Lietuvos matematikos rinkinys
Lietuvos matematikų draugijos darbai, ser. B
53 t., 2012, 56–61

ISSN 0132-2818
www.mii.lt/LMR/

Nepusiausvyro perdavimo testo galios aproksimavimas

Šarūnas Germanas[1], **Audronė Jakaitienė**[2]

[1] *Vilniaus universitetas, Matematikos ir informatikos fakultetas*
Naugarduko g. 24, LT-03225 Vilnius

[2] *Vilniaus universitetas, Matematikos ir informatikos institutas*
Akademijos g. 4, LT-08663 Vilnius

E. paštas: sarunas.germanas@mf.vu.lt, audrone.jakaitiene@mii.vu.lt

Santrauka. Straipsnio tikslas yra pasiūlyti naują nepusiausvyro perdavimo testo (NPT) galios aproksimaciją Puasono skirstiniu. Parodoma, kad NPT aproksimacija Puasono skirstiniu yra tinkamesnė nei normaliuoju skirstiniu, kai binominio skirstinio sėkmės tikimybė p artėja į nulį, o imties dydis n artėja į begalybę. Palyginus siūlomo metodo ir kitų autorių rezultatus, gauta, kad kai galios reikšmė yra tarp 0.8 ir 0.95, o imties dydis didesnis arba lygus 50, siūlomas metodas turi mažesnes paklaidas nei normaliosios aproksimacijos modelis.

Raktiniai žodžiai: nepusiausvyras perdavimo testas, galia, Puasono skirstinys.

Įvadas

Nepusiausvyro perdavimo testas (angl. transmission disequilibrium test) buvo pirmą kartą aprašytas 1993 metais Spielman ir kt. [10]. Spielman, aprašydamas NPT, panaudojo McNemaro kriterijų [8]. NPT yra plačiai taikomas ieškant asociacijų tarp genetinio žymens ir fenotipo (ligos) remiantis šeimų analize. Jei tiriamoje populiacijoje nėra asociacijos tarp geno ir genetinio žymens yra teigiama, kad abi genetinės sritys (geno ir žymens) yra sankibos pusiausvyroje. Situacija, kurioje egzistuoja asociacija tarp geno ir genetinio žymens, yra žinoma kaip nepusiausvyra sankiba [1]. Esant nepusiausvyrai sankibai, galima įtarti, kad šioje sankiboje esantys genetiniai žymenys tiksliai reprezentuoja tiriamą ligą. Dėl to galima nustatyti polinkį į šią ligą individui, kuriam liga dar nepasireiškė. Tokiu būdu galima rekomenduoti imtis profilaktinių priemonių, kad būtų galima išvengti ligos pasireiškimo. Dėl patobulėjusių genomo tyrimo metodų, pastaruoju metu susidomėjimas nepusiausvyra sankiba labai padidėjo. Nors surinkti genetinius duomenis šeimomis yra sudėtingiau nei iš atsitiktinių individų.

NPT galia yra testo parametras, kuris parodo, kokia tikimybė, kad bus atmesta neteisinga hipotezė. Labai svarbu paskaičiuoti šį parametrą, kadangi jis parodo testo patikimumą. NPT galia dažniausiai buvo skaičiuojama aproksimuojant ją normaliuoju skirstiniu. Didžiausia problema buvo paskaičiuoti normaliojo skirstinio vidurkį ir dispersiją. NPT galia buvo aproksimuota skirtingiems paveldėjimo modeliams [11, 5]. Pasiūlytas metodas, kuris nereikalavo nepriklausomumo tarp tėvų alelių perdavimų (įtraukė tėvų heterozigotiškumo savybę) [4]. McGinnis pateikė imties dydžio skaičiavimo formulę, paremtą normaliuoju skirstiniu, kai yra duota galia. Taip pat šis

autorius pateikė alelio paveldėjimo tikimybių formules [7]. Tam, kad būtų galima paskaičiuoti imties dydį, reikalingą pasiekti galią lygią 0.8, buvo paskaičiuotas NPT statistikos vidurkis [3]. Tokiu būdu buvo paskaičiuotas NPT necentriškumo parametras. Taip pat galios aproksimavimas buvo patikrintas skaičiuojant empirinę galią [5, 2, 4]. Buvo pasiūlytas naujas NPT statistikos variantas, kurio galia buvo palyginta su Spielman pasiūlyta NPT statistika naudojant empirinės galios skaičiavimą [6].

Siūlomame metode atsisakoma aproksimacijos normaliuoju skirstiniu, jį keičiant Puasono skirstiniu, nes NPT statistika gaunama transformuojant binominį atsitiktinį dydį. Normalusis skirstinys pakankamai tiksliai aproksimuoja binominį skirstinį, kai imties dydis (n) didelis, o sėkmės tikimybė (p) artima 0.5. Kai p artėja į 0, o sandauga np išlieka pastovi, Puasono skirstinio paklaidų modulių vidurkis tampa reikšmingai mažesnis nei normaliosios aproksimacijos [9].

1 Nepusiausvyro perdavimo testas ir jo galios aproksimavimas normaliuoju skirstiniu

Genetiniams duomenims, sudarytiems iš triadų, kurias sudaro du sveiki tėvai ir vienas sergantis vaikas, taikome NPT. Tikrinama nulinė hipotezė, kad nėra asociacijos tarp genetinio žymens ir fenotipo, t. y. alelių paveldėjimo tikimybė lygi $\frac{1}{2}$. Tokios hipotezės įvertinimui yra naudojami tik heterozigotinių tėvų duomenys – jų alelių perdavimo dažniai (b ir c reikšmės, 1 lentelė). NPT statistikai įvertinti naudojamas McNemaro kriterijus [8].

Visų alelių skaičius (paveldėtų iš heterozigotinių ir homozigotinių tėvų) žymimas $2n$, kai m žymėsime tik iš heterozigotinių tėvų paveldėtų alelių skaičių, tai yra $b + c = m$. Straipsnyje toliau naudosime parametrą $p = \frac{b}{b+c}$ – alelio, paveldėto iš heterozigotinio tėvo, perdavimo santykinį dažnį – kaip NPT statistikos reikšmingumo matą.

Apibrėžiame NPT statistiką (1), kuri yra apytiksliai pasiskirsčiusi pagal χ kvadratų skirstinį su vienu laisvės laipsniu, kai b reikšmė yra pakankamai artima c reikšmei:

$$\chi_1^2 = \frac{(b - \frac{b+c}{2})^2}{\frac{b+c}{2}} + \frac{(c - \frac{b+c}{2})^2}{\frac{b+c}{2}} = \frac{(b-c)^2}{b+c}. \tag{1}$$

Šaknis iš NPT statistikos apytiksliai pasiskirsčiusi pagal standartinį normalųjį skirstinį, kai daroma prielaida, kad alelio perdavimo tikimybė p yra artima 0.5. Turint

1 lentelė. Perduoti ir neperduoti genetinio žymens aleliai $A1$ ir $A2$ iš tėvų ($2n$) sergantiems vaikams (n).

		Neperduoti aleliai		
		$A1$	$A2$	Iš viso
Perduoti aleliai	$A1$	a	b	$a+b$
	$A2$	c	d	$c+d$
	Iš viso	$a+c$	$b+d$	$2n$

a – kiek kartų $A1A1$ perdavė $A1$ sergančiam vaikui,
b – kiek kartų $A1A2$ perdavė $A1$ sergančiam vaikui,
c – kiek kartų $A1A2$ perdavė $A2$ sergančiam vaikui,
d – kiek kartų $A2A2$ perdavė $A2$ sergančiam vaikui.

pasikliovimo lygmenį α, gali būti paskaičiuota kritinės srities ribos reikšmė $x_{\frac{\alpha}{2}}$:

$$x_{\frac{\alpha}{2}} = \Phi^{-1}\left(\frac{\alpha}{2}\right), \tag{2}$$

kur Φ^{-1} žymi atvirkštinę standartinio normaliojo skirstinio pasiskirstymo funkciją arba kvantilių funkciją. Toliau, naudojant minėtą kvantilį, pateikiama NPT galios aproksimacija, kurią apskaičiavome iš McGinniso modelio [7]. Iš minėtame straipsnyje pateiktos imties dydžio formulės išreiškiame galią:

$$Galia \approx \Phi\left(\frac{x_{\frac{\alpha}{2}} - \sqrt{m}(2p-1)}{2\sqrt{p(1-p)}}\right) + \Phi\left(\frac{x_{\frac{\alpha}{2}} + \sqrt{m}(2p-1)}{2\sqrt{p(1-p)}}\right), \tag{3}$$

kur Φ yra standartinio normaliojo skirstinio pasiskirstymo funkcija, p yra apibrėžtas kaip alelio $A1$ perdavimo santykinis dažnis.

2 NPT galios aproksimavimas Puasono skirstiniu

Puasono skirstinys tiksliai aproksimuoja binominį skirstinį, kai parametro p reikšmė artėja į nulį, o tuo pačiu sandauga np lieka pastovi, todėl n turi artėti į begalybę tokiu pat greičiu, kaip p artėja į nulį. Jeigu aproksimuojama NPT galia, kai alelio perdavimo santykinis dažnis artėja į nulį, tada n neartėja į begalybę. Pastarosios situacijos NPT galios aproksimavimo sąlygos yra labiau tinkamos Puasono skirstiniui nei normaliajam.

Straipsnyje siūloma NPT statistikos aproksimacija Puasono skirstiniu. Aproksimuojama pati NPT statistika, o ne jos šaknis kaip normaliosios aproksimacijos atveju, nes nagrinėjami neneigiami sveikieji skaičiai.

Teisingos nulinės hipotezės atveju (santykinis alelio perdavimo dažnis $p = 0.5$) šaknis iš NPT statistikos apytiksliai yra pasiskirsčiusi pagal normalųjį skirstinį su vidurkiu $\mu = 0$ ir dispersija $\sigma^2 = 1$. Pati NPT statistika yra apytiksliai pasiskirsčiusi pagal centrinį χ kvadratų skirstinį su 1 laisvės laipsniu. Tada hipotezės atmetimo srities riba apskaičiuojama:

$$x_{1-\alpha} = F^{-1}_{\chi 1;0}(1-\alpha), \tag{4}$$

kur $F^{-1}_{\chi 1;0}$ yra atvirkštinė χ kvadratų pasiskirstymo funkcija arba kvantilių funkcija.

Esant teisingai alternatyviai hipotezei (alelio perdavimo santykinis dažnis p nėra lygus 0.5), NPT statistika aproksimuojama Puasono skirstiniu su parametru λ, kuris turi būti lygus NPT statistikos vidurkiui. NPT statistikos vidurkis paskaičiuojamas taip:

$$\lambda = E(NPT) = E\left(\frac{(2b-m)^2}{m}\right) = \frac{4}{m}E(b^2) - 4Eb + m, \tag{5}$$

$$\lambda = \frac{4}{m}\big(m(m-1)p^2 + mp\big) - 4mp + m = 4(m-1)\big(p^2 - p\big) + m - 1 + 1, \tag{6}$$

$$\lambda = (m-1)\big(4p^2 - 4p + 1\big) + 1 = (2p-1)^2(m-1) + 1. \tag{7}$$

NPT galia aproksimuojama Puasono skirstiniu:

$$Galia \approx 1 - F_{Poi(\lambda)}(x_{1-\alpha}) = 1 - F_{Poi((2p-1)^2(m-1)+1)}\big(F^{-1}_{\chi 1;0}(1-\alpha)\big). \tag{8}$$

Gavome paprastą formulę NPT galios skaičiavimui. Esminis šios formulės skirtumas nuo (3) formulės yra kitokia pasiskirstymo funkcija. Belieka patikrinti, kuo skiriasi minėti metodai savo rezultatų tikslumu.

Normaliuoju skirstiniu pagrįsti NPT galios aproksimavimai yra tikslūs (paklaida mažesnė už 0.03), kai alelių perdavimo santykiniai dažniai artėja į 0.5 ir tuo pačiu imties dydis yra didelis. Kai alelių santykiniai dažniai artėja į 0 arba į 1, ir imties dydis mažėja, tai minėtos aproksimacijos akivaizdžiai pradeda nuklysti nuo teisingų reikšmių (paklaida viršyja 0.07). Tai bus parodyta vėliau paklaidų lentelėse.

3 Rezultatų analizė

Norint patikrinti siūlomo metodo tinkamumą, skaičiuojama NPT galios aproksimacija generuotiems duomenims bei gautos paklaidos lyginamos su McGinnis'o metodo paklaidomis, nes siekiama palyginti NPT galios aproksimaciją normaliuoju ir Puasono skirstiniais. O McGinnis'o aproksimacija yra pagrįsta normaliuoju skirstiniu. Fiksuotam imties dydžiui n, skaičiuojama galia skirtingiems alelio perdavimo santykiniams dažniams $p = 0.01 \cdot i$, $i = 1, 2, \ldots, 100$.

Dėl didelio žymenų kiekio galios skaičiavimui pasirenkamas pasikliovimo lygmuo $\alpha = 10^{-7}$. Panašios reikšmės pasirenkamos ir kitų autorių [3, 5].

Abiejų metodų (mūsų ir McGinnis'o) paklaidos yra vertinamos skaičiuojant nukrypimus nuo empirinės galios reikšmių. Empirinė galia yra skaičiuojama generuojant 10^6 atsitiktinių dydžių pasiskirsčiusių pagal NPT statistiką prie nulinės hipotezės ir atitinkamai tiek pat prie alternatyvios hipotezės. Tada skaičiuojamas $1 - \alpha$-asis empirinis kvantilis pirmajam duomenų rinkiniui, gaunama kritinės srities riba. Skaičiuojama, kiek kartų antrojo duomenų rinkinio atsitiktinių dydžių patenka į kritinę sritį. Patenkančių į kritinę sritį atsitiktinių dydžių skaičius padalintas iš viso atsitiktinių dydžių skaičiaus yra empirinė galia.

Antroje lentelėje matome, kad visais atvejais siūlomo autorių metodo paklaidos yra didesnės nei McGinnis'o metodo. Pastarieji rezultatai gauti skaičiuojant paklaidas visoms galios reikšmėms. Visgi, dauguma galios reikšmių yra mažos, ypač esant nedidelėms imties dydžio reikšmėms (50 ar 100). Todėl 3-oje lentelėje pateikiamos paklaidos, kai empirinė galia yra didelė – tarp 0.8 ir 0.95.

Kaip matome iš lentelės, kai imties dydis yra lygus 100, siūlomas metodas prie nurodytų sąlygų turi daugiau nei du kartus mažesnį paklaidos vidurkį. Naujo metodo standartinis nuokrypis tuo pačiu atveju taip pat yra mažesnis. Kitais atvejais

2 lentelė. Galios aproksimacijos paklaidų palyginimas. Naudota 10^6 generuotų atsitiktinių dydžių.

	Imties dydis			
	$m = 50$	$m = 100$	$m = 200$	$m = 400$
McGinnis modelis				
Vidurkis	0.0737473	0.04683858	0.02715634	0.01800833
SN	0.1126549	0.0773375	0.05434401	0.04434304
Autorių modelis				
Vidurkis	0.08710264	0.0507079	0.03298077	0.01855796
SN	0.1270005	0.092885	0.0721331	0.0511473

3 lentelė. Galios aproksimacijos paklaidų palyginimas. Naudota 10^6 generuotų atsitiktinių dydžių. Empirinė galia yra tarp 0.8 ir 0.95.

	Imties dydis			
	$m = 50$	$m = 100$	$m = 200$	$m = 400$
McGinnis modelis				
Vidurkis	0.2358666	0.1524292	0.09040485	0.125796
SN	0.1144312	0.0845436	0.05328105	0.07438322
Autorių modelis				
Vidurkis	0.2106943	0.07213755	0.03291665	0.04654384
SN	0.1083862	0.07937233	0.02134065	0.0209857

siūlomo metodo paklaidų vidurkis ir standartinis nuokrypis yra taip pat mažesni už McGinnis'o modelio.

4 Išvados ir pasiūlymai

Šiame straipsnyje buvo parodyta, kad kai imties dydis didesnis arba lygus 50, galia yra tarp 0.8 ir 0.95, Puasono skirstinys labiau nei normalusis skirstinys tinka NPT galios aproksimavimui. Imties dydis ir galia priklauso nuo alelio perdavimo santykinio dažnio ir pasikliovimo lygmens. Ateityje būtų naudinga išvesti formulę imties dydžiui skaičiuoti, kai yra duota galia. Tam reikia išreikšti Puasono pasiskirstymo funkcijos parametrą λ per galią ir $x_{1-\alpha}$.

Literatūra

[1] D.J. Balding, M. Bishop and C. Cannings. *Handbook of Statistical Genetics*. Wiley-Interscience, 2007.

[2] W.-M. Chen and H.-W. Deng. A general and accurate approach for computing the statistical power of the transmission disequilibrium test for complex disease genes. *Genetic Epidemiology*, **21**:53–67, 2001.

[3] D.M. Evans, A.P. Morris, L.R. Cardon and P.C. Sham. A note on the power to detect transmission distortion in parentchild trios via the transmission disequilibrium test. *Behav. Genet.*, **36**:947–950, 2006.

[4] M.M. Iles. On calculating the power of a TDT study – comparison of methods. *Ann. Hum. Genet.*, **66**:323–328, 2002.

[5] M. Knapp. A note on power approximations for the transmission/disequilibrium test. *Am. J. Hum. Genet.*, **64**:1177–1185, 1999.

[6] D. Londono, S. Buyske, S.J. Finch, S. Sharma, C.A. Wise and D. Gordon. TDT-HET: A new transmission disequilibrium test that incorporates locus heterogeneity into the analysis of family-based association data. *BMC Bioinf.*, **13**(13), 2012.

[7] R. McGinnis. General equations for p_t, p_s, and the power of the TDT and the affected-sib-pair test. *Am. J. Hum. Genet.*, **67**:1340–1347, 2000.

[8] Q. McNemar. Note on the sampling error of the difference between correlated proportions or percentages. *Psychometrika*, **12**(2):153–157, 1947.

[9] M.S. Raff. On approximating the point binomial. *J. Am. Stat. Ass.*, **51**(274):293–303, 1956.

[10] R.S. Spielman, R.E. McGinnis and W.J. Ewens. Transmission test for linkage disequilibrium: The insulin-dependent diabetes mellitus (iddm). *Am. J. Hum. Genet.*, **52**:506–516, 1993.

[11] I.V. Zorkoltseva and T.I. Axenovich. Analysis of allelic association: estimation of the power of the TDT. *Russian Journal of Genetics*, **39**(8):948–954, 2003.

SUMMARY

Power approximation of the transmission disequilibrium test

S. Germanas, A. Jakaitienė

In this paper we apply Poisson distribution in order to approximate the power of the Transmission Disequilibrium Test (TDT). In this research we calculated the power of the TDT for different values of sample size n and different values of allele transmission frequencies p. The research showed that Poisson approximation have smaller errors than Normal approximation when the power is between 0.8 and 0.95 and sample size m is equal or bigger than 50.

Keywords: transmission disequilibrium test, power, Poisson distribution.