'}) < \infty$. Tada T_n^* yra asimptotiškai normalus su nuliniu vidurkiu ir dispersija σ^2,*

$$\sigma^2 = \frac{2}{n(n-1)}[2(n-2)C_1 + C_2], \tag{5}$$

čia $C_1 = \mathbf{E}H^*(Y^{'})$ ir $C_2 = \mathbf{E}[H(Y, Y^{'}) - a]^2$.

Tačiau praktiškai gana sudėtinga nustatyti statistikos T_n ribinį skirstinį (3) forma. Pagrindinė problema čia yra susijusi su integralinio operatoriaus (4) tikrinių reikšmių skaičiavimu. **2.1–2.3 poskyriuose** mes pabandėme išspręsti šią problemą. Antras metodas N-atstumo statistikos ribiniam skirstiniui gauti yra pagrįstas empirinių procesų ir jų funkcionalų silpnojo konvergavimo teorija (Rosenplatt, 1952; Durbin, 1973; Tyurin, 1977; Martynov, 1978). Kartu su analizinėmis kvadratinės formos koeficientų formulėmis įvairiems stipriai neigiamai apibrėžtiems branduoliams $L(x, y)$ pateikiami kai kurie praktiniai rezultatai pasiūlytų kriterijų taikymui vienmačiu ir dvimačiu atvejais. Iš pradžių pasiūlyti metodai nagrinėjami vienmačiu atveju, po to apibendrinti. Plačiau ištirtas dvimatis atvejis.

Bendru atveju, siekdami išvengti T_n skirstinio priklausomybės nuo $G(x)$, iš pradžių taikant Rosenblato (Rosenblatt, 1952) arba Bickelio-Breimano (Bickel ir

Breiman, 1983) transformaciją transformuosime pradinę imtį $X_1, \ldots, X_n$, $X_i = (X_{i,1}, ..., X_{i,p})$, $i = 1, .., n$, į imtį $t_1, \ldots, t_n$, $t_i \in [0,1]^p$,
$t_i = (t_{i,1}, ..., t_{i,p})$, $i = 1, .., n$.

Esant nulinei hipotezei, transformuotoji imtis turės tolygų vienetiniame hiperkube $C^p = [0,1]^p$ skirstinį, o statistika T_n tolygumo hypotezei tikrinti turės pavidalą

$$T_n = -n \int_{[0,1]^{2p}} L(x,y)\, d(F_n(x) - x_1 \cdots x_p)\, d(F_n(y) - y_1 \cdots y_p), \qquad (6)$$

čia $F_n(x)$, $x = (x_1, ..., x_p)$, yra p-matė empirinė pasiskirstymo funkcija, sudaryta imties $t_1, \ldots, t_n$ pagrindu.
Asimptotinis statistikos T_n skirstinys nustatytas sekančioje teoremoje.

3 teorema. *Nulinės hipotezės atveju statistikos T_n ribinis skirstinys sutampa su kvadratinės formos Q skirstiniu,*

$$Q = \sum_{i,j=1}^{\infty} a_{ij} \sqrt{\alpha_i \alpha_j}\, \zeta_i \zeta_j, \qquad (7)$$

čia ζ_i yra nepriklausomi standartiniai normalieji atsitiktiniai dydžiai,

$$a_{ij} = -\int_{[0,1]^{2p}} L(x,y)\, d\psi_i(x)\, d\psi_j(y), \quad x, y \in \mathbb{R}^p. \qquad (8)$$

Čia α_i ir $\psi_i(x)$ yra integralinio operatoriaus A tikrinės reikšmės ir funkcijos,

$$Af(x) = \int_{[0,1]^p} K(x,y) f(y) dy, \qquad (9)$$

čia

$$K(x,y) = \prod_{i=1}^{p} \min(x_i, y_i) - \prod_{i=1}^{p} x_i y_i,$$

$x = (x_1, ..., x_p)$, $y = (y_1, ..., y_p)$.

Pagrindinės problemos taikant 3 teoremą yra susijusios su integralinio operatoriaus (9) tikrinių reikšmių ir funkcijų skaičiavimu. Skyriaus pabaigoje pateikti kai kurie šios teoremos taikymo rezultatai įvairiems neigiamai apibrėžtiems branduoliams $L(x,y)$ dvimačiu atveju. Pagrindiniai principai buvo paimti iš kovariacijos operatoriaus (9) tikrinių reikšmių skaičiavimo metodo, pasiūlyto (Tyurin, 1977).

Sudėtinės suderinamumo hipotezės tikrinimas

Tegul $X_1, \ldots, X_n$ yra atsitiktinio dydžio X su tolydžia pasiskirstymo funkcija $F(x)$ nepriklausomų stebėjimų imtis. Parametrinė nulinė suderinamumo hipotezė turi pavidalą

$$H_0 : F(x) \in \mathbf{\Lambda} = \{G(x,\theta),\ x \in \mathbb{R}^p,\ \theta \in \Theta \subset \mathbb{R}^d\},$$

čia $\mathbf{\Lambda}$ yra parametrinė pasiskirstymo funkcijų šeima.

Kriterijaus statistika hypotezės H_0 tikrinimui, sukonstruota naudojant N-atstumą su branduoliu $L(x,y)$, yra formos

$$T_n = -n \int_{R^p} \int_{R^p} L(x,y)\, d(F_n(x) - G(x,\hat{\theta}_n))\, d(F_n(y) - G(y,\hat{\theta}_n)), \quad (10)$$

čia $\hat{\theta}_n$ yra nežinomo parametro θ įvertis, apskaičiuotas darant prielaidą, kad X turi skirstinį iš $\mathbf{\Lambda}$, $F_n(x)$ – empirinė pasiskirstymo funkcija.

Pagrindiniai **2.2 poskyrio** rezultatai yra susiję su statistikos (10) asimptotiniu skirstiniu. Detaliau ribinio T_n skirstinio problema aptariama **2.2.1 poskyryje** vienamačiu atveju. Kai kurie praktiniai rezultatai yra pateikti, kai $\mathbf{\Lambda}$ – Gauso arba eksponentinių pasiskirstymo funkcijų šeima. Daugiamačiu atveju nagrinėjamas tik normalumo kriterijus (**2.2.2 poskyris**). Nors šiuo atveju asimptotinis kriterijaus statistikos skirstinys nėra nustatytas, kritinė testo sritis gaunama naudojant Monte Karlo imitavimą.

Pagal analogiją su paprastosios hipotezės tikrinimo problema, aptarta ankstesniame skyriuje, nagrinėsime statistiką T_n po pradinės imties transformacijos

$$T_n = -n \int_0^1 \int_0^1 L(x,y)\, d(F_n^*(x) - x)\, d(F_n^*(y) - y), \quad (11)$$

čia $F_n^*(x)$ yra empirinė pasiskirstymo funkcija, sudaryta imties $t_1, \ldots, t_n$, $t_i = G(X_i, \hat{\theta}_n)$, $i = 1, 2, ..., n$, pagrindu.

Darant tam tikras prielaidas, susijusias su parametro $\theta \in \Theta \subset \mathbb{R}^d$ įverčiu bei hipotetinio pasiskirstymo funkcijos savybėmis, asimptotinis statistikos T_n skirstinys yra nustatytas 4 teoremoje.

4 teorema. *Tegul $\hat{\theta}_n$ yra parametro θ maksimalaus tikėtinumo įvertis. Esant nulinei hipotezei, statistikos T_n (11) ribinis skirstinys sutampa su kvadratinės formos*

$$Q = \sum_{k=1}^{\infty} \sum_{j=1}^{\infty} a_{kj} \zeta_k \zeta_j \quad (12)$$

skirstiniu, čia ζ_k, $k = 1, 2, ...$, yra nepriklausomi standartiniai normalieji atsitiktiniai dydžiai, o koeficientai a_{ij} turi pavidalą

$$a_{ij} = -\sqrt{\lambda_i \lambda_j} \int_0^1 \int_0^1 L(x, y) d\psi_i(x) d\psi_j(y). \tag{13}$$

Čia λ_k ir $\psi_k(x)$ yra integralinio operatoriaus A tikrinės reikšmės ir funkcijos,

$$Af(x) = \int_0^1 K(x, y) f(y) dy, \tag{14}$$

čia

$$K(x, y) = \min(x, y) - xy - q(x, \theta) I^{-1}(\theta) q(y, \theta), \tag{15}$$

$q(y, \theta) = G'_\theta(x, \theta)$, $y = G(x, \theta)$, o $I(\theta)$ yra Fišerio informacinė matrica.

Koreliacinės funkcijos (15) forma rodo, kad T_n nėra nepriklausoma nuo skirstinio, nes statistikos skirstinys priklauso nuo G. Dar blogiau yra tai, kad bendru atveju net asimptotiškai negalima išvengti priklausymo nuo nežinomo parametro θ. Tačiau tam tikrais atvejais šios parametrinės priklausomybės išvengti galima. Tai pavyksta padaryti, kai $\mathbf{\Lambda}$ yra pasiskirstymo funkcijų šeima su poslinkio bei mastelio parametrais,

$$\mathbf{\Lambda} = \{G(\frac{x - \theta_1}{\theta_2}),\ \theta_1 \in \mathbf{R},\ \theta_2 > 0\}.$$

Branduolys $K(x, y)$ taip pat nepriklauso nuo nežinomų parametrų dar vienai šeimai $\mathbf{\Lambda}$, būtent pasiskirstymo funkcijų šeimai su mastelio ir formos parametrais

$$\mathbf{\Lambda} = \{G((\frac{x}{\theta_1})^{\theta_2}),\ \theta_1 > 0,\ \theta_2 > 0\}.$$

Ši šeima apima tokius žinomus pasiskirstymus, kaip Veibulo, log-logistinį ir kitus.

Sekančiuose poskyriuose 4 teorema yra pritaikyta Gauso bei eksponentinio suderinamumo hipotezės tikrinimo problemoms. Pateikti atitinkamos diagonalizuotos kvadratinės formos (12) didžiausių koeficientų skaičiavimo rezultatai.

Daugiamačiu atveju kriterijaus statistikos (10) asimptotinio skirstinio nustatymas begalinės kvadratinės formos pavidalu tampa sudėtingu uždaviniu. Pagrindiniai sunkumai čia, kaip ir neparametrinės hipotezės atveju, yra susiję su tam tikro operatoriaus tikrinių reikšmių ir funkcijų skaičiavimu. Tačiau specifinėms šeimoms $\mathbf{\Lambda}$ galima pasiūlyti, kai kurias alternatyvias procedūras, skirtas nustatyti

testo kritinę sritį. **2.2.2 poskyryje** yra nagrinėjamas normalumo kriterijus daugiamačiu atveju. Tarp suderinamumo hipotezių normalumo prielaida yra populiariausia.

Tegul $X_1, \ldots, X_n$ yra atsitiktinio dydžio X su pasiskirstymo funkcija $F(x)$, $x \in \mathbb{R}^p$, nepriklausomų stebėjimų imtis. Nulinė normalumo hipotezė turi pavidalą

$$H_0 : F(x) \in N_p(a, \Sigma),$$

čia a ir Σ – nežinomi normaliojo skirstinio vidurkių vektorius bei kovariacinė matrica.

Statistikos T_n (10), pritaikytos hipotezei H_0 tikrinti, skirstinys priklauso nuo nežinomų parametrų a ir Σ. Kad to išvengtume, darbe yra pasiūlyta pradinė imties standartizacijos procedūra

$$Y_k = \hat{S}^{-1/2}(X_k - \bar{X}),\ k = 1, ..., n, \tag{16}$$

čia $\bar{X}$ ir $\hat{S}$ yra parametrų a ir Σ maksimalaus tikėtinumo įverčiai.

Transformuota imtis asimptotiškai turės p-matį standartinį Gauso skirstinį. Kriterijaus statistikos

$$T_n = -n \int_{R^{2p}} L(x, y)\, d(F_n(x) - \Phi(x))\, d(F_n(y) - \Phi(y)), \tag{17}$$

čia $F_n(x)$ yra empirinė pasiskirstymo funkcija, sukonstruota imties $Y_1, \ldots, Y_n$ pagrindu, $\Phi(x)$ – standartinio p-mačio Gauso dėsnio pasiskirstymo funkcija, didelių reikšmių atveju hipotezė H_0 turi būti atmesta.

Bendras $Y_1, \ldots, Y_n$ skirstinys asimptotiškai nepriklauso nuo parametrų a ir Σ. Todėl statistikos T_n pasiskirstymo kvantilius galima įvertinti Monte Karlo imitavimu.

3. Neparametriniai N-metrikiniai testai

3 skyrius yra skirtas N-metrikų teorijos pritaikymui statistinių homogeniškumo (**3.1 pokyris**), tolygumo hipersferoje S^{p-1} (**3.2 poskyris**), simetrijos bei nepriklausomumo (**3.3 poskyris**) hipotezių tikrinimo problemoms spręsti.

Homogeniškumo kriterijus

Tegul $X_1, \ldots, X_n$ ir $Y_1, \ldots, Y_m$ yra atsitiktinių dydžių X ir Y su nežinomomis tolydžiomis pasiskirstymo funkcijomis $F(x)$ ir $G(x)$ nepriklausomi stebiniai. Nulinė homogeniškumo hipotezė turi pavidalą $H_0 : F(x) = G(x)$.

Kriterijaus statistika, skirta H_0 tikrinti, yra sukonstruota naudojant N-atstu-

mą tarp dviejų empirinių skirstinių

$$T_{n,m} = -\frac{nm}{n+m} \int_{\mathbb{R}^{2p}} L(x,y)\, d(F_n(x) - G_m(x))\, d(F_n(y) - G_m(y)), \quad (18)$$

čia $F_n(x)$, $G_m(x)$ yra empirinės pasiskirstymo funkcijos, sudarytos imčių $X_1, \ldots, X_n$ ir $Y_1, \ldots, Y_m$ pagrindu.

3.1.1 poskyris yra skirtas pasiūlytos statistikos asimptotiniam skirstiniui nulinės hipotezės ir alternatyvos atvejais. Ribinis statistikos (18) skirstinys yra nustatytas taikant V-statistikų teoriją.

Tegul $L(x,y)$ yra N-metrikos stipriai neigiamai apibrėžtas branduolys ir $x_1, x_2, y_1, y_2 \in \mathbb{R}^p$. Pažymėkime

$$H(x_1, y_1, x_2, y_2) := L(x_1, y_2) + L(x_2, y_1) - L(x_1, x_2) - L(y_1, y_2). \quad (19)$$

Tada statistika $T_{n,m}$ (18) gali būti perrašyta kaip V-statistika (Koroljuk ir Borovskich, 1994)

$$T_{n,m} = \frac{nm}{n+m} \int_{\mathbb{R}^{4p}} H(x_1, y_1, x_2, y_2) dF_n(x_1) dG_m(y_1) dF_n(x_2) dG_m(y_2). \quad (20)$$

Esant nulinei hipotezei, kai X ir Y turi tą patį skirstinį $F(x)$, V-statistikos (20) branduolys tenkina išsigimimo savybę

$$\mathbf{E}H(X, Y, x_2, y_2) = \mathbf{E}L(X, y_2) + \mathbf{E}L(x_2, Y) - \mathbf{E}L(X, x_2) - \mathbf{E}L(Y, y_2) = 0.$$

Tegul $X^{'}$ ir $Y^{'}$ yra nepriklausomos atsitiktinių dydžių X ir Y kopijos. Tarkime, kad $\mathbf{E}H^2(X, Y, X^{'}, Y^{'}) < \infty$. Tada pagal spektrinę teoremą erdvėje L_2 egzistuoja ortogonalių funkcijų seka $\psi_1, \psi_2, ...$, $\mathbf{E}\psi_j(X, Y) = 0$, $j \geq 1$, ir skaičių seka $\lambda_1, \lambda_2, ...$, $\forall i\, \lambda_i \in \mathbb{R}$, $\sum_{j=1}^{\infty} \lambda_j^2 = \mathbf{E}H^2(X, Y, X^{'}, Y^{'}) < \infty$, tokia, kad $\lim_{s\to\infty} \|H - H^s\|_{L_2}^2 = 0$, čia $H^s(x_1, y_1, x_2, y_2) = \sum_{j=1}^{s} \lambda_j \psi_j(x_1, y_1) \times \psi_j(x_2, y_2)$.

5 teorema. *Esant nulinei hipotezei bei aukščiau nurodytoms prielaidoms, asimptotinis $T_{n,m}$ skirstinys sutampa su atsitiktinio dydžio T skirstiniu,*

$$T = \sum_{j=1}^{\infty} \lambda_j \sigma_j^2 \zeta_j^2, \quad (21)$$

čia

$$\sigma_j^2 = \int_{\mathbb{R}^p} (\mathbf{E}\psi_{\mathbf{j}}(\mathbf{x_1}, \mathbf{Y}))^2 dF(x_1), \quad j = 1, 2, ...,$$

o ζ_j, $j = 1, 2, ...$, *yra nepriklausomi standartiniai Gauso atsitiktiniai dydžiai.*

Antroje **3.1.1 poskyrio** dalyje statistikos $T_{n,m}$ (18) asimptotinis skirstinys yra nagrinėjamas esant alternatyviai hipotezei. Šiuo atveju tikimybė atmesti nulinę hipotezę, kai testo dydis α yra fiksuotas, artėja prie 1, kai $n, m \to \infty$. Todėl mūsų statistiką $T_{n,m}$ nagrinėkime normalizuotą.

Tegul $\exists x : \ F(x) \neq G(x)$,

$$H(x_1, y_1, x_2, y_2) := \quad \frac{1}{2}\left[L(x_1, y_2) + L(x_2, y_1) + L(x_1, y_1) + L(x_2, y_2)\right] - \\ -L(x_1, x_2) - L(y_1, y_2).$$

Branduolys $H(x_1, y_1, x_2, y_2)$ tenkina simetrijos sąlygas pagal kintamuosius $x_1 \leftrightarrow x_2$ ir $y_1 \leftrightarrow y_2$, todėl statistika $T_{n,m}$ gali būti pateikta kaip V-statistika

$$T_{n,m} = \int_{\mathbb{R}^p}\int_{\mathbb{R}^p}\int_{\mathbb{R}^p}\int_{\mathbb{R}^p} H(x_1, y_1, x_2, y_2) dF_n(x_1) dG_m(y_1) dF_n(x_2) dG_m(y_2). \tag{22}$$

Tegul X, $X^{'}$ ir Y,$Y^{'}$ yra nepriklausomi atsitiktiniai dydžiai su pasiskirstymo funkcijomis $F(x)$ ir $G(x)$. Pažymėkime

$$a := \int_{\mathbb{R}^p}\int_{\mathbb{R}^p}\int_{\mathbb{R}^p}\int_{\mathbb{R}^p} H(x_1, y_1, x_2, y_2) dF(x_1) dG(y_1) dF(x_2) dG(y_2) \tag{23}$$

ir apibrėžkime funkcijas

$$g_1(x) = \mathbf{E}(H(X, Y, X^{'}, Y^{'})|X = x) - a,$$

$$g_2(x) = \mathbf{E}(H(X, Y, X^{'}, Y^{'})|Y = x) - a.$$

Tarkime, kad $\sigma_1^2 = \mathbf{E}g_1^2(X)$ ir $\sigma_2^2 = \mathbf{E}g_1^2(Y)$.

6 teorema. *Jeigu* $\mathbf{E}H^2 < \infty$ *ir* $\sigma_1^2 \neq 0$, $\sigma_2^2 \neq 0$, *tai*

$$(4\sigma)^{-1}(T_{n,m} - a) \xrightarrow{d} \zeta,$$

kai $\min(n, m) \to \infty$, $\frac{n}{m} \to const. \neq 0$, *čia* $\sigma^2 = \mathbf{E}(T_{n,m} - a)^2 = \frac{2}{n}\sigma_1^2 + \frac{2}{m}\sigma_2^2$ *ir* ζ *yra standartinis Gauso atsitiktinis dydis.*

Akivaizdu, kad esant nulinei hipotezei ribinis statistikos $T_{n,m}$ (18) skirstinys priklauso nuo bendro atsitiktinių dydžių X ir Y skirstinio F, kuris yra nežinomas. Šis faktas neleidžia panaudoti 5 teoremos rezultatų praktikoje. **3.1.2–3.1.4 poskyriuose** mes siūlome kai kuriuos metodus šiai problemai spręsti. **3.1.2 poskyryje** ribinio statistikos $T_{n,m}$ skirstinio problema yra aptariama vienamačiu atveju. Daugiamačiu atveju (**3.1.3 poskyris**) praktiškai nustatant pasiūlyto testo kritinę sritį siūloma naudoti butstrepo (bootstrap) arba perstatinių (permutation) metodus.

3.1.4 poskyryje daugiamačiu atveju pasiūlyta konstrukcija, nepriklausanti nuo skirstinio homogeniškumo testo.

Kiekvieną iš dviejų imčių $X \sim (X_1, \ldots, X_n)$ ir $Y \sim (Y_1, \ldots, Y_m)$ atsitiktinai padalykime į dvi lygias dalis ir kiekvieną iš dalių nagrinėkime kaip atskirą nepriklausomą imtį X, X' ir Y, Y'. Kriterijaus statistika dviejų imčių homogeniškumo hipotezei tikrinti bus

$$T_{n,m} = \frac{2}{mn} \sum_{i,j} L(X_i, Y_j) - \frac{1}{n^2} \sum_{i,j} L(X_i, X_j{}') - \frac{1}{m^2} \sum_{i,j} L(Y_i, Y_j'). \quad (24)$$

Asimptotinis $T_{n,m}$ skirstinys yra nustatytas 7 teoremoje.

7 teorema. *Esant nulinei hipotezei,*

$$\frac{T_{n,m}}{\sqrt{\mathbf{D}T_{n,m}}} \xrightarrow{d} \zeta$$

kai $\min(n, m) \to \infty$*, čia* ζ *yra Gauso standartinis atsitiktinis dydis.*

Tolygumo hipersferoje S^{p-1} ***tikrinimo testai***

3.2 poskyris skirtas N-metrikų teorijos taikymui tikrinant sferinių duomenų tolygumo hipotezę.

Tegul $X_1, ..., X_n$, $X_i \in \mathbb{R}^p$, $\|X_i\| = 1$, $i = 1, ..., n$, yra atsitiktinio dydžio X nepriklausomi stebiniai. Šios imties pagrindu norime patikrinti hipotezę, kad X turi tolygų skirstinį S^{p-1}.

Kriterijaus statistika, sukonstruota naudojant N-metriką su branduoliu $L(x, y)$, yra

$$T_n = n \left[\frac{2}{n} \sum_{i=1}^{n} \mathbf{E}_Y L(X_i, Y) - \frac{1}{n^2} \sum_{i,j=1}^{n} L(X_i, X_j) - \mathbf{E}L(Y, Y') \right], \quad (25)$$

čia Y, Y' – nepriklausomi atsitiktiniai dydžiai su tolygiu sferoje S^{p-1} skirstiniu.

3.2.1 poskyris yra skirtas asimptotiniam statistikos T_n (25) skirstiniui, esant nulinei hipotezei. Detaliau yra nagrinėjami apskritimo S^1 ir sferos S^2 atvejai. Šitais atvejais ribinis pasiūlytų testų elgesys yra nustatytas naudojant du metodus. Pirmas metodas yra pagrįstas suderinamumo testų metodikos, aprašytos **2 skyruje**, taikymu. S^1 ir S^2 kartu su duotosiomis imtimis ir kriterijaus statistikomis iš pradžių atitinkamai transformuojami į intervalą $\mathbb{R}^1$ erdvėje bei kvadratą $\mathbb{R}^2$ erdvėje. Po to, vietoj nulinės hipotezės tikrinimo hipersferoje S^{p-1}, $p = 2, 3$, tolygumo hipotezės yra tikrinamos atitinkamame intervale ir kvadrate $\mathbb{R}^{p-1}$, $p = 2, 3$, remiantis transformuotomis imtimis.

Antrame metode taikoma Gine teorija apie Sobolevo invariantinius toly- gumo testus kompaktiškose Rymano daugdarose (Gine, 1975; Jupp, 2005).

Toliau statistiką T_n su branduoliu $L(x, y) = \|x-y\|$ nagrinėsime apskritime bei sferoje su vienetiniu spinduliu. Nulinės hipotezės atveju statistikos T_n ribinis skirstinys nustatytas šiose teoremose.

8 teorema. *Jeigu $X_1, ..., X_n$ yra nepriklausomų stebinių iš apskritime S^1 su vienetiniu spinduliu tolygaus skirstinio imtis, tai*

$$\frac{\pi}{4} T_n \xrightarrow{d} \sum_{k=1}^{\infty} a_k^2 \chi_k^2, \tag{26}$$

čia χ_k^2 yra nepriklausomi atsitiktiniai dydžiai, turintys chi kvadrato skirstinį su dviem laisvės laipsniais,

$$a_k^2 = \frac{1}{2\pi} \int_0^{2\pi} (1 - \frac{\pi}{2} \sin \frac{x}{2}) \cos kx dx.$$

9 teorema. *Jeigu $X_1, ..., X_n$ yra nepriklausomų stebinių, tolygiai pasiskirsčiusių sferoje S^2 su vienetiniu spinduliu, imtis, tai*

$$\frac{3}{4} T_n \xrightarrow{d} \sum_{k=1}^{\infty} a_k^2 \chi_{2k+1}^2, \tag{27}$$

$$a_k^2 = \frac{1}{2} \int_0^{\pi} (1 - \frac{3}{2} \sin \frac{x}{2}) \sin x P_k(\cos x) dx, \tag{28}$$

čia χ_{2k+1}^2 yra nepriklausomi atsitiktiniai dydžiai, turintys chi kvadrato skirstinį su $2k + 1$ laisvės laipsniais ir $P_k(x)$ yra Ležandro polinomai.

3.3 poskyris skirtas N-metrikų teorijos taikymui simetriškumo nulio atžvilgiu vienamačiu atveju bei nepriklausomumo hipotezių tikrinimui dvimačiu atveju. Pasiūlytų kriterijaus statistikų asimptotinis skirstinys nulinės hipotezės atveju yra nustatytas ir sutampa su tam tikros Gauso atsitiktinių dydžių kvadra- tinės formos skirstiniu.

4. Testų galingumo palyginimas

4 skyriuje lyginami pasiūlytos N-metrikos bei kai kurie klasikiniai kriterijai. **4.1 poskyryje**, kaip palyginimo priemonė nagrinėjamas Bahaduro asimptotinis santykinis efektyvumas (Bahadur, 1960; Nikitin, 1995). Paprastumo dėlei mes apsiribojame tik neparametrine suderinamumo hipoteze vienamačiu atveju. **4.2 poskyryje** kriterijų galia yra lyginama Monte Karlo imitavimu. Be paprastos ir sudėtinės suderinamumo hipotezių yra analizuojami homogeniškumo testai vienamačiu ir daugiamačiu atvejais. Ištirtas platus alternatyvių hipotezių diapazonas.

Atliktas teorinis bei empirinis kriterijų palyginimas parodė, kad darbe pasiūlyti testai yra galingi klasikinių kriterijų konkurentai. N-metrikos testai yra suderinti prieš visas alternatyvas, turi paprastą skaičuojamąją išraišką bei palyginti gerą galią prieš bendras alternatyvas. N-metrikos branduolių pasirinkimo galimybė leidžia sukonstruoti jautresnį kriterijų prieš tam tikrų alternatyvų klasę nei klasikiniai kriterijai.

Bendrosios išvados

1. Remiantis N-metrikų teorija, sukonstruoti statistinių suderinamumo, homogeniškumo, simetriškumo bei nepriklausomumo hipotezių tikrinimo kriterijai yra suderinti.

2. Pasiūlytų N-metrikos kriterijaus statistikų asimptotinis skirstinys esant nulinei hipotezei sutampa su Gauso atsitiktinių dydžių begalinės kvadratinės formos skirstiniu. Alternatyvos atveju testinių statistikų ribinis skirstinys yra normalusis.

3. Atliktas teorinis bei empirinis kriterijų palyginimas parodė, kad darbe pasiūlyti testai yra galingi klasikinių kriterijų konkurentai. N-metrikos testai yra suderinti prieš visas alternatyvas, turi paprastą skaičiavimo išraišką bei palyginti gerą galią prieš bendras alternatyvas. N-metrikos branduolių pasirinkimo galimybė leidžia sukonstruoti jautresnį kriterijų tam tikrų alternatyvų klasei nei klasikiniai kriterijai.

4. Bendru atveju N-metrikos statistikos priklauso nuo imties skirstinio. Homogeniškumo hipotezės atveju pasiūlytos testinės statistikos skirstinio priklausomybės nuo įmčių skirstinio galima išvengti su butstrepo (angl. boot-

strap) arba perstatinių (angl. permutation) metodų pagalba.

5. Normalumo hipotezės bei neparametrinės suderinamumo hipotezės didelės dimensijos atveju, kai kriterijaus statistikos asimptotinį skirstinį išvesti analiziškai yra sunku, testo kritinę aibę galima nustatyti Monte Karlo metodu.

Nagrinėjama tematika turi dar nemažai vystymosi galimybių: kriterijaus statistikos asimptotinio skirstinio nustatymas parametrinės suderinamumo hipotezės daugiamačiu atveju, kai pasiskirstymų šeima nėra Gauso; N-metrikos branduolių optimalus pasirinkimo būdas ir t. t. Disertantas planuoja vėliau pratęsti N-metrikų bei statistinių hipotezių tikrinimo tematikos nagrinėjimą.

Autoriaus mokslinių publikacijų disertacijos tema sarašas

Pagrindiniai rezultatai atspausdinti šiuose straipsniuose:

1. Bakshaev, A. 2008. Nonparametric tests based on N-distances, *Lithuanian Mathematical Journal* 48(4): 368–379. ISSN 0363-1672 (ISI Master Journal List).

2. Bakshaev, A. 2009. Goodness of fit and homogeneity tests on the basis of N-distances, *Journal of Statistical Planning and Inference* 139 (11): 3750–3758. ISSN 0378-3758 (ISI Master Journal List).

3. Bakshaev, A. 2010. N-distance tests for composite hypothesis of goodness of fit, *Lithuanian Mathematical Journal* 50(1): 14–34. ISSN 0363-1672 (ISI Master Journal List).

4. Bakshaev, A. 2010. N-distance tests for uniformity on the hypersphere, priimtas spausdinimui į *Nonlinear Analysis, Modelling and Control*. ISSN 1392-5113.

Aprobavimas

Disertacijos rezultatai buvo pristatyti Lietuvos matematikų draugijos konferencijose (2008, 2009 m.), 8th Tartu Conference on Multivariate statistics (Tartu, Estija, 2007 m. birželio 26–29 d.), 22nd Nordic Conference on Mathematical Statistics, NORDSTAT (Vilnius, Lietuva, 2008 m. birželio 16–19 d.)

Disertacijos tema skaitytas pranešimas Maskvos valstybinio universiteto Matematikos ir mechanikos fakulteto seminare „Neparametrinė statistika ir laiko eilutė“ (Maskva, 2007 m., balandis), Matematikos ir informatikos instituto Tikimybių teorijos ir statistikos skyriaus bei Vilniaus Gedimino technikos universiteto Matematinės statistikos katedros seminaruose.

Trumpos žinios apie autorių

Aleksej Bakšajev gimė 1981 m. lapkričio 25 d. Jekaterinburge, Rusijoje.

1988–1999 m. mokėsi Visagino „Atgimimo“ gimnazijoje. 2000 m. baigė A. N. Kolmogorovo mokykla ir įstojo į Maskvos M. V. Lomonosovo universiteto mechanikos ir matematikos fakultetą. 2005 m. su pagyrimu (magna cum laude) baigė pagrindinių studijų matematikos programą ir įgijo matematikos magistro kvalifikacinį laipsnį. 2005–2009 m. – Matematikos ir informatikos instituto doktorantas.

STATISTICAL TESTS BASED ON N-DISTANCES

Scientific problem

In this thesis the problem of verification of classical statistical hypotheses of goodness of fit, homogeneity, symmetry and independence is investigated.

Actuality

In the classical statistical analysis of observations in various studies researchers usually begin their investigations by proposing a distribution for their observations. There are several reasons for that:

- The distribution of the sample data may throw a light on the process that generate the data, if a suggested model for the process is correct, the sample data follow a specific distribution.
- Parameters of the distribution may be connected with important parameters in describing the basic model.
- Knowledge of the distribution of the data allows for application of standard statistical testing and estimation procedures.

Sometimes such assumptions about the form of the distribution are made by analyzing the procedure by which the data was obtained or made arbitrarily, often from considerations of convenience in the statistical methods used. In any case there arises a need to check whether the chosen distribution is true.

The researcher may be interested in the question whether the distribution of observed data has a given fixed form (a simple hypothesis) or belongs to a certain family of distributions (composite hypothesis). In case of multivariate observations, in addition to goodness of fit problems, there arises the problem of testing the hypothesis of the independence of the components of the random vector being observed without knowing the precise form of the marginal distributions. Another class of problems is that of comparing two or several samples among

themselves. These are the so-called homogeneity tests, designed for testing the hypothesis that the samples obtained are identically distributed.

To solve these problems a large number of goodness of fit, homogeneity and independence procedures have appeared over the years, the choice of which is made depending on the structure of the observations, the hypothesis being tested, the efficiency of the test, etc. Choosing the most efficient test of several ones that are available to the researcher is regarded as one of the basic problems of statistics. However, it is well known that for a variety of problems arising in statistical theory and practice the uniformly most powerful tests are unknown. Therefore creation of new test procedures sensitive to a particular type of hypotheses remains actual and in our days.

Klebanov in (Zinger, Klebanov ir Kakosyan, 1989; Klebanov, 2005) introduced a new class of probability metrics - N-distances, which has many useful properties and therefore could be applied to obtaining new powerful and simply computable statistical tests. The construction of such criteria together with investigation of their properties become a topical problem after Klebanov's works.

Research object

This thesis is devoted to statistical criteria based on N-distances for testing classical statistical hypotheses of goodness of fit, homogeneity, symmetry and independence.

Aim and tasks

The main objectives of the thesis are connected with application of N-distance theory to testing classical statistical hypotheses of goodness of fit, homogeneity, independence and symmetry. In particularly, we focus on the following tasks:

- Construction of statistics based on N-metrics for testing mentioned hypotheses.
- Establishing the critical region of proposed criteria, obtaining the asymptotic distribution of test statistics under the null and alternative hypotheses.
- Comparison of proposed N-distance tests with some classical criteria using Asymptotic Relative Efficiency (ARE) by Bahadur (Bahadur, 1960; Nikitin, 1995).

In parallel to the theoretical results the empirical comparison of the power of proposed N-distance tests is investigated.

Research methods

Methods of mathematical statistics, general probability theory and stochastic processes are applied. The proofs of the limit behavior of proposed test statistics

are based on the theory of U-statistics (Lee, 1990; Koroljuk ir Borovskich, 1994) and the properties of the weak convergence of stochastic processes (Bulinskii and Shiryaev, 2005). All the results presented in empirical part of the thesis are produced by the means of Monte Carlo simulations done with the help of R statistical package.

Scientific novelty

Novelty of the results is closely related to the formulated aims and problems. Proposed methods extend, generalize and supplement the results of Klebanov in (Klebanov, 2005), Baringhaus and Franz in (Baringhaus and Franz, 2004) and Szekely and Rizzo in (Szekely ir Rizzo, 2005). In particular, proposed criteria and established asymptotic distributions of test statistics in the problems of goodness of fit, uniformity on the hypersphere, independence (in bivariate case) and symmetry (in univariate case) have not been earlier considered in statistical literature.

Practical value of the results

Proposed statistical criteria could be applied to the real data analysis problems connected with verification of the hypotheses about the considered sample of observations.

Defended propositions

- Propositions on the asymptotic distribution of goodness of fit (simple and composite hypotheses) test statistics based on N-distances under the null and alternative hypotheses.
- Construction and asymptotic behavior of the test statistic in the problem of uniformity on the hypersphere S^{p-1}.
- Propositions on the asymptotic distribution of two-sample test statistics based on N-distances under the null and alternative hypotheses; application of bootstrap and permutation procedures to determination of critical region of proposed tests; construction and asymptotic behavior of distribution-free two-sample test.
- Construction and asymptotic null distribution of tests statistic based on N-distances for criteria of symmetry about zero in univariate case and independence in bivariate case.
- Propositions on the computational form of tests statistics based on N-metrics with different strongly negative definite kernels.
- Comparison of proposed N-distance and classical nonparametric good-

ness of fit tests in univariate case by means of asymptotic relative efficiency by Bahadur.

The scope of the scientific work

The thesis consists of Introduction, four chapters, Conclusions, References and list of authors publications. The total scope of the dissertation is 148 pages.

General conclusions

1. Based on N-distances, the construction of statistical tests of goodness of fit, homogeneity, symmetry and independence were proposed.
2. In the general case the limit null distribution of N-metrics statistics coincides with the distribution of infinite quadratic form of Gaussian random variables. Under the alternative hypothesis, proposed tests statistics are asymptotically normal.
3. The results of the theoretical and empirical power comparison study show that N-metrics tests are powerful competitors to existing classical criteria, in the sense that they are consistent against all alternatives and have relatively good power against general alternatives compared with other tests. The possibility in the selection of the strongly negative definite kernel for N-distance allows to create the test more sensitive to particular type of alternative hypothesis.
4. In the general case proposed N-metrics statistics are not distribution-free. In case of homogeneity hypothesis to avoid this problem bootstrap and permutation approaches are suggested to be used.
5. For normality and nonparametric hypotheses of goodness of fit in high dimensional cases, when it is difficult from computational point of view to determine the limit null distribution of N-distance statistic analytically, the critical region of the test can be established by means of Monte Karlo simulations.

About the author

Aleksej Bakšajev was born in Ekaterinburg, Russia, on 25 of November, 1981.

He studied in "Atgimimo" gimnasium during 1988–1999. In 2000 Aleksej graduated from Kolmogorov school and entered Lomonosov Moscow State university, Faculty of Mechanics and Mathematics. Master of Science in Mathematics with magna cum laude diploma in 2005. in 2005–2009 – PhD student of Institute of Mathematics and Informatics.

Aleksej BAKŠAJEV

STATISTINIŲ HIPOTEZIŲ TIKRINIMAS,
NAUDOJANT N-METRIKAS

Daktaro disertacijos santrauka
Fiziniai mokslai, matematika (01P)

Aleksej BAKŠAJEV

STATISTICAL TESTS
BASED ON N-DISTANCES

Summary of Doctoral Dissertation
Physical Sciences, Mathematics (01P)

2010 01 13. 1,5 sp. l. Tiražas 70 egz.
Vilniaus Gedimino technikos universiteto
leidykla „Technika“,
Saulėtekio al. 11, LT-10223 Vilnius
http://leidykla.vgtu.lt
Spausdino UAB „Baltijos kopija“, Kareivių g. 13B,
09109 Vilnius, *http://www.kopija.lt*