Liet. matem. rink., T. 41, spec. nr., 451–456

Nauji skirstinių simetriškumo testai

Vytautas MANIUŠIS (VU)
e-mail: vtas@uosis.mif.vu.lt

Įvadas

Nagrinėsime statistikų klasę, pagrįstą empirine charakteristine funkcija

$$c_n(t) = \int_{-\infty}^{+\infty} \mathrm{e}^{itx}\mathrm{d}F_n(x) = \frac{1}{n}\sum_{k=1}^{n} \mathrm{e}^{itX_k},$$

arba ją atitinkančiu empiriniu charakteristiniu procesu

$$Y_n(t) = \sqrt{n}\big(c_n(t) - c(t)\big);$$

čia $X, X_1, \ldots, X_n$ – nepriklausomi vienodai pasiskirstę atsitiktiniai dydžiai, turintys pasiskirstymo funkciją $F(x)$ ir charakteristinę funkciją $c(t) = \int_{-\infty}^{+\infty} \mathrm{e}^{itx}\mathrm{d}F(x)$, be to $F_n(x) = \frac{1}{n}\sum_{k=1}^{n} \chi(X_k \leqslant x)$ – empirinė pasiskirstymo funkcija. Kadangi charakteristinė funkcija yra reali tada ir tiktai tada, kai atsitiktinis dydis yra simetriškas nulio atžvilgiu, tai atsitiktinių dydžių simetriškumui tikrinti natūralu naudoti tokią statistikų klasę:

$$S_n(q) = \int_{-\infty}^{+\infty} \big|\mathrm{Im}\, c_n(t)\big|^2 q(t)\mathrm{d}t; \tag{1}$$

čia $q(t)$ – funkcija, tenkinanti tam tikras sąlygas.

Tokio tipo statistika pasiūlyta [1] straipsnyje, kai $q(t)$ – simetrinė nulio atžvilgiu tankio funkcija. Straipsnyje taip pat nurodyti galimi tokios statistikos ribinio skirstinio radimo metodai, tačiau nėra palyginimo su kitais simetriškumo tikrinimo kriterijais.

Šis darbas žymia dalimi remiasi [2], [3] straipsnių rezultatais, kurie leidžia išplėsti galimų funkcijų $q(t)$ klasę, pvz., galima imti $q(t) = |t|^{-1-\alpha}$, $\alpha \in (0; 1)$.

Pirmoje straipsnio dalyje sudaromas naujas kriterijus atsitiktinio dydžio simetriškumui tikrinti. Kriterijus pagrįstas statistika iš (1) klasės, kai funkcija $q(t) = |t|^{-1-\alpha}$. Šiai statistikai žinoma ribinio skirstinio išraiška, tačiau ji yra sudėtingos struktūros, todėl naudojama butstrepo aproksimacija. Antroje dalyje aprašomas grafinis P-reikšmių metodas, ypač tinkantis neparametrinių statistikų palyginimui. Trečioje straipsnio dalyje, modeliuojant kompiuteriu, tiriamas (1) klasės statistikų elgesys, kai funkcija $q(t) = |t|^{-1-\alpha}$.

Šios statistikos lyginamos su ženklų kriterijaus statistika, kai turimas Koši arba normalusis skirstinys. Ženklų kriterijaus naudojimas palyginimui yra pagrįstas, kadangi yra žinomi atvejai, kai laikomi geresniais už ženklų kriterijų kriterijai, pvz. Vilkoksono, yra prastesni, kai alternatyvos yra su „ilgomis uodegomis" ([6], [7]), pvz., toks yra Koši skirstinys. Iš modeliavimo galima matyti kokiems α statistikos yra pakankamai „geros" ir kokiems α jos gali būti „geriausios".

1. Kriterijaus sudarymas

Imkime statistiką iš (1) klasės, o funkciją $q(t) = |t|^{-1-\alpha}$, $\alpha \in (0;1)$. Nesunku įsitikinti, kad šiuo atveju $S_n(q) = cS_{n,\alpha}$, kai

$$S_{n,\alpha} = \frac{1}{n^2} \sum_{j,k=1}^{n} \left(|X_j + X_k|^\alpha - |X_j - X_k|^\alpha\right), \tag{2}$$

$$c = \int_{-\infty}^{+\infty} \sin^2\left(\frac{t}{2}\right) |t|^{-1-\alpha}\, \mathrm{d}t. \tag{3}$$

Teorema ([2, 17 išvada]). *Jei atsitiktinis dydis $X \in \mathbb{R}$ yra simetriškas, tai statistika $nS_{n,\alpha}$ konverguoja pagal pasiskirstymą į atsitiktinį dydį*

$$S_\alpha = \int_{-\infty}^{+\infty} |\operatorname{Im} Y(t)|^2 |t|^{-1-\alpha}\, \mathrm{d}t,$$

jei patenkintos šios sąlygos:

$$\sum_{j=1}^{\infty} \frac{\sqrt{j}}{\rho(2^{-j})} \left(\mathbf{E} \sin^4\left(\frac{X}{2^{j+1}}\right)\right)^{\frac{1}{2}} < \infty, \tag{4}$$

$$\int_{-\infty}^{+\infty} \rho(t) |t|^{-1-\alpha}\, \mathrm{d}t < \infty. \tag{5}$$

Šioje teoremoje Y yra kompleksinis Gauso procesas, kurio vidurkiai $\mathbf{E}Y(t) = 0$ ir kovariacijos $\mathbf{E}Y(t)\overline{Y(s)} = c(t-s) - c(t)c(-s)$, $s, t \in [0;1]$,

$$Y(t) = \int_{-\infty}^{+\infty} \mathrm{e}^{itx}\, \mathrm{d}W(F(x)) - W(1)c(t), \quad \operatorname{Im} Y(t) = \int_{-\infty}^{+\infty} \sin(tx)\, \mathrm{d}W(F(x));$$

čia $W(t)$ – standartinis Vynerio procesas.

Nors teorema ir nurodo ribinį statistikos $nS_{n,\alpha}$ skirstinį, tačiau jis yra sudėtingas, todėl praktikoje tenka naudoti butstrepo aproksimaciją. Pastebėkime, kad $S_{n,\alpha}$ yra V-statistika. Klasikinis n iš n butstrepas šiai statistikai neveikia, tačiau galima naudoti m iš n butstrepą, kai $m = \mathrm{o}(n)$ ([4]).

Nesunku įsitikinti, kad (3) netiesioginis integralas konverguoja, kai $0 < \alpha < 2$.

1 pavyzdys. Imkime Koši skirstinį $C(\alpha, \beta)$; čia α – mediana, 2β – tarpkvartilinis atstumas,

$$F(x) = \frac{1}{2} + \frac{1}{\pi}\operatorname{arctg}\left(\frac{x-\alpha}{\beta}\right) - \text{pasiskirstymo funkcija},$$

$$f(x) = \frac{1}{\pi}\frac{\beta}{\beta^2 + (x-\alpha)^2} - \text{tankio funkcija}.$$

Kadangi Koši skirstinys turi p-uosius momentus, $0 \leqslant p < 1$, todėl sąlyga (4) tenkinama, kai $0 < \alpha < 0,5$.

2. P-reikšmės

Jei atsitiktinio dydžio T pasiskirstymo funkcija $F(x)$, tai atsitiktinis dydis $F(T)$ pasiskirstęs tolygiai intervale $[0; 1]$. Pasinaudosime [5] straipsnyje išdėstyta schema. Tarkime, turime statistiką T ir n šios statistikos realizacijų τ_k, $k = 1, \ldots, n$. P-reikšmę apibrėšime kaip tikimybės patekti į kritinę sritį funkciją. Vienpusio kriterijaus atveju ši tikimybė $p(t) = \mathbf{P}(T \geqslant t) = 1 - \mathbf{P}(T < t) = 1 - F(t)$. Paprastai, kai reikšmingumo lygmuo yra a, statistikos T kritinė reikšmė apibrėžiama kaip didžiausias skaičius t_a, kuriam $\mathbf{P}(T \geqslant t_a) \leqslant a$. Nulinę hipotezę H_0 atmetame, jei statistikos T realizacija $\tau \geqslant t_a$. Tačiau galima imti realizacijos τ P-reikšmę $p(\tau) = 1 - F(\tau)$ ir hipotezę H_0 atmesti, jei $p(\tau) \leqslant a$. Kiekvienai statistikos T realizacijai τ_k, $k = 1, \ldots, n$ galime rasti P-reikšmę $p_k = p(\tau_k) = 1 - F(\tau_k)$. Kadangi $1 - F(T)$ – tolygiai pasiskirstęs intervale $[0; 1]$ atsitiktinis dydis, tai nubrėžus p_k empirinės pasiskirstymo funkcijos

$$\hat{F}(x_i) = \frac{1}{n}\sum_{k=1}^{n} \chi(p_k \leqslant x_i), \quad x_i \in (0; 1)$$

grafiką $(x_i, \hat{F}(x_i))$, galima jį palyginti su 45^o nuolydžio tiese.

Pakankamai geras statistikas, kurių pasiskirstymo funkcijos artimos tolygiajai pasiskirstymo funkcijai, geriau galima ištirti naudojant P-reikšmių skirtumo grafiką $(x_i, \hat{F}(x_i) - x_i)$.

Trečias grafikų tipas skirtas kriterijų galios palyginimui. Tarkime, turime P-reikšmių empirines pasiskirstymo funkcijas $\hat{F}_0(x_i)$ ir $\hat{F}_1(x_i)$, atitinkamai nulinės hipotezės H_0 ir alternatyvos H_1 atvejais. Tada $(\hat{F}_0(x_i), \hat{F}_1(x_i))$ yra galios grafikas. Taškus x_i galima imti, pavyzdžiui, tokius (221 taškas, su papildomais taškais intervalo kraštuose):

$$x_i = 1 \times 10^{-9}; 1 \times 10^{-8}; \ldots; 1 \times 10^{-3}; 2 \times 10^{-3}; 1 \times 10^{-2}; 1,5 \times 10^{-2}; \ldots;$$
$$9,9 \times 10^{-1}; 9,91 \times 10^{-1}; 9.99 \times 10^{-1}.$$

3. Modeliavimas

Lyginami kriterijai, pagrįsti statistika $S_{n,\alpha}$, ir ženklų kriterijus. Iš modeliavimo kompiuteriu matyti, kad pakankamai reikšmingi rezultatai gaunami, kai statistikos realizacijų skaičius r, butstrepo kartojimų skaičius b ir imties dydis n būna apie 1000.

Koši skirstinio $C(\alpha, \beta)$ atveju tarkime, kad žinomas tarpkvartilinis atstumas ir $\beta = 1$. Šio skirstinio simetriškumas tapatus tam, kad mediana $\alpha = 0$. Todėl hipotezė H_0 bus, kad atsitiktinio dydžio skirstinys yra $C(0; 1)$. Alternatyva imkime $C(0, 1; 1)$. Imant didesnes medianas hipotezė ir alternatyva pasidaro lengvai atskiriamos. Modeliavimas atliktas, kai $(r, b, n = 1000)$ $\alpha = 0,1; 0,15; \ldots; 0,45; 0,49$. 1 lentelėje matyti, kaip skirtumas $\max_i |\hat{F}_0(x_i) - x_i|$ priklauso nuo α.

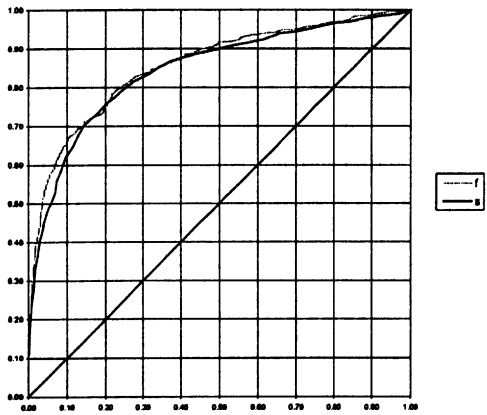
1 lentelė

α	$\max_i \lvert\hat{F}_0(x_i) - x_i\rvert$
0,1	0,096
0,15	0,075
0,2	0,06
0,25	0,05
0,3	0,05
0,35	0,054
0,4	0,058
0,45	0,059
0,49	0,075

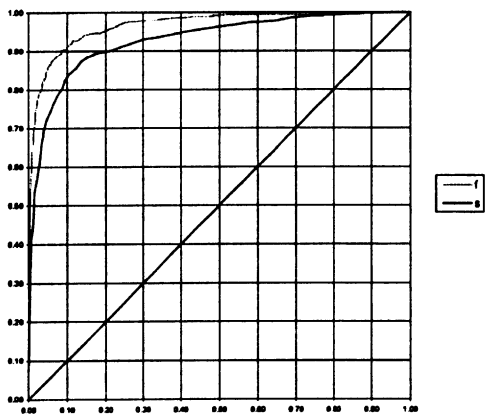
Matome, kad mažiausias skirtumas gaunamas, kai α yra apie 0, 25. Galios grafikas, kai $\alpha = 0, 25$ pateiktas 1 pav.

Normaliojo skirstinio $\mathcal{N}(\mu, \sigma^2)$ atveju tarkime, kad žinoma dispersija ir $\sigma = 1$. Šio skirstinio simetriškumas tapatus tam, kad vidurkis $\mu = 0$. Todėl hipotezė H_0 bus, kad atsitiktinio dydžio skirstinys yra $\mathcal{N}(0; 1)$. Alternatyva imkime $\mathcal{N}(0, 1; 1)$. Imant didesnes medianas hipotezė ir alternatyva pasidaro lengvai atskiriamos. Modeliavimas atliktas, kai $(r, b, n = 1000)$ $\alpha = 0,1; 0,2; \ldots; 1,9; 1,99$. 2 lentelėje matyti, kaip skirtumas $\max_i |\hat{F}_0(x_i) - x_i|$ priklauso nuo α.

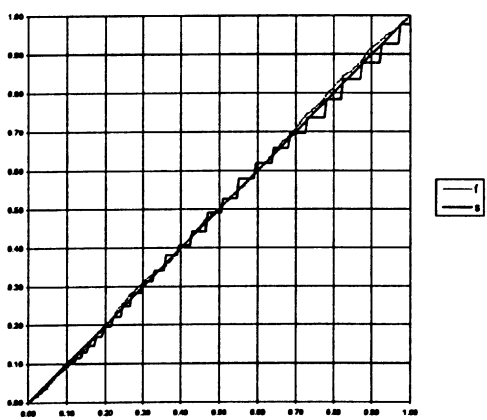
Matome, kad intervale $0 < \alpha < 1$ mažiausias skirtumas gaunamas, kai α yra apie 0, 7. Galios grafikas, kai $\alpha = 0, 7$, pateiktas 2 pav. Intervale $1 \leqslant \alpha < 2$ mažiausias skirtumas gaunamas, kai α yra apie 1, 5. P-reikšmių grafikas, kai $\alpha = 1, 5$ pateiktas 3 pav. Galios grafikas, kai $\alpha = 1, 5$ pateiktas 4 pav. Įdomu yra tai, kad mažiausias skirtumas $\max_i |\hat{F}_0(x_i) - x_i|$ normaliojo skirstinio atveju gaunamas , kai α yra apie 1, 5, ir apskritai modeliavimas rodo, kad panašiems į normalųjį skirstiniams galimas konvergavimas ir intervale $1 \leqslant \alpha < 2$. Kaip matome iš grafikų, nauji skirstinių simetriškumo testai dažnai būna geresni už ženklų kriterijų.



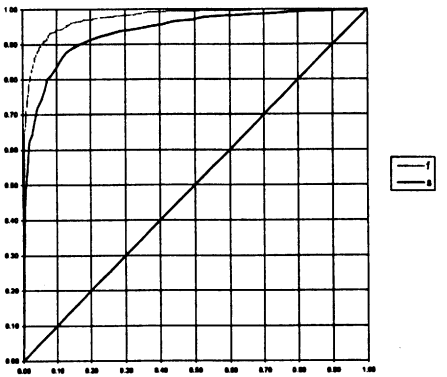
1 pav.



2 pav.



3 pav.



4 pav.

2 lentelė

α	$\max_i \lvert\hat{F}_0(x_i) - x_i\rvert$	α	$\max_i \lvert\hat{F}_0(x_i) - x_i\rvert$
0,1	0,093	1,1	0,023
0,2	0,087	1,2	0,024
0,3	0,05	1,3	0,033
0,4	0,026	1,4	0,033
0,5	0,029	1,5	0,018
0,6	0,027	1,6	0,024
0,7	0,022	1,7	0,035
0,8	0,03	1,8	0,021
0,9	0,03	1,9	0,035
1,0	0,032	1,99	0,035

Literatūra

[1] A. Feuerverger, R.A. Mureika, The empirical characteristic function and its applications, *Ann. Statist.*, **5**(1), 88–97 (1977).

[2] A. Račkauskas, Ch. Suquet, Hölder convergences of multivariate empirical characteristic functions, *Pub. IRMA Lille*, **54**(6), *Preprint* (2001).

[3] A. Račkauskas, Ch. Suquet, Central limit theorem in Hölder spaces, *Probab. and Math. Statist.*, **19**(1), 133–152 (1999).

[4] F. Götze, A. Račkauskas, Adaptive choice of bootstrap sample sizes, Sonderforschungsbereich 343 "Diskrete Strukturen in der Mathematik", 99–071, *Preprint* (1999).

[5] R. Davidson, J.G. MacKinnon, Graphical methods for investigating the size and power of hypothesis tests, *Revised Version of Queen's Institute for Economic Research Discussion Paper No. 903* (1994).

[6] J. Klotz, Alternative efficiencies for signed rank tests, *Ann. Math. Statist.*, **36**(6) 1759–1766 (1965).

[7] H.J. Arnold, Small sample power of the one sample Wilcoxon test for non-normal shift alternatives, *Ann. Math. Statist.*, **36**(6), 1767–1778 (1965).

New tests for symmetry

V. Maniušis

A new class of tests for symmetry of random variables is considered. This class is based on the empirical characteristic function. Tests from this class are compared with the classical Sign test using graphical display of P values. It is easy to extend this approach to the multivariate case.