

VILNIAUS GEDIMINO TECHNIKOS UNIVERSITETAS

Pavel SAMUSENKO

NEPARAMETRINIAI KRITERIJAI
RETŲ ĮVYKIŲ DAŽNIŲ LENTELĖMS

DAKTARO DISERTACIJOS SANTRAUKA

FIZINIAI MOKSLAI,
MATEMATIKA (01P)


Vilnius LEIDYKLA TECHNICA 2012

Disertacija rengta 2008–2012 metais Vilniaus Gedimino technikos universitete.

Mokslinis vadovas

doc. dr. Marijus RADAVIČIUS (Vilniaus Gedimino technikos universitetas, fiziniai mokslai, matematika – 01P).

Disertacija ginama Vilniaus Gedimino technikos universiteto Matematikos mokslo krypties taryboje:

Pirmininkas

prof. habil. dr. Kęstutis KUBILIUS (Vilniaus Gedimino technikos universitetas, fiziniai mokslai, matematika – 01P).

Nariai:

prof. habil. dr. Algimantas Jonas BIKELIS (Vytauto Didžiojo universitetas, fiziniai mokslai, matematika – 01P),

prof. dr. Kęstutis DUČINSKAS (Klaipėdos universitetas, fiziniai mokslai, matematika – 01P),

doc. dr. Mečislavas MEILŪNAS (Vilniaus Gedimino technikos universitetas, fiziniai mokslai, matematika – 01P),

doc. dr. Pranas VAITKUS (Vilniaus universitetas, fiziniai mokslai, matematika – 01P).

Oponentai:

prof. habil. dr. Vilijandas BAGDONAVIČIUS (Vilniaus universitetas, fiziniai mokslai, matematika – 01P),

prof. habil. dr. Jonas Kazys SUNKLODAS (Vilniaus Gedimino technikos universitetas, fiziniai mokslai, matematika – 01P).

Disertacija bus ginama viešame Matematikos mokslo krypties tarybos posėdyje 2013 m. sausio 23 d. 13 val. Vilniaus Gedimino technikos universiteto Senato posėdžių salėje.

Adresas: Saulėtekio al. 11, LT-10223 Vilnius, Lietuva.

Tel.: (8 5) 274 49 52, (8 5) 274 49 56; faksas (8 5) 270 01 12;

el. paštas doktor@vgtu.lt

Disertacijos santrauka išsiuntinėta 2012 m. gruodžio 21 d.

Su disertacija galima susipažinti Vilniaus Gedimino technikos universiteto (Saulėtekio al. 14, LT-10223 Vilnius, Lietuva) ir Vilniaus universiteto Matematikos ir informatikos instituto (Akademijos g. 4, LT-08663 Vilnius, Lietuva) bibliotekose.

VGTU leidyklos „Technika“ 2095-M mokslo literatūros knyga.

VILNIUS GEDIMINAS TECHNICAL UNIVERSITY

Pavel SAMUSENKO

NONPARAMETRIC CRITERIA FOR SPARSE CONTINGENCY TABLES

SUMMARY OF DOCTORAL DISSERTATION

PHYSICAL SCIENCES,
MATHEMATICS (01P)


LEIDYKLA
Vilnius TECHNIKA 2012

Doctoral dissertation was prepared at Vilnius Gediminas Technical University in 2008–2012.

Scientific Supervisor

Assoc Prof Dr Marijus RADAVIČIUS (Vilnius Gediminas Technical 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 (Vilnius Gediminas Technical University, Physical Sciences, Mathematics – 01P).

Members:

Prof Dr Habil Algimantas Jonas BIKELIS (Vytautas Magnus University, Physical Sciences, Mathematics – 01P),

Prof Dr Kęstutis DUČINSKAS (Klaipėda University, Physical Sciences, Mathematics – 01P),

Assoc Prof Dr Mečislavas MEILŪNAS (Vilnius Gediminas Technical University, Physical Sciences, Mathematics – 01P),

Assoc Prof Dr Pranas VAITKUS (Vilnius University, Physical Sciences, Mathematics – 01P).

Opponents:

Prof Dr Habil Vilijandas BAGDONAVIČIUS (Vilnius University, Physical Sciences, Mathematics – 01P),

Prof Dr Habil Jonas Kazys SUNKLODAS (Vilnius Gediminas Technical University, Physical Sciences, Mathematics – 01P).

The dissertation will be defended at the public meeting of the Council of Scientific Field of Mathematics in the Senate Hall of Vilnius Gediminas Technical University at 1 p. m. on 23 January 2013.

Address: Saulėtekio al. 11, LT-10223 Vilnius, Lithuania.

Tel. +370 5 274 49 52, +370 5 274 49 56; fax +370 5 270 01 12;

e-mail: doktor@vgtu.lt

The summary of the doctoral dissertation was distributed on 21 December 2012.

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 Vilnius University Institute of Mathematics and Informatics (Akademijos 4, LT-08663 Vilnius, Lithuania).

Ivadas

Problemos formulavimas

Tradiciškai reikalaujama, kad daugumoje (stebėtų) dažnių lentelės langelių tikėtini (pagal nulinės hipotezės prielaidas) dažniai būtų nemažesni už 5. Jeigu šis reikalavimas nėra įvykdytas, tai suderinamumo statistikų skirstinių aproksimacija χ^2 skirstinių gali būti nepakankamai tiksli, ir tokia dažnių lentelė vadinama išsklaidyta, arba retų įvykių dažnių lentele.

Problemos, susijusios su retų įvykių dažnių lentelėmis, yra plačiai aptartos mokslinėje literatūroje. Yra pasiūlyta daug metodų: alternatyvieji skirstinių aproksimavimo būdai, ranginių kintamųjų dažnių lentelių glodinimas, parametrinė ir neparametrinė saviranka, Bayeso metodai ir kiti. Tačiau jie nepritaikomi arba yra neefektyvūs atliekant neparametrinę išsklaidytų dažnių lentelių analizę.

Disertacijoje nagrinėjamas neparametrinių kriterijų pagrįstumo retų įvykių dažnių lentelėms uždavinys ir tiriama tų kriterijų galios priklausomybė nuo retų įvykių lentelių išsisklaidymo lygio.

Darbo aktualumas

Pastaruoju metu sukauptos informacijos kiekiai yra didžiuliai, todėl su didelės dimensijos duomenų problema susiduriama gana dažnai. Kategoriniams duomenims ši problema ypač aktuali. Ryšiai tarp tolydžių (kiekybinių) kintamųjų paprastai išreiškiami vidurkais ir kovariacinėmis matricomis, todėl jiems nusakyti užtenka eilės $O(n^2)$ skaliarinių parametru. Čia n yra duomenų dimensija. Kategoriniams duomenims skaliarinių parametru skaičius bendruoju atveju didėja eksponentiškai atžvilgiu n . Todėl net gana nedidelės dimensijos kategorinių duomenų dažnių lentelės turi daug tuščių langelių arba langelių su mažu dažniu, taigi yra išsklaidytos. Išsklaidytos dažnių lentelės dažnai aptinkamos genetinių sekų analizėje, apklausose su didesniais klausimynais, tekstų ir natūralios kalbos analizėje bei tekstų apdorojime ir kitur.

Bendruoju atveju išsklaidytoms dažnių lentelėms pagrįstasis tikėtinų dažnių įvertinys neegzistuoja. Ranginių kintamųjų duomenims galima pritaikyti grupavimą arba glodinimą su branduoliu ir tokiu būdu gauti pagrįstuosius tikėtinų dažnių įvertinius su sąlyga, kad jie yra pakankamai glodūs. Nominaliųjų kintamųjų duomenims, kurie ir yra pagrindinis šio darbo tyrimo objektas, nei glodinimas su branduoliu, nei grupavimas tiesiogiai, deja, nepritaikomi.

Tyrimų objektas

Disertacijos tyrimų objektai yra išsklaidytų kategorinių (nominaliųjų) duomenų apibrėžimas, retų įvykių dažnių lentelės. Darbe nagrinėjamos jų profilio (spektrinės) statistikos, laipsninės divergencijos pagrindu apibrėžtos suderinamumo statistikos, neparametrinių kriterijų pagrįstumas ir jų savybės

baigtinių imčių atveju.

Darbo tikslas ir uždaviniai

Šio darbo tikslas yra ištirti klasikinių suderinamumo statistikų ir testų asimptotines bei baigtinės imties savybes ir pasiūlyti klasikinių testų patobulinimus.

Šiame darbe yra sprendžiami tokie uždaviniai:

1. Pasiūlyti klasikinių kriterijų patobulinimus, kurie padidintų jų jautrumą išsklaidytų nominaliųjų duomenų skirstinio nukrypimams nuo nulinės hipotezės.
2. Įrodyti pasiūlytų kriterijų pagrįstumą.
3. Kompiuterinio modeliavimo būdu palyginti klasikinių ir siūlomų testų veikimą baigtinių imčių atveju esant įvairioms alternatyvoms.

Tyrimų metodai

Asimptotinis retų įvykių duomenų modeliavimas remiasi latentinio skirstinio modeliu, didelio skaičiaus retų įvykių (angl. LNRE – large number of rare events) struktūrinio skirstinio modeliu bei empiriniu Bayeso metodu. Suderinamumas yra matuojamas laipsnine divergencija. Rezultatai tikėtinumo santykio kriterijui su dažnių lentelės profilio (spektrinėmis) statistikomis, įrašytomis vietoje stebėtų dažnių, remiasi bendra tikėtinumo teorija ir modeliu pagrįsta klasterine analize. Statistinių kriterijų pagrįstumui (nepagrįstumui) įrodyti naudojamos Chebyshevo tipo nelygybės. Kriterijų elgesys baigtinių imčių atveju yra tiriamas Monte Karlo metodu. Skaičiavimai atlikti naudojant R paketą.

Mokslinis darbo naujumas

Darbe yra parodyta, kad išsklaidytiems kategoriniams duomenims tikėtinumo santykio ir Pearsono χ^2 statistikos gali būti neinformatyvios: jos jau neatspindi nulinės hipotezės ir duomenų suderinamumo, t. y. jų neatitikimo masto. Taigi, jos gali būti nepagrįstos net tada, kai egzistuoja paprastas pagrįstasis testas. Įvestas naujas retų įvykių asimptotinis modelis, kuris remiasi (išplėstiniu) empiriniu Bayeso metodu. Pasiūlyti testai, kuriuose naudojamas grupavimas, Markovo grandinių Monte Karlo metodas bei Puasono skirstinių (baigtinių) mišinių modelis. Įrodytas testų, kuriuose glodinimui atliekamas grupavimas, pagrįstumas.

Praktinė darbo rezultatų reikšmė

Darbe pasiūlyti neparametriniai kriterijai gali būti taikomi didelio matavimo kategorinių (nominaliųjų) duomenų statistinei analizei. Tokie duomenys dažnai pasitaiko apklausose su dideliais klausimynais, natūraliosios kalbos bei tekstų tyrimuose, genetinių sekų analizėje ir pan. Siūlomi kriterijai yra gana paprastai realizuojami, greitai apskaičiuojami ir jiems nereikia jokių specifinių prielaidų

apie tiriamus kategorinius (nominaliuosius) duomenis.

Ginamieji teiginiai

1. Labai išsklaidytų kategorinių duomenų atveju tiek tikėtumo santykio statistika, tiek ir Pearsono χ^2 statistika gali būti neinformatyvios: jos jau neatspindi nulinės hipotezės ir duomenų neatitikimo masto.
2. Retų įvykių asimptotinis modelis, sudarytas remiantis empiriniu Bayeso metodu, leidžia ir nominaliesiems duomenims pritaikyti latentinio skirstinio modelį.
3. Remiantis empirine Bayeso metodologija, glodinimas naudojant grupavimą, Markovo grandinių Monte Karlo glodinimas bei Puasono skirstinių (baigtinių) mišinių modelio taikymas pagerina klasikinių kriterijų galią ypač reguliariųjų alternatyvų atveju.
4. Esant gana bendroms sąlygoms kriterijai, kuriuose glodinimui atliekamas grupavimas, yra pagrįsti.

Darbo rezultatų apibavimas

Disertacijos rezultatai paskelbti 4 straipsniuose recenzuojamuose periodiniuose mokslo leidiniuose ir pristatyti 5 mokslinėse konferencijose, iš kurių 3 – tarptautinės.

Disertacijos struktūra

Disertaciją sudaro įvadas, keturi pagrindiniai skyriai, bendrosios išvados, literatūros sąrašas, autoriaus publikacijų disertacijos tema sąrašas ir 1 priedas. Disertacijos apimtis yra 106 puslapiai, 29 lentelės, 31 paveikslas ir 62 literatūros šaltiniai.

1. Išretintų dažnių lentelių apžvalga ir apibrėžtys

Per pastaruosius dešimtmečius susidomėjimas daugiamačių kokybinių duomenų analize labai padidėjo, buvo išplėtoti efektyvūs didelio matavimo stebėtų dažnių lentelių statistinės analizės metodai. Jie išsamiai išdėstyti vadovėliuose, monografijose, pvz., Habermano, Bishopo, Fienbergo, Reado ir Cressie, Agresti, Congdono, jiems skirtos Agresti, Fienbergo, Agresti ir Hitchcocko apžvalgos.

Tačiau didelio matavimo kategorinių (ypač nominaliųjų) duomenų problema lieka neišspręsta ir darosi vis aktualesnė.

Tiriamų kategorinių duomenų ir siūlomo (suformuluoto nulinėje hipotezėje) modelio suderinamumas paprastai matuojamas divergencija (angl. divergence), jos pagrindu yra sudaromos kriterijaus statistikos. Divergencijos visada yra neneigiamos ir yra lygios 0 tik sutapimo atveju. Atskirais atvejais divergencijos

tenkina atstumo aksiomas. Cressie ir Reedas plačiai išnagrinėjo kriterijų, kurie remiasi laipsninės divergencijos statistikomis, savybes. Gerai žinomos statistikos, Pearsono χ^2 statistika X^2 ir (dviguba neigiama logaritmuota) tikėtinumo santykio statistika G^2 , taip pat priklauso laipsninės divergencijos statistikų šeimai.

Formuluojant suderinamumo kriterijus naudojamas kriterijaus statistikos skirstinys. Apibendrinant Pearsono rezultatus yra parodyta, kad esant teisingai nulinei hipotezei ir fiksuotam dažnių lentelės dydžiui, laipsnines divergencijos statistikos turi asimptotinę χ^2 skirstinį, kai imties dydis neapbrėžtai auga. Tačiau, praktikoje imties dydis yra baigtinis. Todėl natūraliai iškyla klausimas apie aproksimacijos χ^2 skirstinių tikslumą.

Praktinę nykščio taisyklę, taikomą norint įvertinti, kada χ^2 aproksimacija yra pakankamai tiksli, pirmasis pasiūlė Cochranas. Yra reikalaujama, kad dažnių lentelėje mažiausias tikėtinas dažnis turėtų būti ne mažesnis už 5. Įvairios kitos praktinės taisyklės yra išvardintos Hu disertacijoje.

Kategorinių duomenų, kurie neatitinka šių taisyklių, praktikoje pasitaiko gana dažnai: apklausose su didesniais klausimynais, genetinių sekų analizėje, natūralios kalbos tyrimuose, tekstų apdorojime ir kitur. Jie vadinami išretintais (angl. sparse), arba retų dažnių, duomenimis. Konkrečių išsklaidytų lentelių pavyzdžiai pateikti knygoje Agresti, XactStat dokumentacijoje.

Savo ruožtu yra siūloma įvairių praktinių būdų šiai problemai spręsti dažnių lentelės langelių sujungimas, teigiamos konstantos pridėjimas, apsiribojimas tik langelių su didesniais stebėtais dažniais nagrinėjimu ir kt. Kaip teisingai pabrėžia Baglivo, šie metodai gali iškraipyti statistinės analizės rezultatus, ir jų taikymas reikalauja (teorinio) pagrindimo.

Koehlerio ir Larntzo Monte Karlo metodu atliktas tyrimas liudija, kad paprastai taikomos dažnių lentelėms jų išretinimo (išsisklaidymo) nustatymo taisyklės yra per daug konservatyvios, tačiau mažai tikėtina, kad pavyktų surasti paprastą universalią taisyklę. Iš tikro χ^2 aproksimacijos tikslumą nulemia ne vien tik tikėtinų dažnių didumas ar mažumas, bet greičiau jų nesubalansuotumas.

Teoriniai išretintų dažnių lentelių tyrimai paremti retų dažnių arba išretinimo asimptotika (angl. sparse asymptotics), kai dažnių lentelės langelių skaičius n auga kartu su imties dydžiu N . Įvairių statistikų, susijusių su dažnių lentelėmis, tarp jų ir suderinamumo statistikų, asimptotinis elgesys, kai $n \rightarrow \infty$, buvo nagrinėjamas ir anksčiau. Morisas bei Tumanyan savo darbuose parodė, kad χ^2 aproksimacija Pearson statistikai X^2 ir tikėtinumo santykio statistikai G^2 galioja ir retų įvykių asimptotikos atveju, jeigu tik dažnių lentelės langelių mažiausias tikėtinas dažnis artėja į ∞ pakankamai greitai. Györfi ir Vajda apibendrina šiuos rezultatus suderinamumo statistikoms, apibrėžtomis naudojant Csiszár's ϕ -divergencijas, atskiru atveju – laipsnines divergencijas.

Jeigu tikėtini dažniai elgiasi gana reguliariai (yra panašios eilės – subalansuoti), tai asimptotinis normalumas išlieka ir tuomet, kai n yra tos pačios eilės kaip N . Tačiau šiuo atveju asimptotinio normalumo centruojančios ir normuojančios sekos priklauso nuo nežinomų tikėtinų lentelės dažnių. Tokiu atveju, pabrėžia

Khmaladzė, klasikiniai kriterijai jau nėra laisvi nuo skirstinio (angl. distribution free).

Kolchinas, Sevastjanovas ir Chistyakovas, Ivchenko ir Medvedevas aptaria kriterijus, kurie paremti tuščių dėžučių skaičiaus, arba paskirstymo (angl. allocation, occupation), statistikomis, kartais dar vadinamomis dažnių lentelių spektrinėmis arba profilio statistikomis.

Ši trumpa apžvalga rodo, kad išretinimo dažnių lentelėse efektas iš tikrųjų pasireiškia (tik) tuomet, kai $N = O(n)$, $n \rightarrow \infty$ ir/arba tikėtini dažniai ląstelėse yra tam tikra prasme nereguliarūs, nesubalansuoti (labai išsisklaidę). Norint retų įvykių (išretinimo) asimptotikoje su $N = O(n)$ gauti priimtinius (bendrus ir suvokiamus) rezultatus kaip tik reikia prielaidų apie tikėtinų dažnių asimptotinio elgesio koki nors reguliarumą. Disertacijoje aptariami 3 išretinimo asimptotikos modeliai: latentinio skirstinio, struktūrinio skirstinio ir naujai siūlomas išplėstinis empirinis Bajeso modelis.

Dažnių lenteles, kai $N = o(n)$, $n \rightarrow \infty$ vadinsime labai išretintomis.

Puasono ir polinominės imtys

Tarkime, kad $\mathbf{y} = \mathbf{y}(n) := (y_1, \dots, y_n) \sim \text{Poisson}_n(\mu)$ ir $\mathbf{y}$ komponentės yra nepriklausomos (Puasono imties schema) arba

$$\mathbf{y} \sim \text{Multinomial}_n(N, \mathbf{p}), \quad (1)$$

(polinominė imties schema).

Čia y yra stebėti dažniai, n yra ląstelių skaičius dažnių lentelėje, polinominėje schemeje $N = y_+ := \sum_{i=1}^n y_i$ yra imties dydis, $\mu := (\mu_1, \dots, \mu_n)$ yra tikėtinų y reikšmės, $\mathbf{p} := (p_1, \dots, p_n)$ yra atitinkamų tikimybių vektorius, $p_i := \mu_i / \mu_+$, $i = 1, \dots, n$, $\mu_+ := \sum_{j=1}^n \mu_j$.

Laipsninės divergencijos statistikos

Tarkime, kad $\mu^\circ := (\mu_1^\circ, \dots, \mu_n^\circ)$ yra duotas teigiamas vektorius, $\mathbf{p}^\circ := \mu^\circ / \mu_+^\circ$. Nagrinėjama paprastoji hipotezė su sudėtingąja alternatyva:

$$H_0 : \mu = \mu^\circ \text{ versus } H_1 : \mu \neq \mu^\circ. \quad (2)$$

Nulinės hipotezės H_0 ir stebėtų dažnių lentelės $\mathbf{y}$ suderinamumą galima matuoti divergencijos (nukrypimo) statistika. Laipsninės divergencijos statistikų (angl. power divergence statistics) savybės yra tyrinėtos Cressie ir Reado

darbuose. Laipsninės divergencijos statistika:

$$D(\lambda) = \frac{2}{\lambda(\lambda+1)} \sum_{i=1}^k \left\{ y_i \left[\left(\frac{y_i}{\mu_i^\circ} \right)^\lambda - 1 \right] - \lambda [y_i - \mu_i^\circ] \right\},$$

čia $\lambda \in (-\infty, \infty)$ (atvejai $\lambda = 0$ ir $\lambda = -1$ apibrėžiami pagal tolydumą kaip atitinkamos ribos). Kai $\lambda = 1$ ir $\lambda = 0$, gaunamos atitinkamai Pearsono χ^2 statistikos ir (neigiama dviguba logritmuota) tikėtinumo santykio statistika. Disertacijoje dar nagrinėjamos modifikuota Pearsono χ^2 statistika $X_m^2 = D(-2)$, modifikuota tikėtinumo santykio statistika $G_m^2 = D(-1)$, Freemano-Tukey'aus statistika $F^2 = D(-1/2)$ ir Cressie ir Reado rekomenduojama statistika $D(3/2)$, kuri darbe dar vadinama mažų imčių (angl. small sample) statistika. Cressie ir Reado rezultatai rodo, kad ši statistika geriausiai tinka, kai imties dydis N yra mažas ir yra mažai arba nieko nežinoma apie galimus alternatyvius modelius.

Latentinio skirstinio modelis

Tarkime, kad ranginio kintamojo reikšmių tikėtini dažniai $\mu = (\mu_1, \dots, \mu_n)$ yra pasiskirstę pagal latentinę pasiskirstymo funkciją: $F : [0, 1] \rightarrow [0, 1]$:

$$\mu_i = \mu_i(n) := \mu_+ \cdot (F(i/n) - F((i-1)/n)), \quad i = 1, \dots, n. \quad (3)$$

Šiuo atveju išretinimo lygis (mastas) yra išreiškiamas tikėtinų dažnių vidurkiu $\rho = \rho(n) := \mu_+/n$. Tipinė retų įvykių asimptotikos prielaida yra $\rho = O(1)$.

Latentinio skirstinio (3) modelis su tolygiais pagal n apribojimais latentinio pasiskirstymo tankio $f = F'$ glodumui yra netinkamas nominaliesiems duomenims, nes jiems statistinės išvados neturi priklausyti nuo langelių (kategorijų) numeravimo tvarkos.

Nominaliesiems duomenims retų įvykių asimptotika gali būti nusakyta remiantis struktūriniu skirstiniu.

Struktūrinio skirstinio modelis

Kartais natūralu daryti prielaidą, kad statistinėms išvadoms langelių numeravimo tvarka yra nesvarbi. Tada pradinis hipotezių tikrinimo (2) uždavinys gali būti formuluojamas struktūrinio skirstinio terminais.

Tarkime, $\hat{F} = \hat{F}^{(n)}$ yra stebinių $\{\mu_1, \dots, \mu_n\}$ empirinė pasiskirstymo funkcija.

Apibrėžimas. Tegul $F_\rho(t) := \hat{F}^{(n)}(\rho t)$, $\rho = \rho(n) > 0$, ir tarkime, kad F_ρ silpnai konverguoja į pasiskirstymo funkciją F , kai $n \rightarrow \infty$. Tada F vadinamas lentelės tikėtinų dažnių μ (arba tiesiog lentelės dažnių $\mathbf{y}$) struktūriniu pasiskirstymo funkcija su normuojančiu daugikliu $\rho = \rho(n)$.

Struktūrinio skirstinio hipotezė: $H_0 : F = F^\circ$ versus $H_1 : F \neq F^\circ$. Čia F° yra duota pasiskirstymo funkcija.

2. Klasikinių kriterijų nepagrįstumas labai išretintiems duomenims

Nagrinėkime polinominę (1) schemą su tikimybių vektoriumi $\mathbf{p} = \mathbf{p}(n) := (p_1, \dots, p_n) \in \mathcal{P}^{(n)}$, čia $\mathcal{P}^{(n)}$ yra duota tikimybinių skirstinių aibėje

$J_n := \{1, \dots, n\}$ klasė. Tegul $\mathbf{p}^\circ = \mathbf{p}^\circ(n)$ yra duotas tikimybinių skirstinys aibėje J_n , kuris nepriklauso klasei $\mathcal{P}^{(n)}$. Sprendžiamas hipotezių tikrinimo uždavinys:

$$H_0 : \mathbf{p} = \mathbf{p}^\circ \text{ vs } H_1 : \mathbf{p} \in \mathcal{P}^{(n)}. \quad (4)$$

Šio poskyrio tikslas yra parodyti, kad klasikiniai kriterijai labai išretintoms dažnių lentelėms gali būti nepagrįsti. Bendruoju atveju to tikimasi, nes modelio parametrų skaičius, kai $N = O(n)$, yra proporcingas imties dydžiui. Tačiau klasikinių kriterijų nepagrįstumas išlieka ir tuomet, kai klasės $\mathcal{P}^{(n)}$ skirstiniai akivaizdžiai skiriasi nuo $\mathbf{p}^\circ$.

Apibrėžimas. Suderinamumo kriterijus, kuris paremtas statistika

$T_n := T(\mathbf{y}(n))$, vadinamas pagrįstuoju hipotezių tikrinimo (4) uždavinyje, jeigu egzistuoja seka c_n tokia, kad $\mathbb{P}_0 \{T_n > c_n\} + \mathbb{P}_{\mathbf{p}} \{T_n < c_n\} \rightarrow 0$, $\forall \mathbf{p} \in \mathcal{P}^{(n)}$, $n \rightarrow \infty$. Priešingu atveju kriterijus vadinamas nepagrįstu.

Čia $\mathbb{P}_0$ (atitinkamai $\mathbb{P}_{\mathbf{p}}$) žymi tikimybinių skirstinių, nusakytą nuline hipoteze H_0 (alternatyva H_1 su duotu $\mathbf{p} \in \mathcal{P}^{(n)}$). Analogiškai apibrėžiami ir vidurkiai $\mathbb{E}_0$, $\mathbb{E}_{\mathbf{p}}$ bei dispersijos $\mathbb{D}_0$, $\mathbb{D}_{\mathbf{p}}$.

Teiginys. Tarkime, kad Pearsono χ^2 statistikai X^2 ir kuriam nors $\mathbf{p} \in \mathcal{P}_n$ galioja

$$\Delta_n := \mathbb{E}_{\mathbf{p}} X^2 - \mathbb{E}_0 X^2 < 0 \quad (5)$$

ir

$$\rho_n^2 := \Delta_n^2 / D_n^2 \rightarrow \infty \quad (n \rightarrow \infty) \quad (6)$$

su $D_n := \sqrt{\mathbb{D}_0 X^2} + \sqrt{\mathbb{D}_{\mathbf{p}} X^2}$. Tada Pearsono χ^2 kriterijus yra nepagrįstas.

Kita vertus, kriterijus, kuris remiasi statistika $T_n := |\chi^2 - (n - 1)|$, yra pagrįstas, jeigu (6) formulė galioja visiems $\mathbf{p} \in \mathcal{P}_n$.

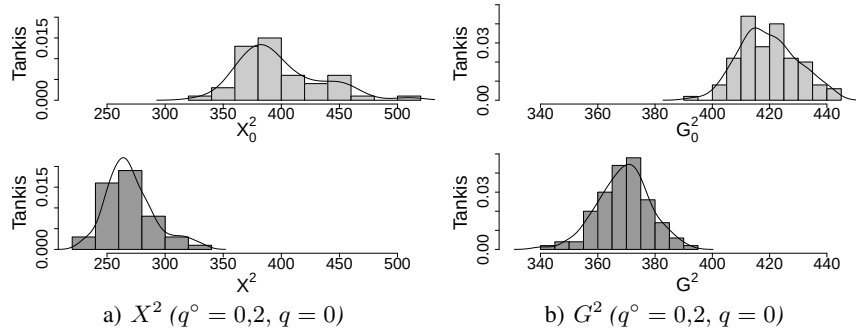
Pavyzdys: $\beta > 1$, $q_0 \in (0, 1/2)$ yra konstanta, $m = \lceil N^\beta \rceil$, $n = 2m$,

$$p_j^\circ = q_0/m, \forall j \leq m, \quad p_j^\circ = (1 - q_0)/m, \forall j > m,$$

$$p_j = 0, \forall j \leq m, \quad p_j = 1/m, \forall j > m.$$

Tuomet $N/n = o(1)$ ir $\rho_n = -(Nq_0(1 - q_0))^{1/2} [1 + O(N/n)]$.

Monte Karlo imitacinio modeliavimo rezultatai su $n = 600$ ir $N = 200$, pateikti 1 paveikslėlyje, rodo, kad aprašytas reiškinys stebimas ir esant baigtiniams



1 pav. Statistikų histogramos prie hipotezės ir alternatyvos:

a) X^2 statistika; b) G^2 statistika

n ir N . Be to, jis, matyt, galioja ir tikėtinumo santykio kriterijui G^2 . Kriterijui G^2 polinominėje schemoje jis įrodytas (esant gana griežtoms sąlygoms) Radavičiaus ir Samusenko darbe, o Puasono schemoje – kitame šių autorių leidinyje.

3. Suderinamumo kriterijų išretintoms lentelėms pagrįstumas

Khmaladze pastebėjo, kad struktūrinis skirstinys gali būti traktuojamas kaip latentinis sumaišantysis skirstinys empiriniame Bayeso metode. Žemiau yra pateiktas šio metodo išplėtimas įtraukiant į Bayeso modelį ir nulinės hipotezės parametrus. Tariama, kad $\{(\mu_i^\circ, \mu_i), i = 1, \dots, n\}$ yra atsitiktinių dydžių poros (γ°, γ) su reikšmėmis $\mathbb{R}_+^2$ ir skirstiniu $Q = Q^{(n)}$, $n \rightarrow \infty$, nepriklausomos kopijos.

Taigi, stebimi duomenys $\mathcal{D} = \mathcal{D}^{(n)} := \{(\mu_i^\circ, y_i), i = 1, \dots, n\}$ yra aprašomi tokiu tikimybinio modeliu:

- y_i sąlyginis skirstinys, kai žinoma (μ_i°, μ_i) , yra Puasono su parametru μ_i ,
- $\{(\mu_i^\circ, \mu_i), i = 1, \dots, n\}$ yra nepriklausomos ir vienodai pasiskirsčiusios (iid) poros su bendru skirstiniu $Q = Q^{(n)}$ ($n \rightarrow \infty$).

Tada, kai skirstinys Q nepriklauso nuo n , atsitiktinio dydžio γ° (marginalinis) skirstinys Q° esant teisingai nulinei hipotezei H_0 sutampa su μ° struktūriniu skirstiniu, o atsitiktinio dydžio γ (marginalinis) skirstinys Q^γ esant teisingai alternatyvai H_1 sutampa su μ struktūriniu skirstiniu.

Pasiūlyta tikimybinių schema yra labai lanksti. Ji leidžia nagrinėti įvairias, taip pat ir neregulias, retų įvykių asimptotikas ir bendresnius hipotezių tikrinimo uždavinius. Pavyzdžiui, struktūrinio skirstinio nulinė hipotezė reiškia, kad $Q^\circ = Q^\gamma$, o nulinės hipotezės ir alternatyvos susiejimas atsitiktinės poros (γ°, γ) skirstinyje leidžia neprarandant informacijos išdėstyti stebinius natūralia

μ° reikšmių didėjimo tvarka ir tuo remiantis pritaikyti latentinio skirstinio modelį ir nominaliesiems duomenims.

Tarkime, kad $\mathcal{P} = \mathcal{P}^{(n)}$ žymi tam tikrą galimų skirstinių $Q = Q^{(n)}$ klasę, o $\mathbb{P}_P = \mathbb{P}_P^{(n)}$ žymi duomenų $\mathcal{D}^{(n)}$ skirstinį, kai atsitiktinės poros (γ°, γ) generavimui naudojamas skirstinys $Q = P$. Tegul $d(Q; P)$ yra duotas skirstinių Q ir P , $Q, P \in \mathcal{P}$, neatitikimo (divergencijos) matas. Jam galioja $d(Q; P) \geq 0$, $d(P; P) = 0$.

Su žinomu $P^{(n)} \in \mathcal{P}^{(n)}$ ir pasirinktu $\delta = \delta(n) > 0$ nagrinėkime tokią hipotezių tikrinimo uždavinį:

$$H_0 : d(Q^{(n)}, P^{(n)}) = 0 \text{ versus } H_1 : d(Q^{(n)}, P^{(n)}) \geq \delta(n), \forall n. \quad (7)$$

Prielaida C. Sakykime, kad Q yra žinomas skirstinys, o δ yra seka iš (7) formulį. Tarkime, kad remiantis duomenimis $\mathcal{D}^{(n)}$ pavyko sukonstruoti tokią neatitikimo mato $d(Q; P)$ įvertinį $\hat{d}(Q; P)$, kad su tam tikru $\tau = \tau(n) \in (0, 1)$ ir su visais $P^{(n)} \in \mathcal{P}^{(n)}$

$$\mathbb{P}_Q^{(n)} \{ \hat{d}(Q^{(n)}; P^{(n)}) > (1 - \tau(n))\delta(n) \} \rightarrow 0,$$

$$\mathbb{P}_P^{(n)} \{ \hat{d}(Q^{(n)}; P^{(n)}) < d(Q^{(n)}; P^{(n)}) - \tau(n)\delta(n) \} \rightarrow 0, \quad n \rightarrow \infty.$$

Lema. Tarkime, kad galioja C prielaida. Tada kriterijus

$$\mathcal{K} := \{ \hat{d}(Q^{(n)}; P^{(n)}) > (1 - \tau(n))\delta(n) \}$$

yra pagrįstas, kai $n \rightarrow \infty$, tikrinant (7) uždavinį.

Įprasta neatitikimo matą d parinkti iš laipsninės divergencijos šeimos. Standartinis universalus būdas įvertinti $d(Q; P)$ yra nežinomą skirstinį Q pakeisti jo įvertiniu $\hat{Q}$. Tačiau, kaip parodyta ankstesniame skyriuje, tradiciniai suderinamumo kriterijai ne visada yra tinkami išretintoms lentelėms. Todėl yra siūloma duomenis glodinti arba grupuoti remiantis išplėstiniu Bayeso modeliu.

Glodinimas. Radavičiaus ir Samusenko darbe glodinimas realizuojamas aprašant atsitiktinės poros (γ°, γ) skirstinį Q baigtiniu skirstinių mišiniu. Dvimačiai mišinio komponentų skirstiniai yra sudaryti iš nepriklausomų marginalinių skirstinių, marginalinis γ° skirstinys yra normalusis, o marginalinio γ skirstinio parametrinė forma nėra nusakyta, jis modeliuojamas neparametriškai. Skirstiniui Q įvertinti naudojamas didžiausio tikėtumo įvertinys $\hat{Q}_{ML}$. Monte Karlo metodu atlikti pradiniai tyrimai rodo, kad šis metodas gali būti daug galingesnis už klasikinius kriterijus išretintų dažnių lentelių atveju. Tačiau minėto metodo teorinis tyrimas yra sudėtingas.

Sekančiame skyriuje taikoma glodinimo procedūra, pavadinta Markovo

grandinių Monte Karlo glodinimu, kuri remiasi semiparametrinio Radavičiaus ir Židanavičiūtės glodinimo idėja.

Grupavimas. Grupavimo metodas paremtas pastebėjimu, kad išplėstinio Bayeso modelio atveju stebėtų porų $\{(\mu_i^\circ, y_i), i = 1, \dots, n\}$ numeravimo tvarka nekeičia informacijos, ir todėl nemažinant bendrumo galima laikyti, kad seka $(\mu_i^\circ, i = 1, \dots, n)$ yra nemažėjanti.

Tarkime, kad $\Delta = \Delta^{(n)} := \{\Delta_k, k = 1, \dots, K\}$ yra intervalo $(0, \mu_+^\circ]$ suskaidymas į nesikertančius intervalus $\Delta_k = (t_{k-1}, t_k]$, kurių ilgis yra $|\Delta_k| := t_k - t_{k-1}$, $0 = t_0 < t_1 < \dots < t_K := \mu_+^\circ$. Apibrėžkime $\mu_{+j}^\circ := \sum_{i=1}^j \mu_i^\circ$,

$$\mu_{[k]}^\circ := \sum_{i=1}^n \mu_i^\circ \mathbf{1}\{\mu_{+i}^\circ \in \Delta_k\}, \quad (8)$$

$$\mu_{[k]} := \sum_{i=1}^n \mu_i \mathbf{1}\{\mu_{+i}^\circ \in \Delta_k\}, \quad y_{[k]} := \sum_{i=1}^n y_i \mathbf{1}\{\mu_{+i}^\circ \in \Delta_k\}.$$

Tegul $Q^{(n)}$ ir $P^{(n)}$ yra atitinkamai duomenų

$$\{(\mu_i^\circ, \mu_i^\circ), i = 1, \dots, n\} \quad \text{ir} \quad \{(\mu_i^\circ, \mu_i), i = 1, \dots, n\} \quad (9)$$

empiriniai skirstiniai. Skirstinių $Q^{(n)}$ ir $P^{(n)}$ neatitikimas matuojamas laipsnine divergencija, apskaičiuota suskaidymui Δ :

$$d(Q^{(n)}; P^{(n)}) = \sum_{k=1}^K \mu_{[k]}^\circ \phi_\lambda(\mu_{[k]} / \mu_{[k]}^\circ). \quad (10)$$

Čia pažymėta, kad

$\phi_\lambda(t) := (t^{\lambda+1} - (\lambda+1)(t-1) - 1) / (\lambda(\lambda+1))$ (atvejai $\lambda(\lambda+1) = 0$ apibrėžiami pagal tolydumą). Iš (10) išraiškos tiesiogiai gaunamas $d(Q^{(n)}; P^{(n)})$ įvertinys:

$$\widehat{d}(Q; P) := \sum_{k=1}^K \mu_{[k]}^\circ \phi_\lambda(y_{[k]} / \mu_{[k]}^\circ). \quad (11)$$

Pasirinkime, kad $\eta_u \sim \text{Poisson}(u)$ ir

$$\begin{aligned} A(n) &:= \sum_{k=1}^K a(\mu_{k+}^\circ), \quad a(v) := v \mathbb{E} \phi_\lambda(\eta_v/v), \\ V^2(n) &:= \sum_{k=1}^K \sigma^2(\mu_{k+}^\circ, \mu_{k+}), \quad V_0^2(n) := \sum_{k=1}^K \sigma^2(\mu_{k+}^\circ, \mu_{k+}^\circ), \\ \sigma^2(v; u) &:= v^2 \mathbb{E}(\phi_\lambda(\eta_u/v) - \phi_\lambda(u/v))^2. \end{aligned}$$

Teorema. Tarkime, $Q^{(n)}$ ($P^{(n)}$) yra duomenų empiriniai skirstiniai (9) formulėje. Neatitikimo matas tarp $Q^{(n)}$ ir $P^{(n)}$ yra apibrėžtas (10) formulėje ir įvertintas (11) formule. Jeigu $V_0(n) + V(n) = o(\delta(n) - A(n))$, $n \rightarrow \infty$, tai kriterijus

$$\mathcal{K} := \left\{ \widehat{d(Q; P)} > A(n) + \varkappa_1(\delta(n) - A(n)) \right\}$$

yra pagrįstas (kai $n \rightarrow \infty$, konstanta $\varkappa_1 \in (0, 1)$) (7) uždaviniui tikrinti.

4. Modeliavimo rezultatai

Monte Carlo metodas plačiai taikomas statistinių procedūrų savybėms tirti baigtinių imčių atveju bei tų procedūrų efektyvumui palyginti. Klasikinių suderinamumo kriterijų galios retų įvykių dažnių lentelėms tyrimai Monte Carlo metodu aprašyti Koehlerio ir Larntzo, Cressie ir Reedo, Habermano, Hu, Finklerio, ir kitų autorių.

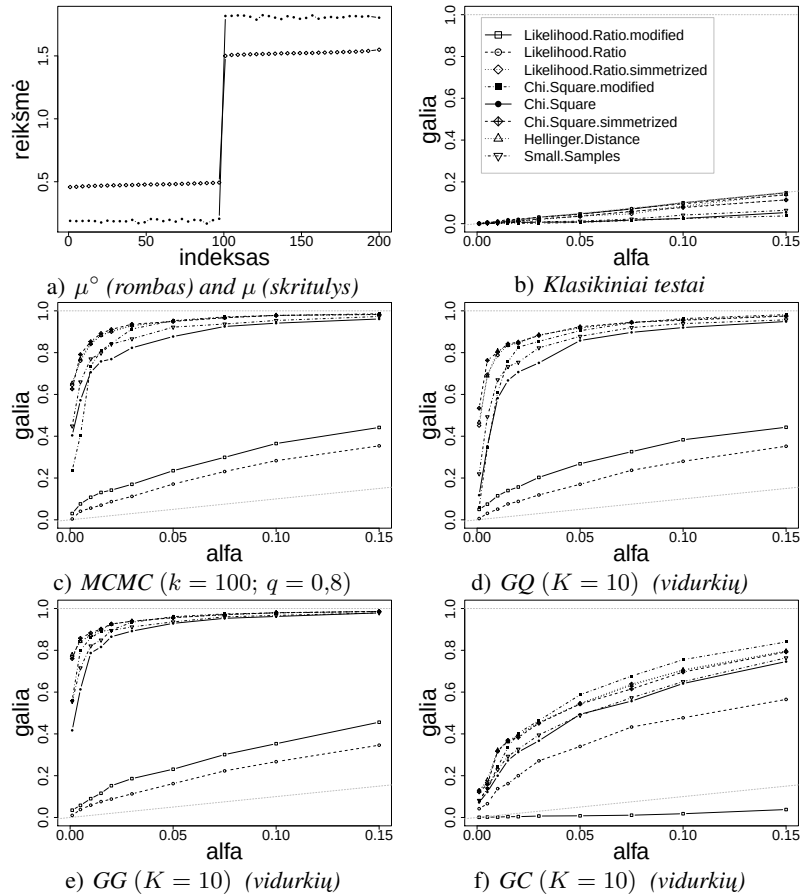
Monte Carlo metodas taikomas ir kaip alternatyvus tradiciniams asimptotiniams skleidiniams būdas, greta tikslųjų testų, kriterijų p-reikšmėms apskaičiuoti (aproksimuoti).

Disertacijoje Monte Carlo metodas naudojamas nagrinėjamų testo statistikų skirstinių aproksimacijai. Tų aproksimacijų pagrindu yra įvertinama kriterijų galia.

Lyginami suderinamumo kriterijai

Kompiuterinio modeliavimo būdu tiriami suderinamumo kriterijai, kurie remiasi laipsninės divergencijos statistikomis, apibrėžtomis (11) formule su $\lambda \in \{2, 5/3, 1/2, 1, 0, -1\}$ sugrupuotiems arba suglodintiems, taikant Markovo grandinių Monte Carlo metodą (žr. žemiau), kategoriniams duomenims. Grupavimas atliekamas stebėtų dažnių lentelės langeliams, sunumeruotiems μ° , tikėtinų dažnių esant nulinei hipotezei H_0 , didėjimo tvarka sujungiant gretimus langelius. Tyrime taikomi tokie grupavimo metodai:

1. Grupės sudarytos taip, kad turėtų (apytiksliai) vienodus tikėtinus dažnius (GC metodas).

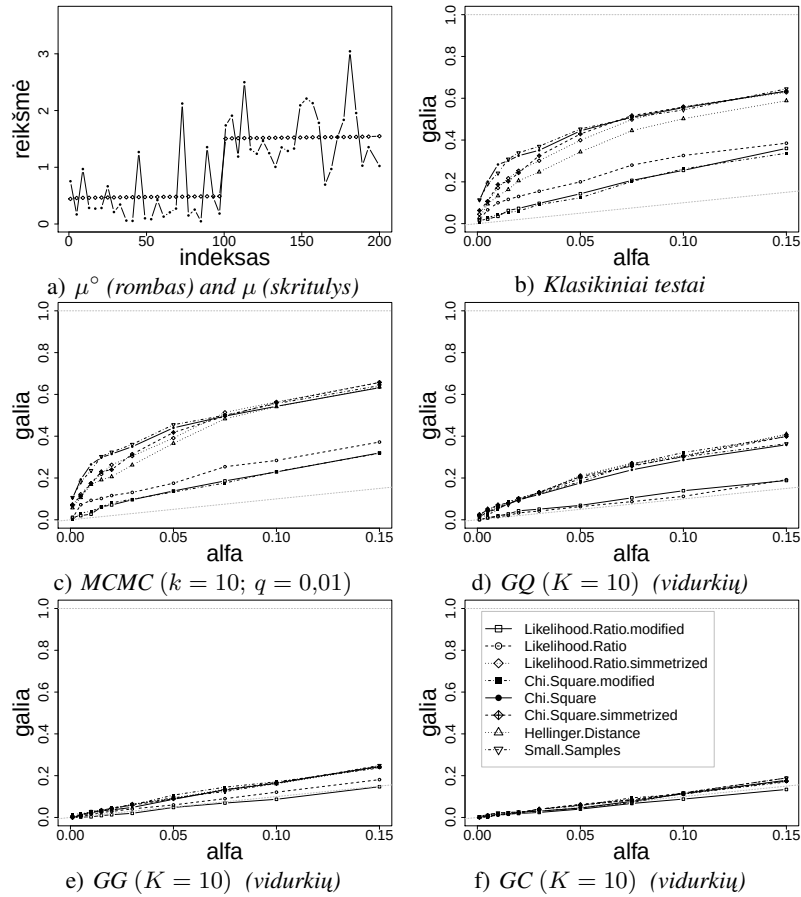


2 pav. Suderinamumo testas (vidurkių). MCMC, GQ, GG ir GC kriterijų, taikomų 2S03 modeliui, galių palyginimas

2. Grupės sudarytos iš (apytiksliai) vienodo langelių skaičiaus (GQ metodas).
3. Grupės sudarytos iš (apytiksliai) vienodo langelių skaičiaus, bet kiekvienas langelis turi svorį, nusakytą gama skirstiniu (GG metodas: langeliams, kurių stebėtas dažnis ženkliai skiriasi nuo grupės vidurkio $\mu_{[k]}^{\circ}$ (8), priskiriamas mažesnis svoris).

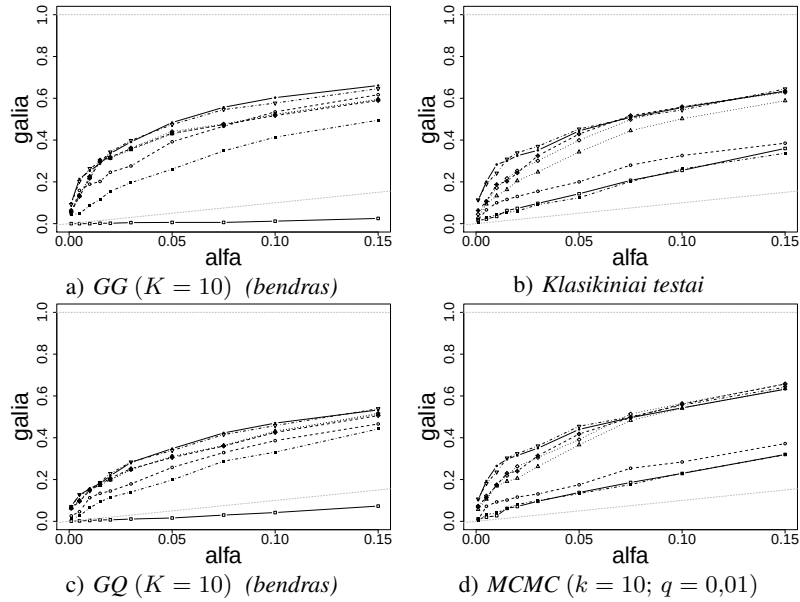
Paprastai pagal statistikas matuojamas vidutinių reikšmių langeliuose arba grupėse, kai teisinga nulinė hipotezė ir kai teisinga alternatyva, neatitikimas. Kriterijai, naudojantys tokias statistikas, turi mažą galią, kai alternatyvos

nukrypimai nuo nulinės hipotezės yra nereguliarūs (žr. dviejų laiptelių su dispersija modelį).



3 pav. Suderinamumo testas (vidurkių). MCMC, GQ, GG ir GC kriterijų, taikomų 2SV025 modeliui, galių palyginimas

Markovo grandinių Monte Carlo (MGMC) glodinimo algoritmas yra specifinė neatsitiktinė (suvidurkinta) semiparametrinių glodinimo procedūrų, aprašytų Faddy ir Joneso bei Radavičiaus ir Židanavičiūtės, versija. Jis priklauso nuo 2 parametrų – r ir k : Parametras $r \in (0, 1)$ parodo, kokia lentelės dažnių dalis yra perskaičiuojama. parametras k parodo, kiek kartų ciklas turi būti kartojamas. Bendras glodinimo efektas priklauso nuo parametrų sandaugos: $r \cdot k$.



4 pav. Suderinamumo testas (bendras). MCMC, GQ, GG ir GC kriterijų, taikomų 2SV025 modeliui, galių palyginimas

Išretintų lentelių generavimo modeliai

Kompiuteriniame eksperimente išretintų lentelių tikėtini dažniai μ° ir μ yra generuoti kaip nepriklausomi gama atsitiktiniai dydžiai:

$$\mu_i \sim \text{gama}(a(i), v), \quad \mu_i^\circ \sim \text{gama}(a^\circ(i), v^\circ), \quad i = 1, \dots, n. \quad (12)$$

Čia $\text{gama}(a, v)$ yra gama skirstinys, nusakytas vidurkiu a ir dispersija v , $a^\circ(\cdot)$ ir $a(\cdot)$ yra paprastos (laiptuotos) funkcijos.

Disertacijoje yra pateikti siūlomų naujų kriterijų, skirtų išretintoms lentelėms, bei klasikinių testų galios palyginimo rezultatai trijų tipų generuotiems modeliams:

- Dviejų laiptelių (angl. Two step) modelis yra vienas iš paprasčiausių, jis jau buvo naudojamas 2 skyriuje klasikinių kriterijų nepagristumo retų įvykių dažnių lentelėse iliustracijai.
- Suskaidyti dviejų laiptelių modeliai, kai alternatyva nuo hipotezės skiriasi tik viename iš laiptelių.
- Dviejų laiptelių su dispersija (angl. Two step with dispersion) modelis, kurio $a^\circ(\cdot)$ ir $a(\cdot)$ nesiskiria, alternatyva nuo nulinės hipotezės skiriasi tik

atsitiktiniais nukrypimais, kurių vidutinį dydį reguliuoja dispersija v .

Čia pateikiame tik dviejų laiptelių modelio 2S03 (žr. 2 pav. a)) ir dviejų laiptelių su dispersija modelio 2SV025 su dispersija $v = 0,25$ (žr. 3 pav. a)) bei jiems gautų testų palyginimo rezultatų grafines iliustracijas.

Rezultatai

Kriterijų galios palyginimo rezultatai dviejų laiptelių modeliui yra pavaizduoti 2 paveiksle. Matyti, kad visi siūlomi kriterijai (2 pav. c)–f)), skirti išretintoms lentelėms, yra daug galingesni už klasikinius, kurie faktiškai neskiria alternatyvos nuo nulinės hipotezės (2 pav. b)). Tačiau dviejų laiptelių su dispersija atveju, kai alternatyva yra nereguliari, yra stebimas priešingas reiškinys: klasikiniai vidurkių suderinamumo testai (3 pav. b)) yra kur kas galingesni už vidurkių suderinamumo testus sugrupuotiems duomenims (3 pav. d)–f)). Iš siūlomų naujų testų tik testas su MGMC glodinimu nenusileidžia klasikiniams (3 pav. c)).

Rezultatai, pateikti 4 paveiksle, rodo, kad sudėtinių grupavimo kriterijų (4 pav. a) ir c)) (kuriuose naudojamos kartu vidurkių ir dispersijų suderinamumo statistikos) galia ženkliai išauga, lyginant su vidurkių grupavimo kriterijais, o kriterijus, naudojantis GG grupavimą (4 pav. a)), yra galingiausias iš visų nagrinėjamų testų.

Bendrosios išvados

Sprendžiant disertacijos uždavinius, buvo gauti tokie rezultatai:

1. Labai išsklaidytų kategorinių duomenų atveju tiek tikėtinumo santykio statistika, tiek ir Pearsono χ^2 statistika gali būti neinformatyvios: jos jau neatspindi nulinės hipotezės ir duomenų neatitikimo masto. Taigi, kriterijai, kurie remiasi šiomis statistikomis, gali būti nepagrįsti net ir tuomet, kai alternatyva ženkliai skiriasi nuo nulinės hipotezės.
2. Išretintų kategorinių duomenų grupavimas, naudojant išplėstinį Bayeso modelį, bei jų glodinimas, taikant Markovo grandinių Monte Carlo metodą, kai alternatyva yra nereguliari, gali labai padidinti klasikinių testų galią.
3. Esant bendroms sąlygoms, testai, kuriuose naudojamas grupavimas, yra pagrįsti.
4. Grupavimo (glodinimo) efektas priklauso nuo grupavimo (glodinimo) metodo ir jo parametrų (grupių skaičiaus, kartojimų skaičiaus ir t. t.).
5. Esant nereguliarijai alternatyvai, kai tikėtinų dažnių nukrypimai nuo nulinės hipotezės reikšmių yra išsisklaidę chaotiškai (pavyzdžiui, yra

nepriklausomi centruoti atsitiktiniai gama dydžiai), geresnę galimybę atskirti hipotezę nuo alternatyvos turi grupavimo testai, kurie įvertina ir vidurkių, ir dispersijų suderinamumą grupėse.

Autoriaus mokslinių publikacijų disertacijos tema sąrašas

Recenzuojamuose mokslo žurnaluose

Radavičius, M.; Samusenko, P. 2012. Nonparametric testing for sparse nominal data, *Nonlinear Analysis: modeling and control*. ISSN 1392-5113. (Thomson ISI Web of Science). (Accepted for printing)

Radavičius, M.; Samusenko, P. 2011. Profile statistics for sparse contingency tables under poisson sampling, *Austrian Journal of Statistics* 40: 115–123. ISSN 1026-597X.

Radavičius, M.; Samusenko, P. 2011. Inconsistency of chi-square test for sparse categorical data under multinomial sampling, *Lietuvos Matematikos Rinkiny: LMD darbai* 52: 327–331. ISSN 0132-2818.

Radavičius, M.; Samusenko, P. 2010. Profile statistics for sparse contingency tables, *Computer Data Analysis and Modeling. Complex Stochastic Data and Systems, held in Minsk, Belarus* Minsk. BSU. 9(2): 55–58. ISBN 978-985-476-848-9.

Apie autorių

Pavel Samusenko gimė Vilniuje, 1983 m. sausio 19 d.

2005 m. baigė inžinerinės informatikos bakalauro studijas (Vilniaus Gedimino technikos universiteto Fundamentinių mokslų fakultete). 2007 m. įgijo taikomosios statistikos magistro laipsnį (Vilniaus Gedimino technikos universiteto Fundamentinių mokslų fakultete). 2008–2012 m. – doktorantūros studijos Vilniaus Gedimino technikos universiteto Matematinės statistikos katedroje.

Nuo 2007 m. dirba Vilniaus Gedimino technikos universitete: 2007–2011 m. – dirba asistentu, nuo 2011 m. – lektoriumi.

NONPARAMETRIC CRITERIA FOR SPARSE CONTINGENCY TABLES

Formulation of the problem

In the subfield of numerical analysis, a sparse table is a table populated primarily with zeros. The concept of sparsity is useful in combinatorics and application areas such as network theory, which have a low density of significant data or connections. Statistical inference problems caused by sparsity of contingency tables are widely discussed in the literature. Traditionally, the expected (under null the hypothesis) frequency is required to exceed 5 in almost all cells of the contingency table. If this condition is violated, the χ^2 approximations of goodness-of-fit statistics may be inaccurate and the table is said to be sparse Agresti. Several techniques have been proposed to tackle the problem: exact tests, alternative approximations, parametric and nonparametric bootstrap, Bayes approach and other methods, but they all are not applicable or have some limitations in nonparametric statistical inference of very sparse contingency tables.

The dissertation is devoted to the problem of consistent nonparametric testing for sparse contingency tables and the investigation of sparsity effect on the test power.

Topicality of the work

Recently amounts of information are very extensive, therefore problems related to a large dimension and/or sparsity of data arise rather frequently. The sparsity problem is especially topical for categorical data. Relationships between continuous variables are usually described by covariance matrices. Thus, the number of model parameters increases quadratically with n , the dimension of the data. For categorical data, the number of unknown parameters grows exponentially with n . Consequently, even for a moderate number of categorical variables, many cells in the contingency table are empty or have small counts. In fact, for categorical data, the number of cells in the corresponding contingency table is even more important characteristic of sparsity than the dimensionality k itself. Sometimes the number of cells (the number of unknown parameters) is even greater than the sample size.

In general, consistent estimator for expected cell counts of sparse contingency table does not exist. For categorical data of ordered variables, kernel smoothing or grouping enables one to obtain consistent estimators of expected cell counts under additional assumptions of their smoothness. For nominal data, the main object of our study, both kernel smoothing and grouping are not directly applicable.

Research object

The research objects are definitions of sparsity, sparse contingency tables of nominal data, goodness-of-fit statistics based on power divergences, consistency and finite-sample properties of nonparametric tests.

The aim and tasks of the work

The aim of this work is to investigate asymptotic and finite-sample behavior of classical goodness-of-fit statistics and tests for sparse contingency tables and to propose improvements of the classical tests.

The tasks of this work are:

1. To propose improvements of the classical tests in order to increase their sensitivity to deviations from null hypotheses for sparse nominal data.
2. To prove the consistency of the proposed tests.
3. To compare the finite-sample performance of the classical and proposed tests under alternatives of various types by means of computer simulations.

Applied methods

Sparse asymptotics modeling is based on latent distribution model, structural distribution model for large number or rare events (LNRE, by Khmaladze) and the empirical Bayes approach. Goodness-of-fit is measured by power divergences. Results for likelihood ratio criterion with profile (spectral) statistics substituted for cell counts rely on general likelihood theory and model-based cluster analysis. Chebyshev type inequalities are used to establish the consistency (inconsistency) of the tests. Finite-sample performance of the tests is studied by Monte-Carlo simulations. All calculations are performed using the R software.

Scientific novelty

It is shown that, for (very) sparse data, the likelihood ratio statistic and Pearson's χ^2 statistic may become noninformative: they do not anymore measure the goodness-of-fit of null hypotheses to data. For instance, they can be inconsistent even in cases where a simple consistent test does exist.

A new sparse asymptotics model relying on (extended) empirical Bayes approach is introduced. Tests based on MCMC smoothing, on smoothing by grouping or on (finite) mixtures of Poisson distributions are proposed. The consistency of the tests based on grouping is proved.

Practical value of the work results

The proposed nonparametric tests can be applied for statistical inference of high-dimensional categorical (nominal) data frequently met in surveys with large questionnaires, natural language and text processing, genetic data, etc. The tests

are easy to implement, they are computationally efficient and do not impose any specific requirements on the categorical (nominal) data they are applied to.

Statements presented for defence

1. For (very) sparse data, the likelihood ratio statistic and Pearson's χ^2 statistic may become noninformative: they do not anymore measure the goodness-of-fit of the null hypotheses to data.
2. Sparse asymptotics based on (extended) empirical Bayes approach enables one to apply latent distribution model to sparse nominal data.
3. In the empirical Bayes setting, MCMC smoothing, smoothing by grouping and modeling by finite mixtures of Poisson distributions can improve the power of classical tests especially for regular alternatives.
4. Under general conditions, the tests based on grouping are consistent.

Approval of the work results

Two classical goodness-of-fit statistics were studied through simulated data. For the latter the introduced estimator studied in this work based on the profile statistics of the contingency table was tested and proposed for a very sparse contingency tables. It is easy to implement, computationally efficient and do not impose any specific requirements on the data.

The scope of the scientific work

The dissertation consists of the introduction, four chapters, general conclusions, the references, the list of author's publications and one appendix. The total scope of the dissertation is 106 pages, 29 tables, 31 figures and 62 items of reference.

In Chapter 1, an overview of the problem is presented and basic definitions are introduced. Chapter 2 demonstrates the inconsistency of classical tests in case of (very) sparse categorical data for both multinomial and Poisson sampling scheme. In Chapter 3, extended Bayes model is introduced. It provides a basis for smoothing and grouping of sparse (nominal) data. The consistency of criteria based on grouping is proved. Finite-sample behavior of the classical and proposed criteria is studied in Chapter 4. Details of the computer simulation results are given in Appendix.

General conclusions

Having solved the tasks listed in the introduction the following results were obtained:

1. For very sparse contingency tables, the classical tests may have vary low power.

2. MCMC smoothing or grouping using the extended empirical Bayes model may considerably increase the power of the classical goodness-of-fit tests for sparse categorical data.
3. Under general conditions, the tests based on grouping are consistent.
4. The effect of grouping (smoothing) significantly depends on the grouping (smoothing) method as well as on its parameters (number of groups, number of iterations, etc.).
5. For the irregular alternatives that differ from the null hypothesis in non-smooth and chaotic manner, the tests which take into account discrepancies not only between the means but also between the variances are more powerful.

About the author

Pavel Samusenko was born in Vilnius on the 19rd of January, 1983.

He obtained the bachelor degree in Engineering informatics at the Vilnius Gediminas Technical University in 2005, and a master's degree in statistics at the Vilnius Gediminas Technical University in 2007. In 2008–2012 he'd been undertaking the doctoral studies in Mathematics at the Vilnius Gediminas Technical University.

He has been working at the Vilnius Gediminas Technical University since 2007.

Pavel SAMUSENKO

NEPARAMETRINIAI KRITERIJAI
RETŲ ĮVYKIŲ DAŽNIŲ LENTELĖMS

Daktaro disertacijos santrauka
Fiziniai mokslai, Matematika (01P)

NONPARAMETRIC CRITERIA
FOR SPARSE CONTINGENCY TABLES

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

2012 12 21. 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 „Ciklonas“,
J. Jasinskio g. 15, LT-01111 Vilnius