VILNIAUS UNIVERSITETAS
ŠIAULIŲ AKADEMIJA

MATEMATIKOS MAGISTRO STUDIJŲ PROGRAMA
Didžiųjų duomenų analitikos specializacija

MONIKA GELEŽINYTĖ

Magistro studijų baigiamasis darbas

KULBAKO-LEIBLERO INFORMACIJOS TAIKYMAS TIKRINANT STATISTINES HIPOTEZES

Darbo vadovas: doc. dr. Vaidotas Kanišauskas

Šiauliai, 2021

Studijuojančiojo, teikiančio baigiamąjį darbą, GARANTIJA

WARRANTY of Final Thesis

Vardas, pavardė *Name, Surname*	**Monika Geležinytė**
Padalinys *Faculty*	**Šiaulių akademija** ***Šiauliai Academy***
Studijų programa *Study Programme*	**Matematikos magistro studijų programa** ***Master's Degree Program in Mathematics***
Darbo pavadinimas *Thesis topic*	**Kulbako-Leiblero informacijos taikymas tikrinant statistines hipotezes** ***The Application of Kulback-Leibler Information for the Testing of Statistical Hypotheses***
Darbo tipas *Thesis type*	**Baigiamasis darbas** ***Final Thesis***

Garantuoju, kad mano baigiamasis darbas yra parengtas sąžiningai ir savarankiškai, kitų asmenų indėlio į parengtą darbą nėra. Jokių neteisėtų mokėjimų už šį darbą niekam nesu mokėjęs.

I guarantee that my thesis is prepared in good faith and independently, there is no contribution to this work from other individuals. I have not made any illegal payments related to this work.

Šiame darbe tiesiogiai ar netiesiogiai panaudotos kitų šaltinių citatos yra pažymėtos literatūros nuorodose.

Quotes from other sources directly or indirectly used in this thesis, are indicated in literature references.

Aš, Monika Geležinytė, pateikdamas (-a) šį darbą, patvirtinu (*pažymėti*) ☐

Embargo laikotarpis
Embargo Period

Prašau nustatyti šiam baigiamajam darbui toliau nurodytos trukmės embargo laikotarpį:
I am requesting an embargo of this thesis for the period indicated below:

☐ _____ mėnesių / *months*
(embargo laikotarpis negali viršyti 60 mėn. / *an embargo period shall not exceed 60 months*).

☐ Embargo laikotarpis nereikalingas / *no embargo requested.*

Embargo laikotarpio nustatymo priežastis / *Reason for embargo period*:

TURINYS

ĮVADAS

Darbo tikslas – ištirti, kaip susieta Kulbako-Leiblero informacija su statistinių hipotezių tikrinimu.

Uždaviniai:

1. Įsisavinti Kulbako informacinę teoriją, kur taikoma Kulbako-Leiblero informacija statistinėms hipotezėms tikrinti.
2. Nustatyti polinominio skirstinio Kulbako-Leiblero informacijos taikymo statistinėms hipotezėms tikrinti procedūrinius žingsnius.
3. Pritaikyti polinominiam skirstiniui skirtus Kulbako-Leiblero informacijos procedūrinius žingsnius binominio skirstinio paprastosioms ir sudėtingosioms statistinėms hipotezėms tikrinti. Gautas išvadas iliustruoti pavyzdžiais su konkrečiais duomenimis.
4. Parodyti, kaip taikant Kulbako-Leiblero informaciją yra tikrinamos paprastosios ir sudėtingosios polinominio skirstinio statistinės hipotezės, gautas procedūras iliustruojant pavyzdžiais su realiais duomenimis.

Informacijos teorija yra matematikos teorijos šaka, apimanti tikimybių ir matematinės statistikos kryptis. Pirmą kartą ši teorijos šaka buvo pristatyta viename iš R. A. Fišerio darbų 1925 m. (žiūr. [8] istoriografiją).

Informacijos teorija buvo nuolat plečiama. Vieną didžiausių indėlių į šios matematikos šakos plėtrą įdėjo Solomonas Kulbakas, kuris 1959 m. išleido knygą „Information theory and statistics “ (pakartotinas leidimas 1978 m. [9]). Ši teorija glaudžiai susijusi su didžiaisiais nuokrypiais ir hipotezių tikrinimu. H. Chernoff buvo pirmasis, kuris pritaikė didžiuosius nuokrypius hipotezėms tikrinti, kai buvo stebimi nepriklausomi ir vienodai pasiskirstę stebėjimai [3]. S. Kulbako [9] ir H. Chernoff [3] darbuose susiduriama su Neimano-Pirsono kriterijumi, kurio vienos rūšies klaidos tikimybė, kai kita fiksuota, asimptotiškai nusakoma Kulbako-Leiblero informacija, kai neapibrėžtai didėja stebėjimų skaičius. Nemažai mokslininkų tęsė asimptotinius paprastųjų hipotezių tikrinimo tyrimus, kuriuose atsiranda Kulbako-Leiblero ir jo atskiro atvejo, Černovo, informacijos. Iš šių mokslininkų verta paminėti ukrainietį J. N. Linkovą [6], [7] ir lietuvius V. Kanišauską su L. Dronova-Platbarzde ([10], [12], [13]), kurie nemažai išplėtojo minėtus taikymus. Panašius hipotezių tikrinimo uždavinius nagrinėjo Igoris Vajda, Fridrichas Liese

ir Wolfgang Stummer ([4], [5], [1], [2], [14]), kurie teikė pirmenybę iškilių funkcijų (kurioms priklauso ir Kulbako-Leiblero informacija) tyrimams ir taikymams.

Šiame magistro darbe daugiausia dėmesio ir bus skiriama Kulbako-Leiblero informacinei teorijai: susipažinsime ir panagrinėsime knygoje gautus rezultatus bei pritaikysime juos binominiam ir polinominiam skirstiniams.

Magistro darbą sudaro: įvadas, teorija, tyrimas ir literatūra. Teorinėje dalyje trumpai supažindinama su statistinėmis hipotezėmis ir detaliau aprašoma Kulbako-Leiblero informacijos taikymo teorija. Dalyje „Tyrimas“ yra parodomas Kulbako-Leiblero informacijos pritaikymas binominiam ir polinomiam skirstiniams: nustatomi Kulbako-Leiblero polinominio skirstinio informacijos taikymo procedūriniai žingsniai statistinėms hipotezėms tikrinti; parodoma, kaip pritaikomi polinominiam skirstiniui skirti Kulbako-Leiblero informacijos procedūriniai žingsniai binominio skirstinio paprastų ir sudėtingų statistinių hipotezių tikrinimui; parodoma, kaip taikant Kulbako-Leiblero informaciją yra tikrinamos paprastosios ir sudėtingosios polinominio skirstinio statistinės hipotezės, gautas procedūras iliustruojant pavyzdžiais su konkrečiais duomenimis. Darbo pabaigoje pateikiamos išvados.

I. TEORINĖ DALIS

1. Statistinės hipotezės

1.1 Paprastosios hipotezės

Informacija pateikta remiantis [11] knyga. Hipotezių žymėjimai perdaryti, suderinant su [9] knygos žymėjimais, nes jie dažniau naudojami magistriniame darbe.

Tarkime, kad turime binarinį statistinį eksperimentą $\boldsymbol{\mathcal{E}} = (\mathcal{X}, B, \{P_2, P_1\})$ su imtimi $X \in \mathcal{X}$. Iš anksto nėra žinoma, kuris tikrasis imties X skirstinys: P_2 ar P_1.

1 apibrėžimas. Bet koks teiginys apie tikrąjį imties X skirstinį P^X vadinamas hipoteze. Kadangi mūsų atveju tėra dvi galimybės, tai tikriname tik dvi hipotezes:

$H_2: P^X = P_2,\ H_1: P^X = P_1$

(užrašas $P^X = P_i$ reiškia teiginį: tikrasis X skirstinys P^X yra P_i).

2 apibrėžimas. Hipotezė, sudaryta iš vienos galimybės (taško), vadinama paprastąja. Tokiu būdu sprendžiamas dviejų paprastųjų hipotezių tikrinimo uždavinys.

3 apibrėžimas. Hipotezių H_2 ir H_1 tikrinimo (nerandomizuotu) statistiniu kriterijumi vadiname kiekvieną statistiką $\delta = \delta(X)$, įgyjančią dvi reikšmes: 0 ir 1. Kai $\delta(X)$ įgyja reikšmę 0, tada hipotezė H_2 priimama, o H_1 atmetama, o kai įgyja reikšmę 1, tada hipotezė H_1 priimama, o H_2 atmetama.

Akivaizdu, kad taip apibrėžtas statistinis kriterijus $\delta(X)$ padalija galimas $\mathcal{X}$ reikšmes X į dvi nesikertančias aibes $\mathcal{X}_2 = \mathcal{X}/W$, kur priimama hipotezė H_2 (atmetama H_1), ir $\mathcal{X}_1 = W$, kur priimama hipotezė H_1 (atmetama H_2).

4 apibrėžimas. Sritis W, kurioje priimama hipotezė H_1 (atmetama H_2), vadinama kritine sritimi.

Kiekvienas statistinis kriterijus $\delta = \delta(X)$ gali būti išreikštas per W taip:

$$\delta(X) = \mathbb{I}_W(x) = \begin{cases} 1, jei\ x \in W, \\ 0, jei\ x \notin W. \end{cases}$$

Taigi hipotezė H_2 atmetama, jei konkreti imties reikšmė $x \in W$. Todėl kriterinės srities W nusakymas ekvivalentus kriterijaus $\delta(X)$ apibrėžimui.

Kriterijaus $\delta(X)$ (kritinės srities W) gerumą charakterizuoja skaičiai

$$\alpha_2 = \alpha_2(\delta) = P_2(\delta(X) = 1) = P_2(X \in W) = M_2\delta(X),$$

$$\alpha_1 = \alpha_1(\delta) = P_1(\delta(X) = 0) = P_1(X \notin W) = 1 - P_1(X \in W) = M_1\big(1 - \delta(X)\big),$$

kur M_i – vidurkis pagal P_i.

5 apibrėžimas. Skaičius α_2 (atitinkamai α_1), vadinamas kriterijaus $\delta = \delta(X)$ 2-osios (atitinkamai 1-osios) rūšies klaidos tikimybe, žymi tikimybę atmesti hipotezę H_2 (atitinkamai H_1), kai ji teisinga.

Kriterijus $\delta = \delta(X)$ bus tuo geresnis, kuo mažesnės jo abiejų rūšių klaidų tikimybės. Tačiau, jei stebėjimų X apimtis fiksuota, negalima valdyti abiejų rūšių klaidų tikimybių kartu. Vienas iš būdų gauti optimalų kriterijų – fiksuojant vienos rūšies klaidos tikimybę minimizuoti kitą. Tuo tikslu įvedama kriterijų klasė:

$$K_\alpha = \{\delta = \delta(X) : \alpha_1(\delta) \leq \alpha\},$$

kur $\alpha \in (0, 1)$ – duotas mažas skaičius, vadinamas kriterijaus δ reikšmingumo lygmeniu.

1.2 Sudėtingųjų hipotezių tikrinimas

Formulės ir apibrėžimai paimti iš [11] knygos. Tarkime, kad turime statistinį eksperimentą $\mathcal{E} = (\mathcal{X}, B, P_\theta, \theta \epsilon \Theta)$, su imtimi $X \in \mathcal{X}$. Tegul P^X – tikrasis X skirstinys. Norima patikrinti hipotezes:

$$H_2 : P^x \in P_2 = \{P_\theta, \quad \theta \epsilon \Theta_2\},$$
$$H_1 : P^x \in P_1 = \{P_\theta, \quad \theta \epsilon \Theta_1\},$$

kur $\Theta_1 \cup \Theta_2 = \Theta$ ir $\Theta_1 \cap \Theta_2 = \emptyset$.

Akivaizdu, kad šios hipotezės gali būti užrašomos pavidalu:

$$H_2 : \theta \epsilon \Theta_2, \; H_1 : \theta \epsilon \Theta_1,$$

kur parametras θ toks, kad $P^x = P_\theta$.

Tokios hipotezės, kuriose aibės Θ_2 ir Θ_1 sudarytos daugiau negu iš vieno elemento, yra vadinamos sudėtingosiomis.

Analogiškai paprastosioms hipotezėms sudėtingosios hipotezės tikrinamos statistiniu kriterijumi $\delta = \delta(X)$ – statistika, įgyjančia reikšmes 0 ir 1, arba kritine sritimi W.

Šiuo atveju antros ir pirmos rūšies klaidų tikimybės apibrėžiamos pagal formules:

$$\alpha_2(\theta) = \alpha_2(\theta, W) = P_\theta(X \in W), \quad \theta \epsilon \Theta_2,$$
$$\alpha_1(\theta) = \alpha_1(\theta, W) = P_\theta(X \notin W) = 1 - P_\theta(X \in W), \quad \theta \epsilon \Theta_1,$$

Naudinga įvesti kriterijaus (kritinės srities W), galios funkciją

$$\beta_W(\theta) = P_\theta(X \in W), \theta \epsilon \Theta,$$

kuri nurodo tikimybę atmesti hipotezę H_0, kai tikroji parametro reikšmė θ. Dabar antros ir pirmos rūšies klaidų tikimybės užrašomos taip:

$$\alpha_2(\theta) = \beta_W(\theta), \theta \epsilon \Theta,$$
$$\alpha_1(\theta) = 1 - \beta_W(\theta), \theta \epsilon \Theta_1.$$

1 apibrėžimas. Skaičius $\alpha \epsilon (0, 1)$, kuris nusakytas formule

$$\alpha_2(\theta) < \alpha, \theta \epsilon \Theta_2,$$

vadinamas kriterijaus reikšmingumo lygmeniu. Skaičius α paprastai parenkamas artimas nuliui: 0,1; 0,05; 0,025; 0,01.

2 apibrėžimas. α reikšmingumo lygmens W kritinės srities kriterijus vadinamas tolygiai galingiausiu (tarp α reikšmingumo lygmens kriterijų su kritinėmis sritimis $\widetilde{W}$, jei

$$\beta_W(\theta) = \max_{\widetilde{W}} \beta_{\widetilde{W}}(\theta),\ \theta \in \Theta_1.$$

3 apibrėžimas. α reikšmingumo lygmens W kritinės srities kriterijus vadinamas nepaslinktuoju, jei

$$\beta_W(\theta) = \boldsymbol{P}_\theta(X \in W) \geq \alpha, \theta \in \Theta_1.$$

1.3 Nulinė hipotezė

Formulės ir apibrėžimai paimti iš [11] knygos.

Iš hipotezių taikymo praktikos matyti, kad alternatyvioje hipotezėje H_1 yra įrašomas teiginys apie tikrinamąjį objektą, kuris nebuvo statistiškai patikrintas, o hipotezė H_2 įprastai žymi tai, kas anksčiau jau buvo žinoma. Akivaizdu, kad praktiką domina, ar nulinė hipotezė yra pakankamai patikima. Matematiškai šis klausimas suformuluojamas taip: ar parinktam reikšmingumo lygmeniui α hipotezė H_2 atmetama, ar priimama. Tuo tikslu apibrėžiama kritinė sritis W, su α susieta formule:

$$P(X^n \in W | H_2) = \alpha,$$

kuri žymi tikimybę atmesti hipotezę H_0, kai ji teisinga. Praktikas visada turi tik imties realizaciją $x^n = (x_1, \dots, x_n)$. Jei $x^n \in W$, tada hipotezė H_2 (su tikimybe α) atmetama. Jei $x^n \notin W$, tada hipotezė H_2 (su tikimybe $1 - \alpha$) priimama ir sakoma, kad stebėjimų duomenys neprieštarauja senai hipotezei H_2.

Dažniausiai yra suformuluojama tik (nulinė hipotezė) H_2, o alternatyvą H_1 reikia laikyti jos priešingybe.

Tokiuose uždaviniuose hipotezės H_2 tikrinimo kritinė sritis W paprastai turi vieną iš šių trijų pavidalų:

1. $W = \{X^n : K(X^n) > t_{1-\alpha}\}$,
2. $W = \{X^n : K(X^n) < t_\alpha\}$,
3. $W = \{X^n : K(X^n) < t_{\frac{\alpha}{2}}\ arba\ , K(X^n) < t_{1-\frac{\alpha}{2}}\}$,

kur t_p yra $K(X^n)$ skirstinio p eilės kvantilis. Čia $K = K(X^n)$ – tam tikra statistika, kurios skirstinys su sąlyga, kad teisinga hipotezė H_2, yra gerai žinomas.

Jei trečiu atveju statistikos $K(X^n)$ skirstinio tankio funkcija simetriška $0y$ ašies atžvilgiu, tai kritinė sritis W įgyja pavidalą:

$$W = \{X^n : |K(X^n)| > t_{1-\frac{\alpha}{2}}\},$$

kur $t_{1-\frac{\alpha}{2}}$ statistikos $K(X^n)$ skirstinio $1 - \frac{\alpha}{2}$ eilės kvantilis.

2. Kulbako-Leiblero informacinė teorija

Šiame skyriuje apibrėžimai ir teiginiai paimti iš [9] literatūros šaltinio 3-124 psl. Šiame ir tolimesniuose skyreliuose log žymėjimas reiškia logaritmą su pagrindu e.

2.1 Tikimybinių matų absoliutus tolydumas

Skyrelio medžiaga parengta naudojantis [9] literatūros šaltiniu (1 skyrius, 2 poskyris).

Tarkime, kad yra tikimybinė erdvė $(X, \mathcal{B}(X), \mu_i),\ i = 1,\ 2.$ $x \in X$ elementų rinkinys, o visų yra X pogrupių σ-algebra žymima $\mathcal{B}(X)$, kurių tikimybės matas yra $\mu_i,$ i $= 1,\ 2.$ Pora $(X, \mathcal{B}(X))$, kur X yra elementų erdvė, o $\mathcal{B}(X)$ jos poaibių σ – algebra, vadinama mačia erdve. Elementai X gali būti vienmačiai ar daugiamačiai, diskretūs ar tolydūs.

Sakykime, kad tikimybės matai μ_1 ir μ_2 yra absoliučiai tolydūs vienas kito atžvilgiu, žym. $\mu_1 \ll \mu_2$, tai yra, nėra rinkinio (įvykio) $E \in Y$, kuriam $\mu_1(E) = 0$ ir $\mu_2(E) \neq 0$, arba $\mu_1(E) \neq 0$ ir $\mu_2(E) = 0$.

2.2 Radono-Nikodimo išvestinė

Skyrelio medžiaga parengta naudojantis [9] literatūros šaltiniu (1 skyrius, 2 poskyris).

Tegul λ yra toks tikimybės matas, kad $\mu_1 \ll \lambda,\ \mu_2 \ll \lambda,$; pavyzdžiui, $\lambda = (\mu_1 + \mu_2)/2$. Tada egzistuoja funkcijos $f_i(x),\ i = 1{,}2,$ vadinamos apibendrintais tikimybės tankiais, apibrėžtais iki nulinės λ tikimybės, $0 < f_i(x) < \infty.\ i = 1{,}2$ tokiais, kad

$$\mu_i(E) = \int_E f_i(x) d\lambda(x),\ \ i = 1{,}2,$$

visiems $E \in \mathcal{B}(X)$.

Simbolis $[\lambda]$, tariamas kaip „iki mato λ nulinės tikimybės“, susijęs su X elementais, reiškia, kad teiginys yra teisingas, išskyrus rinkinį E, tokį, kuriame $E \in \mathcal{B}(X)$ ir $\lambda(E) = 0$.

Funkcija $f_i(x)$ dar vadinama Radono-Nikodimo išvestine ir rašoma

$$d\mu_i(x) = f_i(x)\lambda(x) \text{ arba } f_i(x) = d\mu_i/d\lambda.$$

Jei tikimybės matas μ yra absoliučiai nepertraukiamas λ tikimybės mato atžvilgiu ir tikimybės matas v yra absoliučiai tolydus tikimybės mato μ atžvilgiu, tai tikimybės matas v taip pat yra absoliučiai tolydus tikimybės λ atžvilgiu, o Radono-Nikodymo išvestinė tenkina formulę

$$\frac{dv}{d\lambda} = \frac{dv}{d\mu} \cdot \frac{d\mu}{d\lambda}[\lambda].$$

Jei $H_i,\ i = 1{,}2,$ tai yra hipotezė, kad X (bendram kintamajam naudojame X, o konkrečiai X reikšmei - x) yra iš statistinės populiacijos su tikimybės matu μ_i, tai iš Bayeso teoremos, arba sąlyginės tikimybės teoremos išplaukia, kad

$$P(H_i|x) = \frac{P(H_i)f_i(x)}{P(H_1)f_1(x) + P(H_2)f_2(x)}[\lambda],$$

iš kurio gaunama

$$log\frac{f_1(x)}{f_2(x)} = log\frac{P(H_1|x)}{P(H_2|x)} - log\frac{P(H_1)}{P(H_2)}[\lambda],$$

kur $P(H_i), i = 1,2$, yra H_i tikimybė, o $P(H_1|x)$ yra antrinė H_i tikimybė arba sąlyginė H_i tikimybė, atsižvelgiant į $X = x$.

Anksčiau minėtos lygybės logaritmų pagrindas yra nereikšmingas, iš esmės gali būti bet koks. Jeigu nenurodyta kitaip, naudojami natūralūs arba Naperijos logaritmai (su pagrindu e).

Dešiniojoje pusėje yra H_1 skirtumas $X = x$ prieš stebėjimą ir po stebėjimo. Šis skirtumas, kuris gali būti teigiamas arba neigiamas, gali būti laikomas informacija, gaunama stebint $X = x$, o tikimybės santykio logaritmas $log[\frac{f_1(x)}{f_2(x)}]$ apibrėžiamas kaip informacija $X = x$ už hipotezės H_1 nauda prieš H_2.

2.3 Kulbako-Leiblero informacijos apibrėžimas

Skyrelio medžiaga parengta naudojantis [9] literatūros šaltiniu (1 skyrius, 2 poskyris).

1 apibrėžimas. Vidutinė informacija hipotezės H_1 nauda prieš H_2 atžvilgiu, kai $x \in \mathcal{B}(X)$, duotam μ_1, yra

$$I(1:2;E) = \frac{1}{\mu_1(E)}\int_E log\frac{f_1(x)}{f_2(x)}d\mu_1(x) = \frac{1}{\mu_1(E)}\int_E f_1(x)log\frac{f_1(x)}{f_2(x)}d\lambda(x),$$
$$\mu_1 > 0,\ \mu_1(E) = 0,$$

su $d\mu_1(x) = f_1(x)d\lambda(x)$.

Kai E yra visa erdvė X, žymime $I(1:2)$, o ne $I(1:2:X)$. Tada vidutinė informacija tikrinant hipotezės H_1 naudą prieš H_2 lygi:

$$I(1:2) = \int log\frac{f_1(x)}{f_2(x)}d\mu_1(x) = \int f_1(x)log\frac{f_1(x)}{f_2(x)}d\lambda(x) = \int log\frac{P(H_1|x)}{P(H_2|x)}d\mu_1(x) - log\frac{P(H_1)}{P(H_2)}.$$

$I(1:2)$ vadinama Kulbako-Leiblero informacija.

2.4 Kulbako-Leiblero informacijos savybės

Skyrelio medžiaga parengta naudojantis [9] literatūros šaltiniu (2 skyrius, 1-4 poskyriai).

1) $I(1:2)$ yra beveik teigiamai apibrėžtas; tai yra $I(1:2) \geq 0$, su lygybe tada ir tik tada, jei $f_1(x) = f_2(x)[\lambda]$.

2) $I(1:2;\ X,Y) = I(1:2;\ X) + I(1:2;\ X|Y) = I(1:2;\ Y) + I(1:2;\ X|Y)$,

čia $I(1:2;\ X,Y) = \int_E\ f_1(x,y) log \frac{f_1(x,y)}{f_2(x,y)} dxdy.$

3) Jei X ir Y nepriklausomiems pagal H_i, $i\ =\ l, 2$, tai

$$I(1:2;\ X,Y)\ =\ I(1:2;\ X)\ +\ I(1:2;\ Y).$$

4) Visiems $\lambda(E) > 0$,

$$\int_E\ f_1(x)\ log \frac{f_1(x)}{f_2(x)} d\lambda(x) \geq \left(\int_E\ f_1(x)\, d\lambda(x) \right) log \frac{\int_E\ f_1(x)\, d\lambda(x)}{\int_E\ f_2(x)\, d\lambda(x)} = \mu_1(E) log \frac{\mu_1(E)}{\mu_2(E)},$$

su lygybe tada ir tik tada, kai $\frac{f_1(x)}{f_2(x)} = \frac{\mu_1(E)}{\mu_2(E)} [\lambda]$, $x \in E$.

5) Jei $E_i \in \mathcal{B}(X), i = 1,2. E_i \cap E_j = 0, i \neq j$ ir $X = \cup_i E_i$, tai

$$I(1:2) \geq \sum_i \mu_1(E_i) log \frac{\mu_1(E_i)}{\mu_2(E_i)},$$

su lygybe tada ir tik tada, kai $\frac{f_1(x)}{f_2(x)} = \frac{\mu_1(E)}{\mu_2(E)} [\lambda]$, $x \in E_i, i = 1,2, \dots$.

6) Teisingos nelygybės:

$I(1:2; X|Y) \geq I(1:2; X)$ su lygybe tada ir tik tada, kai $I(1:2; X|Y) = 0$;

$I(1:2; X|Y) \geq I(1:2; Y)$ su lygybe tada ir tik tada, kai $I(1:2; X|Y) = 0$;

$I(1:2; X, Y) \geq I(1:2; Y|X)$ su lygybe tada ir tik tada, kai $I(1:2; X) = 0$;

$I(1:2; X, Y) \geq I(1:2; X|Y)$ su lygybe tada ir tik tada, kai $I(1:2; Y) = 0$.

7) Teisinga nelygybė $I(1:2; X) \geq I(1:2; Y)$, $Y = T(x)$ su lygybe tada ir tik tada, jei

$$\frac{f_1(x)}{f_2(x)} = \frac{g_1\big(T(x)\big)}{g_2\big(T(x)\big)} [\lambda].$$

8) Jei μ_1 ir μ_2 yra bet kurie du homogeniniai m matai, tai $I(1:2; X) \geq I(1:2; Y)$, su lygybe tada ir tik tada, jei statistika $Y = T(x)$ yra pakankama m atžvilgiu.

9) Jei f yra skaičioji X funkcija, tai būtina ir pakankama sąlyga, kad egzistuoja tokia išmatuojama funkcija g iš Y, kad $f\ = gT$ yra ta, kad f būtų išmatuojama $\mathrm{T}^{-1}(T)$; jei egzistuoja tokia funkcija g, tai ji yra vienintelė.

10) $I(1:2; T^{-1}(G)) = I(1:2; G)$ visiems $G \in T$ tada ir tik tada, jei $I(1:2; X) = I(1:2; Y)$; tai yra, tik tada, jei $Y = T(x)$ yra pakankama statistika.

11) $\int \varphi(x) f_1(x) log \frac{f_1(x)}{f_2(x)}\ d\lambda(x) \geq \int \psi(y) g_1(y) log \frac{g_1(x)}{g_2(x)} d\gamma(y)$, su lygybe tada ir tik tada, jei $Y = T(x)$ yra pakankama statistika

2.5 $\mathbf{J(1:2)}$ informacijos savybės

Skyrelio medžiaga parengta naudojantis [9] literatūros šaltiniu (2 skyrius, 5 poskyris).

2 apibrėžimas. Informacinis atstumas *J(1: 2)* apibrėžiamas pagal formulę:

$$J(1:2) = I(1:2) + I(2:1) = \int_E (f_1(x) - f_2(x)) log \frac{f_1(x)}{f_2(x)} d\lambda(x).$$

1) Jei X ir Y nepriklausomi, tai

$$J(1:2;\, X,Y) = J(1:2;\, X) + J(1:2;\, Y).$$

2) Bet kuriems X ir Y

$$J(1:2;\, X,Y) = J(1:2;\, X) + J(1:2;\, X|Y) = J(1:2;\, Y) + J(1:2;\, X|Y).$$

3) $J(1:2)$ yra beveik teigiamai apibrėžtas; tai yra $J(1:2) \geq 0$, su lygybe tada ir tik tada, jei $f_1(x) = f_2(x)[\lambda]$.

4) Visiems $\lambda(E) > 0$,

$$\int_E (f_1(x) - f_2(x))\, log \frac{f_1(x)}{f_2(x)} d\lambda(x) \geq \left(\int_E f_1(x)\, d\lambda(x) - \int_E f_2(x)\, d\lambda(x) \right) log \frac{\int_E f_1(x)\, d\lambda(x)}{\int_E f_2(x)\, d\lambda(x)},$$

$$= (\mu_1(E) - \mu_2(E)) log \frac{\mu_1(E)}{\mu_2(E)}$$

su lygybe tada ir tik tada, kai $\frac{f_1(x)}{f_2(x)} = \frac{\mu_1(E)}{\mu_2(E)}[\lambda]$, $x \in E$.

5) Jei $E_i \in Y, i = 1,2. E_i \cap E_j = 0, i \neq j$ ir $X = \cup_i E_i$, tai

$$J(1:2) \geq \sum_i \mu_1(E_i) - \mu_2(E_i)) log \frac{\mu_1(E_i)}{\mu_2(E_i)}$$

su lygybe tada ir tik tada, kai $\frac{f_1(x)}{f_2(x)} = \frac{\mu_1(E)}{\mu_2(E)}[\lambda]$, $x \in E_i, i = 1,2, \dots$.

6) Teisingos nelygybės:

$J(1:2; X|Y) \geq J(1:2; X)$ su lygybe tada ir tik tada, kai $J(1:2; X|Y) = 0$;

$J(1:2; X|Y) \geq J(1:2; Y)$ su lygybe tada ir tik tada, kai $J(1:2; X|Y) = 0$;

$J(1:2; X,Y) \geq J(1:2; Y|X)$ su lygybe tada ir tik tada, kai $J(1:2; X) = 0$;

$J(1:2; X,Y) \geq J(1:2; X|Y)$ su lygybe tada ir tik tada, kai $J(1:2; Y) = 0$.

7) $J(1,2; X) \geq J(1,2; Y)$, tada ir tik tada, kai galioja ši lygybė: $\frac{f_1(x)}{f_2(x)} = \frac{g_1(T(x))}{g_2(T(x))}[\lambda]$.

8) Jei μ_1 ir μ_2 yra bet kurie du homogeniško matų m rinkinio nariai, tai $J(1:2; X) \geq J(1:2; Y)$, su lygybe tada ir tik tada, jei statistika $Y = T(x)$ pakankama homogeninei aibei m.

9) $J(1:2; T^{-1}(G)) = J(1:2; G)$ visiems $G \in T$ tada ir tik tada, jei $J(1:2; X) = J(1:2; Y)$; tai yra, tik tada, jei $Y = T(x)$ yra pakankama statistika.

10) Teisinga nelygybė

$$\int \varphi(x)(f_1(x) - f_2(x)) log \frac{f_1(x)}{f_2(x)}\, d\lambda(x) \geq \int \psi(y)(g_1(y) - g_2(y)) log \frac{g_1(x)}{g_2(x)} d\gamma(y),$$

su lygybe tada ir tik tada, jei $Y = T(x)$ yra pakankama statistika.

2.6 Geriausio tankio $f^*(x)$, atitinkančio $f_1(x)$, radimas

Skyrelio medžiaga parengta naudojantis [9] literatūros šaltiniu (3 skyrius, 2 poskyris).

Kadangi $I(1:2) \geq 0$, su lygybe tada ir tik tada, kai $f_1(x) = f_2(x)[\lambda]$, tai, akivaizdu, kad reikia nustatyti tam tikrą papildomą apribojimą $f_1(x)$, jei norima rasti „artimiausią“ tikimybinį tankį, kuris skiriasi nuo tankio, atitinkančio tikimybinį matą μ_2. Reikalaujama, kad $f_1(x)$ būtų toks, kad $I(1:2)$ būtų mažiausia, su sąlyga kad $\int T(x) f_1(x) d\lambda(x) = \theta$, kur θ yra konstanta, o $Y = T(x)$ – išmatuojama statistika. Daugeliu atvejų θ yra daugiamačių populiacijų parametras. Jis taip pat gali atspindėti kai kurias kitas norimas populiacijų savybes.

Pagrindinis principas yra tas, kad $f_2(x)$ yra susietas su nulinės hipotezės populiacijos rinkiniu, o $f_1(x)$ susijęs su alternatyvios hipotezės populiacijos rinkiniu. Imties reikšmės naudojamos nustatyti „panašumui“ tarp imties, kaip galimo alternatyvios hipotezės populiacijos rinkinio nario ir artimiausios nulinės hipotezės populiacijos rinkinio, įvertinant mažiausiai pakitusią informaciją $I(1:2)$. Nulinė hipotezė bus atmesta, jei apskaičiuotas mažiausias informacijos atstumas $I(1:2)$ yra reikšmingai didesnis. Reikalavimai ekvivalentūs minimizacijai dydžio

$$\int \left(f_1(x) \log \frac{f_1(x)}{f_2(x)} + kT(x) f_1(x) + I f_1(x) \right) d\lambda(x),$$

esant pastoviems, laisvai pasirinktiems koeficientams *k* ir *l*.

Pažymima $g(x) = \frac{f_1(x)}{f_2(x)}$, tada nagrinėjamą dydį galima perrašyti pavidalu

$$\int (g_1(x) \log g(x) + kT(x) g(x) + I g(x)) d\mu_2(x).$$

Apibrėžiama pagalbinė funkcija

$$\varphi(t) = t log t + kTt + lt, t_0 = e^{-kT(x)-l-1}.$$

Tada $\varphi(t) = \varphi(t_0) + (t - t_0)\varphi'^{(t_0)} + \frac{1}{2}(t - t_0)^2 \varphi''(t_1)$, kur t_1 yra intervale tarp t ir t_0.

$$\varphi(t_0) = -t_0,\ \varphi''(t_1) = \frac{1}{t_1} > 0.$$

Tada $\int \varphi\big(g(x)\big) d\,\mu_2(x) = -\int e^{--kT(x)-l-1} d\mu_2(x) + \frac{1}{2}\int (g(x) - e^{-kT(x)-l-1})^2 \frac{d\mu_2(x)}{h(x),}$,

kur $h(x)$ yra intervale tarp $g(x)$ ir $e^{-kT(x)-l-1}$.

Galima pastebėti, kad $\int \varphi\big(g(x)\big) d\,\mu_2(x) \geq -\int e^{--kT(x)-l-1} d\mu_2(x)$,

tada ir tik tada, kai $g(x) = e^{-kT(x)-l-1}[\lambda]$.

Taigi minimali reikšmė yra

$$f_1(x) = f^*(x) = f_2(x) e^{-kT(x)-l-1}[\lambda]$$

ir tai reiškia, kad

$$I(*:2) + k\theta + l = -\int f_2(x)e^{-kT(x)-l-1}\mathrm{d}\lambda(\mathrm{x}).$$

Jeigu $-k$ pakeisti į τ ir nurodyti, kad $M_2(\tau) = \int f_2(x)e^{-\tau T(x)}\mathrm{d}\lambda(\mathrm{x})$, $M_2(\tau) < \infty$, išeina, kad minimali besiskirianti vidutinė informacija yra

$$I(*:2) = \theta\tau - logM_2(\tau),$$

kur $\theta = \int T(x)f^*(x)d\lambda(\mathrm{x}) = \int \frac{T(x)f_2(x)e^{-\tau T(x)}\mathrm{d}\lambda(\mathrm{x})}{M_2(\tau)} = \frac{\left(\frac{d}{d\tau}\right)M_2(\tau)}{M_2(\tau)}$,

visiems τ, esantiems intervalo viduje, kuriame $M_2(\tau)$ yra baigtinė. Toliau žymima τ kaip $\tau(\theta)$, kai bus svarbu nurodyti, kad τ yra θ funkcija.

1 TEOREMA. Jei $f_1(x)$ ir duotas $f_2(x)$ yra apibendrinti dominuojančių tikimybės matų rinkinio tankiai, $Y = T(x)$ yra išmatuojama statistika, tokia, kad egzistuoja $\theta = \int T(x)f_1(x)d\lambda(x)$ ir egzistuoja $M_2(\tau) = \int f_2(x)e^{\tau T(x)}d\lambda(x)$ su tam tikru intervalu τ; tada

$$I(1:2) \geq \theta\tau - logM_2(\tau) = I(*:2), \qquad \theta = \frac{d}{d\tau}logM_2(\tau),$$

tada ir tik tada, kai $f_1(x) = f^*(x) = \frac{e^{\tau T(x)}f_2(x)}{M_2(\tau)[\lambda]}$.

Tokiam $f^*(x)$ galima apskaičiuoti informaciją:

$$J(*,2) = \int(f*(x) - f_2(x))log\frac{f^*(x)}{f_2(x)}d\lambda(x) = (\theta - E_2\big(T(x)\big))\tau,$$

kur $E_2(T(x)) = \int T(x)f_2(x)\,d\lambda(x)$.

Tarkime, $f_1(x)$, $f_2(x)$, $f(x)$ yra apibendrinti tikimybinių matų rinkinio tankiai. Kadangi $I(1:2)$ yra beveik teigiamai apibrėžtas − tai yra $I(1:2) \geq 0$, su lygybe tada ir tik tada, jei $f_1(x) = f_2(x)[\lambda]$, išeina, kad

$$\int f(x)\log\frac{f(x)}{f_2(x)}d\lambda(x) + \int f(x)\log\frac{f_2(x)}{f_1(x)}d\lambda(x) = \int f(x)\log\frac{f(x)}{f_1(x)}d\lambda(x) \geq 0$$

arba

$$\int f(x)\log\frac{f(x)}{f_2(x)}d\lambda(x) \geq \int f(x)\log\frac{f_1(x)}{f_2(x)}d\lambda(x),$$

su lygybe, tada ir tik tada, kai $f_1(x) = f_2(x)[\lambda]$.

Jei pirmoje teoremoje $T(x) = log\frac{f_1(x)}{f_2(x)}$, tai minimumas besiskiriančios informacijos bus lygus

$$I\big(f:f_2\big) = \int_E f(x)log\frac{f(x)}{f_2(x)}d\lambda(x),$$

su sąlyga, kad $\theta = \int T(x)f(x)d\lambda(x) = \int f(x)log\frac{f_1(x)}{f_2(x)}d\lambda(x)$.

Iš čia

$$\min I\big(f:f_2\big) = \theta\tau - logM_2(\tau),$$

$$M_2(\tau) = \int f_2(x) e^{\tau log \frac{f_1(x)}{f_2(x)}} d\lambda(x) = \int f_1(x)^\tau f_2(x)^{1-\tau} d\lambda(x),$$

$$\theta = \frac{d}{d\tau} log M_2(\tau) = \frac{\int (f_1(x))^\tau (f_2(x))^{1-\tau} log \frac{f_1(x)}{f_2(x)} d\lambda(x)}{\int f_1(x)^\tau f_2(x)^{1-\tau} d\lambda(x)},$$

$$f(x) = \frac{\exp\left(\tau \log \frac{f_1(x)}{f_2(x)}\right) f_2(x)}{M_2(\tau)} = \frac{f_1(x)^\tau f_2(x)^{1-\tau}}{M_2(\tau)}.$$

Jei $\theta = 0$, tai egzistuoja τ_0, $0<\tau_0<1$, toks, kad

$$f_0(x) = \frac{f_1(x)^{\tau_0} f_2(x)^{1-\tau_0}}{M_2(\tau_0)},$$

tai $I(f_0 : f_2) = -log M_2(\tau_0) = -\log m_2, \ \ m_2 = \inf_{0<\tau<1} M_2(\tau)$,

$$0 = \int_E f_0(x) log \frac{f_1(x)}{f_2(x)} d\lambda(x) = \int_E f_0(x) log \frac{f_0(x)}{f_2(x)} d\lambda(x) + \int_E f_0(x) log \frac{f_1(x)}{f_0(x)} d\lambda(x)$$

arba

$$I(f_0 : f_2) = \int_E f_0(x) log \frac{f_0(x)}{f_2(x)} d\lambda(x) = \int_E f_0(x) log \frac{f_0(x)}{f_1(x)} d\lambda(x) = I(f_0 : f_1).$$

3 apibrėžimas. Informacija $I(f_0 : f_2) = -log M_2(\tau_0) = -\log m_2, \ \ m_2 = \inf_{0<\tau<1} M_2(\tau)$, kur $M_2(\tau) = \int f_1(x)^\tau f_2(x)^{1-\tau} d\lambda(x)$, vadinama Černovo informacija.

2.7 Hipotezių tikrinimas su Kulbako-Leiblero informacija

Skyrelio medžiaga parengta naudojantis [9] literatūros šaltiniu (5 skyrius, 1-3 poskyriai).

Kai bus nagrinėjama Kulbako-Leiblero informacija, turint n nepriklausomų stebėjimų Q_n, pati informacija bus žymima pavidalu ($I(*: 2; Q_n)$, o jei apibrėžiami dydžiai keičiasi, tada už θ imamas įvertis $\hat{\theta}(x)$, o už τ imamas $\hat{\tau}(x) = \tau(\hat{\theta}(x))$ taip, kad

$$T(x) = \hat{\theta}(x) = \frac{d}{d\tau} log M_2(\tau), \text{ kai } \tau = \hat{\tau}(x) = \tau(\hat{\theta}(x)),$$

$$I(*: 2; \ Q_n) = \hat{\theta}(x)\tau(\hat{\theta}(x)) - log M_2\big(\tau(\hat{\theta}(x))\big) .$$

$I(*: 2; Q_n)$ taikymo taisyklė:

Kuo didesnė $I(*: 2; Q_n)$ reikšmė, tuo mažiau „panašumo“ tarp imties su stebėjimais Q_n ir populiacijos su tankiu $f_2(x)$. Kadangi $f_2(x)$ atitinka hipotezę H_2 (nulinę hipotezę), tai kuo didesnė $I(*: 2; Q_n)$ reikšmė, tai tuo labiau hipotezė H_1 skiriasi nuo H_2.

Statistika $\hat{I}(*: H) = \min_{f_2 \in H} \hat{I}(*: 2; Q_n)$ vad. minimalia hipotezių atskyrimo informacija tikrinant nulinę hipotezę H_2 prieš hipotezę H_1, atmetant H_2 tuo atveju, jei

$$P(\hat{I}(*: H_2) - \hat{I}(*: H_1) \geq c \mid H_2) \leq \alpha$$

su tam parinkta konstanta c.

$\hat{I}(*: H_2) - \hat{I}(*: H_1) \geq c$ nurodo kritinę sritį – sritį atmesti H_2, kai ji teisinga (imtis priklauso populiacijai H_2).

$\hat{I}(*:H)$ vadinsis minimaliosios atskiriamosios informacijos statistika ir ji bus naudojama nulinei hipotezei H_2 prieš alternatyvią hipotezę H_1 tikrinti, atmetant H_2 tuo atveju, jei $\{\hat{I}(*:H_2) - \hat{I}(*:H_1) \geq c|H_2\} \leq \alpha$. Atitinkamai parinkus konstantą c, prie kurios dydis $\hat{I}(*:H_2)$ turi viršyti $\hat{I}(*:H_1)$, tam, kad hipotezė H_2 būtų atmesta, galima reguliuoti I rūšies klaidos tikimybės dydį, t.y. atmetimo H_2 tikimybę, kai imtis yra iš H_2 populiacijos. Pastebima, kad ši procedūra garantuoja gauti kriterijų su reikiamomis savybėmis, kai kalbama apie II rūšies klaidos tikimybę, t.y. tikimybę priimti nulinę hipotezę H_2, kai imtis yra iš H_1 populiacijos.

Prieš nagrinėjant minimaliosios atskiriamosios informacijos statistiką, gali būti naudinga iliustruoti tą procedūrą. Toliau pateiktuose pavyzdžiuose nepaisyta tikimybių ir atsižvelgta tik į $\hat{I}(*:H_2) - \hat{I}(*:H_1) \geq c$ išraišką, tai yra, kritinę sritį arba imties reikšmes, kurios pagrindu atmetama nulinė hipotezė.

Pavyzdys. Tarkime, kad yra stebėjimas x, kuris iš tikrųjų gali būti n nepriklausomų stebėjimų pavyzdys, ir, norint patikrinti paprastąją nulinę hipotezę H_2, stebėjimas atliekamas iš populiacijos, kurios tankis $f_2(x)$ lyginamas su paprastąja alternatyvia hipoteze H_1. Stebėjimas atliekamas iš populiacijos, kurios tankis $f_1(x)$. Turint statistiką $T(x) = log\left(\frac{f_1(x)}{f_2(x)}\right)$, pagal vertinimo procedūrą yra $\hat{\theta} = log\left(\frac{f_1(x)}{f_2(x)}\right)$. Apibrėždami $N_2(\hat{\tau}_2)$ ir $N_1(\hat{\tau}_1)$, mes gauname

$$\hat{I}(*:H_2) = \hat{\tau}_2 \log\frac{f_1(x)}{f_2(x)} - \log M_2(\hat{\tau}_2),\ M_2(\tau) = \int (f_1(x))^{\tau}\left(f_2(x)\right)^{1-\tau} d\lambda(x),$$

$$log\frac{f_1(x)}{f_2(x)} = \frac{\int (f_1(x))^{\tau_2}\left(f_2(x)\right)^{1-\tau_2} log\frac{f_1(x)}{f_2(x)} d\lambda(x),}{\int (f_1(x))^{\tau_2}\left(f_2(x)\right)^{1-\tau_2} d\lambda(x)} = \frac{N_2(\hat{\tau}_2)}{M_2(\hat{\tau}_2)}.$$

Analogiškai išeina

$$\hat{I}(*:H_1) = \hat{\tau}_1 \log\frac{f_1(x)}{f_2(x)} - \log M_1(\hat{\tau}_1),\ M_1(\tau) = \int (f_1(x))^{1+\tau}\left(f_2(x)\right)^{-\tau} d\lambda(x),$$

$$log\frac{f_1(x)}{f_2(x)} = \frac{\int (f_1(x))^{1+\tau_1}\left(f_2(x)\right)^{-\tau_1} log\frac{f_1(x)}{f_2(x)} d\lambda(x),}{\int (f_1(x))^{1+\tau_1}\left(f_2(x)\right)^{-\tau_1} d\lambda(x)} = \frac{N_1(\hat{\tau}_1)}{M_1(\hat{\tau}_1)}.$$

Kadangi $\frac{N_2(\hat{\tau}_2)}{M_2(\hat{\tau}_2)} = \frac{N_1(\hat{\tau}_1)}{M_1(\hat{\tau}_1)} = \log\left(\frac{f_1(x)}{f_{2(x)}}\right)$, išeina, kad

$$\hat{\tau}_2 = \hat{\tau}_1 + 1, M_2(\hat{\tau}_2) = M_1(\hat{\tau}_1).$$

Atitinkamai

$$\hat{I}(*:H_2) - \hat{I}(*:H_1) = \hat{\tau}_2 log\frac{f_1(x)}{f_2(x)} - logM_2(\hat{\tau}_2) - (\hat{\tau}_2 - 1)log\frac{f_1(x)}{f_2(x)} + logM_1(\hat{\tau}_1) = log\frac{f_1(x)}{f_2(x)},$$

su kritine sritimi $\frac{f_1(x)}{f_2(x)} \geq c$. Tai yra galingiausia kritinė sritis, kaip rodo pagrindinė Neimano-Pirsono lema. Su tokia kritine sritimi apibrėžtas kriterijus vadinamas Neimano-Pirsono kriterijumi.

2.8 Polinominis skirstinys

2.8.1 c kategorijų ar klasių diskretaus skirstinio hipotezės

Skyrelio medžiaga parengta naudojantis [8] literatūros šaltiniu.

4 apibrėžimas. Yra tokia nepriklausomų bandymų seką, kad kiekviename bandyme gali įvykti nesutaikomi įvykiai $A_1, A_2, \dots, A_c$. Įvykio A_1 tikimybė kiekviename bandyme yra lygi $p_i > 0$ $(p_1 + p_2 + \cdots + p_c = 1)$ ir nepriklauso nuo kitų bandymų rezultatų. Per m bandymų įvykusių įvykių A_i skaičius žymimas X_i, $i = 1,2, \dots, c$. Toks atsitiktinio vektoriaus $X = (X_1, \dots, X_{c-1})$ tikimybinis skirstinys vadinamas polinominiu ir yra žymimas $X \sim P(m, p_1, \dots, p_c)$.

$$P(X_1 = m_1, X_2 = m_2, \dots, X_{c-1} = m_{c-1}) = \frac{m!}{m_1! \dots m_c!} {p_1}^{m_c} \dots {p_k}^{m_c},$$

kur $m \geq m_i \geq 0$ ir $m_1 + \cdots + m_c = m$.

Yra dviejų paprastųjų statistinių hipotezių H_1 ir H_2, atitinkančių c kategorijų ar klasių, diskretūs skirstiniai:

$$H_i: p_{i1,} p_{i2}, \cdots, p_{ic}, p_{i1} + p_{i2} + \cdots + p_{ic} = 1, i = 1, 2.$$

Dabar Kulbako-Leiblero informacija apskaičiuojama formulėmis:

$$I(1:2) = p_{11} \log \frac{p_{11}}{p_{21}} + p_{12} \log \frac{p_{12}}{p_{22}} + \cdots + p_{1c},$$

$$I(2:1) = p_{21} \log \frac{p_{21}}{p_{11}} + p_{22} \log \frac{p_{22}}{p_{12}} + \cdots + p_{1c} \log \frac{p_{2c}}{p_{1c}}.$$

Ir atitinkamai

$$J(1,2) = I(1:2) + I(2:1) = (p_{11} - p_{21}) \log \frac{p_{11}}{p_{21}} + (p_{12} - p_{22}) \log \frac{p_{12}}{p_{22}} + \cdots + (p_{1c} - p_{2c}) \log \frac{p_{1c}}{p_{2c}}.$$

Žinoma, kad $I(1:2) \geq 0$, $I(2:1) \geq 0$, $J(1,2) \geq 0$, kai lygybė tenkinama kiekvienu atveju, tik tada, jei $p_{1i} = p_{2i}$, tai yra, hipotezės reiškia tą pačią populiaciją.

Atitinkamai N atsitiktinių imčių O_N nepriklausomų stebėjimų atveju:

$$I(1:2; O_N) = NI(1:2) = N \sum_{i=1}^{c} p_{1i} \log(\frac{p_{1i}}{p_{2i}}),$$

$$I(2:1; O_N) = NI(2:1) = N \sum_{i=1}^{c} p_{2i} \log(\frac{p_{2i}}{p_{1i}}),$$

$$J(1:2; O_N) = NJ(1:2) = N \sum_{i=1}^{c} (p_{1i} - p_{2i}) \log(\frac{p_{1i}}{p_{2i}}).$$

2.8.2 Jungtinis skirstinys $\mathbf{p}^*(\mathbf{x})$

Skyrelio medžiaga parengta naudojantis [9] literatūros šaltiniu (6 skyrius, 3 poskyris).

Yra N narių polinominis skirstinys su c kategorijomis ar klasėmis:

$$p(x) = p(x_1, x_2, \cdots, x_c) = \frac{N!}{x_1! \cdots x_c!} {p_1}^{x_1} {p_2}^{x_2} \cdots {p_c}^{x_c},$$

kur $p_i > 0,\ i = 1, 2, \cdots,\ p_1 + \cdots + p_c = 1,\ x_1 + x_2 + \cdots + x_c = N.$

Tarkime, kad $p^*(x)$ yra bet koks c kategorijų polinominis skirstinys jungtinis polinominiam skirstiniui su c kategorijomis $p(x)$, atitinkantį 1 teoremą. Taip siekiama išvengti nenumatytų atvejų $p^*(x) \neq 0$ ir $p(x) = 0$.

1 lema. c kategorijų polinominį skirstinį $p(x)$, t. y.

$$p(x) = p(x_1, x_2, \cdots, x_c) = \frac{N!}{x_1! \cdots x_c!} {p_1}^{x_1} {p_2}^{x_2}, \cdots, {p_c}^{x_c}\ ,$$

atitinka jungtinis skirstinys $p^*(x)$ toks, kad

$$E^*(x_i) = \theta_i \text{ ir } \sum_{x_1 + \cdots + x_c - N} p^*(x) \log \frac{p^*(x)}{p(x)}$$

yra minimalus, t. y.

$$p^*(x) = \frac{e^{\tau_1 x_1 + \cdots + \tau_c x_c} p(x)}{(p_1 e^{\tau_1} + \cdots + p_c e^{\tau_c})^N} = \frac{N!}{x_1! \cdots x_c!} ((p_1{}^*)^{x_1} \cdots (p_c{}^*)^{x_c},$$

kur $p^*(x) = \frac{p_i e^{\tau_i}}{(p_1 e^{\tau_1} + \cdots + p_c e^{\tau_c})}, i = 1, 2, \cdots, c,$, yra su tikru τ parametru, o $\theta_i = (\frac{\partial}{\partial_{\tau_i}})\log(p_1 e^{\tau_1} + \cdots + p_c e^{\tau_c})^N$.

2.8.3 Vienos imties atvejis

Skyrelio medžiaga parengta naudojantis [9] literatūros šaltiniu (6 skyrius, 4.4 poskyris).

Tarkime, kad yra atsitiktinė imtis: $x_1, x_2, \cdots, x_c, x_1 + x_2 + \cdots + x_c = N$, su c kategorijų polinominiu skirstiniu ir norima patikrinti nulinę hipotezę H_2, kad imtis yra iš populiacijos, kuriai

$$H_2 \colon (p) = (p_1, p_2, \cdots, p_c),\ p_1 + p_2 + \cdots + p_c = 1,$$

ir palyginti su alternatyvia hipoteze H_1, kad imtis yra iš bet kuris galimas c kategorijų polinominis skirstinys.

Už jungtinį skirstinį, atitinkantį 1 teoremą, galima imti tokį, kuriam parametrai – geriausi nepaslinkti parametro įverčiai $\hat{\theta}_i = N\hat{p}_i{}^* = x_i, i = 1,2, \cdots, c$, kuriuos atitinka

$$\hat{\tau}_i = \log \frac{x_i}{N p_i},\ i = 1,2, \cdots, c.$$

Jungtinė Kulbako-Leiblero informacija šiuo atveju yra:

$$\hat{I}(*\,,2; O_N) = x_1 \log \frac{x_1}{N p_1} + \cdots + x_c \frac{x_c}{N p_c},$$

ir atitinkama

$$\hat{J}(*\,,2; O_N) = N \left[\left(\frac{x_1}{N} - p_1 \right) log\, \frac{x_1}{N p_1} + \cdots + \left(\frac{x_c}{N} - p_c \right) log\, \frac{x_c}{N p_c} \right].$$

Galima įsitikinti, kad

$$2\hat{I}(*,2;O_N) \approx \sum_{i=1}^{c} \frac{(x_i - Np_i)^2}{Np_i} = X^2, \qquad (1)$$

$$\hat{J}(*,2;O_N) \approx \frac{1}{2}\sum_{i=1}^{c} \frac{(x_i - Np_i)^2}{Np_i} + \frac{1}{2}\sum_{i=1}^{c} \frac{(x_i - Np_i)^2}{x_i} = \frac{1}{2}(X^2 + {X_1}^2).$$

Iš (1) formulės matyti, kad jei nulinė hipotezė H_2 teisinga, tai $2\hat{I}(*,2;O_N)$ turi asimptotiškai χ^2(c-1) skirstinį, kas sutampa su klasikiniu χ^2 kriterijaus taikymu statistinėms hipotezėms.

2.8.4 „Vienpusės“ polinominės hipotezės

Skyrelio medžiaga parengta naudojantis [9] literatūros šaltiniu (6 skyrius, 4.5 poskyris).

1) Tarkime, kad yra atsitiktinė N nepriklausomų stebėjimų imtis $x_1, x_2, \cdots, x_c$, $x_1 + x_2 + \cdots + x_c = N$ ir tikrinamos c kategorijų polinominio skirstinio statistinės hipotezės:

$$H_1: p_1 > \frac{1}{c}, \qquad p_1 + p_2 + \cdots + p_c = 1,$$

$$H_2: p_1 = p_2 = \cdots = p_c = \frac{1}{c},$$

Hipotezėms tikrinti sudaromos statistikas – Kulbako-Leiblero informacijos:

$$\hat{I}(H_1: H_2; Q_N) = \sum_{i=1}^{c} x_i \log \frac{cx_i}{N}, kai \;\; x_1 > \frac{N}{c};$$

$$\hat{I}(H_1: H_2; Q_N) = \sum_{i=2}^{c} x_i \log \frac{(c-1)x_i}{N - x_1}, kai \quad x_1 \le \frac{N}{c}.$$

Hipotezių kritinė sritis apibrėžiama taip:

$$W = \{2\hat{I}(H_1: H_2; Q_N) > \chi^2_{1-2\alpha}(c-1)\}.$$

Jei parenkama $p_i = \frac{1}{c}$, $i = 1,2\cdots, c,$ kur laisvės laipsniai yra asimptotinių χ^2-skirstinių laisvės laipsniai nulinei hipotezei H_2.

1 lentelė. Polinominio skirstinio Kulbako-Leiblero informacijų skirstiniai

Komponentas	Informacija I	Laisvės laipsnis
$2 - c \mid x_1$ kategorijoje	$2(x_2 \log \frac{(c-1)x_2}{N - x_1} + \cdots + x_c \log \frac{(c-1)x_c}{N - x_1})$	$c - 2$
$1 - (2 + \cdots + c)$ kategorijoje	$2(x_1 \log \frac{cx_1}{N} + (N - x_1) \log \frac{c(N - x_1)}{N(c-1)})$	1
$2\hat{I}(*: 2; Q_N)$	$2(x_1 \log \frac{cx_1}{N} + \cdots + x_c \log \frac{cx_c}{N})$	$c - 1$

2) Tarkime, kad yra atsitiktinė N nepriklausomų stebėjimų imtis $x_1, x_2, \cdots, x_c$, $x_1 + x_2 + \cdots + x_c = N$ ir tikrinamos c kategorijų polinominio skirstinio statistinės hipotezės:

$$H_1': p_1 = p > \frac{1}{c}, p_2 = p_3 = \cdots = p_c = \frac{1-p}{c-1},$$

$$H_2': p_1 = p_2 = \cdots = p_c = \frac{1}{c}.$$

Hipotezėms tikrinti sudaroma statistika – Kulbako-Leiblero informacija:

$$\hat{I}(H_1': H_2'; Q_N) = \begin{cases} x_1 \log\frac{cx_1}{N} + (N - x_1)\log\frac{(N - x_1)c}{N(c-1)}, x_1 > \frac{N}{c}, \\ 0, x_1 \le \frac{N}{c}. \end{cases}$$

Jei nulinė hipotezė H_2' teisinga, tai $2\hat{I}(H_1': H_2'; Q_N)$ turi asimptotiškai $\chi^2(1)$ skirstinį. Dėl dviejų sąlygų iš $\chi^2(1)$ lentelių α reikšmingumo lygmens reikšmių reikia imti atitinkančius 2α reikšmingumo lygmenis.

3) Tarkime, kad yra atsitiktinė N nepriklausomų stebėjimų imtis $x_1, x_2, \cdots, x_c$, $x_1 + x_2 + \cdots + x_c = N$ ir tikrinamos c kategorijų polinominio skirstinio statistinės hipotezės:

$$H_1'': p_1, p_2, \cdots, p_c, p_1 + p_2 + \cdots + p_c = 1,$$

$$H_2': p_1 = p, \qquad p_2 = \cdots = p_c = \frac{1-p}{c-1}.$$

Nulinė hipotezė H_2' paprastai nenurodo p vertės. Išbandomas x_1 komponentas, priklausantis nuo 2 iki c kategorijų, atsižvelgiant į tai, kad jis asimptotiškai pasiskirsto kaip χ^2 su (c $-$ 2) laisvės laipsniais pagal nulinę hipotezę H_2'. Hipotezėms tikrinti naudojama 2 lentelė.

2 lentelė. Polinominio skirstinio Kulbako-Leiblero informacijų skirstiniai

Komponentas	Informacija	Laisvės laipsnis
$2 - c \mid x_1$ kategorijoje	$2(x_2 \log\frac{(c-1)x_2}{N - x_1} + \cdots + x_c \log\frac{(c-1)x_c}{N - x_1})$	$c - 2$
$1 - (2 + \cdots + c)$ kategorijoje	$2(x_1 \log\frac{x_1}{Np} + (N - x_1)\log\frac{c(N - x_1)}{N(1-p)})$	1
$2\hat{I}(*: 2; Q_N)$	$2(x_1 \log\frac{x_1}{Np} + x_2 \log\frac{(c-1)x_2}{N(1-p)} + \cdots + x_c \log\frac{(c-1)x_c}{N(1-p)})$	$c - 1$

II. TYRIMAS. KULBAKO-LEIBLERO INFORMACINĖS TEORIJOS TAIKYMAI BINOMINIAM SKIRSTINIUI

1. Pagalbiniai rezultatai

1.1 Binominis skirstinys

Apibrėžimas. Sakoma, kad atsitiktinis dydis X turi binominį skirstinį, jei įgyja reikšmes 0, 1, 2, ... , n su tikimybėmis $P(X = k) = C_n^k p^k q^{n-k}$, q=1-p, k=0, 1, ..., n.

Pastebima, kad binominio skirstinio atsitiktinis dydis X gali būti užrašytas pavidalu:

$$X = Y_1 + Y_2 + \cdots + Y_n,$$

kur Y_i yra nepriklausomi Bernulio dydžiai:

$$Y_i = p^k q^{1-k}, \text{ q=1-p, k=0, 1.}$$

1.2 Binominio skirstinio absoliutus mato tolydumas

Poskyrio medžiaga parengta naudojantis [11] knyga (II skyrius, 1 paragrafas).

Sakoma, kad atsitiktinis dydis X yra diskretus ir įgyja reikšmes $x_1, x_2, \dots, x_n$ su tikimybėmis

$$f(x_i) = P(X = x_i),\ i = 1, 2, \dots, n.$$

Tada $\chi = \{x_1, x_2, \dots, x_n\}$, $\mathcal{B}(\chi)$ – visų χ poaibių sistema. Mačiojoje erdvėje $(\chi, \mathcal{B}(\chi))$ apibrėžiamas dar vienas matas λ:

$$\lambda(X = x_i) = \lambda(x_i) \equiv 1, x_i \epsilon\ \chi.$$

Tada Rodono-Nikodimo išvestinė (diskretus tankis) bus

$$\frac{dP}{d\lambda}(x_i) = f(x_i), x_i \epsilon\ \chi.$$

Analogiškai apibrėžiamas binominio skirstinio diskretų tankį vienetinio diskretaus mato λ atžvilgiu:

$$f(x) = C_n^x p^x q^{x-k}, \text{ q=1-p, x } \epsilon\{0, 1, \dots, n\}.$$

1.3 Binominio skirstinio jungtinis tankis

Tarkime, kad yra binominis skirstinys su diskrečiu tankiu $f_2(x) = C_n^x p^x q^{n-x}$, kai $n > 1$ ir $T(x) = x$. Randamas jo jungtinis skirstinys. Pirmiausia randamas $M_2(\tau)$, kuris pagal apibrėžimą apskaičiuojamas formule:

$$\mathrm{M_2(\tau) = \int f_2(x) e^{\tau T(x)} d\lambda(x).}$$

Kadangi skirstinys diskretus, tai

$$\mathrm{M_2(\tau) = \sum_{x=0}^{n} f_2(x) e^{\tau T(x)} = \sum_{x=0}^{n} f_2(x) e^{\tau x} = \sum_{x=0}^{n} C_n^x p^x q^{n-x} e^{\tau x} = (pe^{\tau} + q)^n.}$$

Iš čia

$$f^*(x) = \frac{e^{\tau T(x)} f_2(x)}{M_2(\tau)} = \frac{C_n^x p^x q^{n-x} e^{\tau x}}{(pe^{\tau}+q)^n} = C_n^x (p^*)^x (q^*)^{n-x},$$

kur $p^* = \frac{pe^{\tau}}{pe^{\tau}+q}$, $q^* = \frac{q}{pe^{\tau}+q}$.

1.4 Binominio skirstinio Kulbako-Leiblero informacija

Tarkime, kad yra du tikimybiniai matai P_i, atitinkantys binominio skirstinio tankius $f_i(x) = C_n^x {p_i}^x {q_i}^{n-x}$, i=1, 2, x ϵ { 0, 1, ..., n}.

Pagal Kulbako-Leiblero informacijos apibrėžimą

$$I(1:2) = \int f_1(x) log \frac{f_1(x)}{f_2(x)} d\lambda(x) = \sum_{x=1}^{n} C_n^x {p_1}^x {q_1}^{n-x} log \frac{C_n^x {p_1}^x {q_1}^{n-x}}{C_n^x {p_2}^x {q_2}^{n-x}}$$

$$= \sum_{x=1}^{n} C_n^x x {p_1}^x {q_1}^{n-x} log \frac{p_1}{p_2} + \sum_{x=1}^{n} C_n^x (n-x) {p_1}^x {q_1}^{n-x} log \frac{q_1}{q_2}$$

$$= n\, p_1 \log \frac{p_1}{p_2} + n q_1 \log \frac{q_1}{q_2} = n\left(p_1 \log \frac{p_1}{p_2} + q_1 \log \frac{q_1}{q_2}\right).$$

Tą rezultatą galima buvo gauti ir kitu būdu. Kadangi binominis atsitiktinis dydis užsirašo pavidalu $X = Y_1 + Y_2 + \cdots + Y_n$, kur Y_i yra nepriklausomi Bernulio dydžiai, tai pagal Kulbako-Leiblero informacijos savybes teisinga formulė:

$$I(1:2) = nI_0(1:2),$$

kur $I_0(1:2)$ Bernulio atsitiktinio dydžio Kulbako-Leiblero informacija apskaičiuojama formule:

$$I_0(1:2) = p_1 \log \frac{p_1}{p_2} + q_1 \log \frac{q_1}{q_2}.$$

Vadinasi, binominio skirstinio Kulbako-Leiblero informacija apskaičiuojama formule

$$I(1:2) = n\left(p_1 \log \frac{p_1}{p_2} + q_1 \log \frac{q_1}{q_2}\right).$$

1.5 Binominio skirstinio Černovo informacija

Tarkime, kad yra du tikimybiniai matai P_i, atitinkantys binominio skirstinio tankius $f_i(x) = C_n^x {p_i}^x {q_i}^{n-x}$, i=1, 2, x ϵ { 0, 1, ..., n}. Černovo informacija nusakoma formulėmis:

$$I(f_0 : f_2) = -log M_2(\tau_0) = -\log m_2, \quad m_2 = \inf_{0<\tau<1} M_2(\tau),$$

kur $M_2(\tau) = \int f_1(x)^{\tau} f_2(x)^{1-\tau} d\lambda(x)$.

Atliekami veiksmai:

$$M_2(\tau) = \int f_1(x)^{\tau} f_2(x)^{1-\tau} d\lambda(x) = \sum_{x=1}^{n} (C_n^x {p_1}^x {q_1}^{n-x})^{\tau} (C_n^x {p_2}^x {q_2}^{n-x})^{1-\tau} =$$

$$\sum_{x=1}^{n} C_n^x ({p_1}^{\tau} p_2^{1-\tau})^x (q_1^{\tau} q_2^{1-\tau})^{n-x} = ({p_1}^{\tau} p_2^{1-\tau} + q_1^{\tau} q_2^{1-\tau})^n.$$

$$I(f_0: f_2) = -n \inf_{0<\tau<1} \log(p_1{}^{\tau} p_2^{1-\tau} + q_1{}^{\tau} q_2^{1-\tau}).$$

Pastebima, kad tą formulę galima buvo gauti naudojant Bernulio dydžius ir formulę:

$$I(f_0: f_2) = nI_0(f_0: f_2),$$

kur $I_0(1: 2)$ Bernulio atsitiktinio dydžio Černovo informacija apskaičiuojama formule:

$$I_0(f_0: f_2) = -\inf_{0<\tau<1} \log(p_1{}^{\tau} p_2^{1-\tau} + q_1{}^{\tau} q_2^{1-\tau}) .$$

Vadinasi, binominio skirstinio Černovo informacija apskaičiuojama formule:

$$I(f_0: f_2) = -n \inf_{0<\tau<1} \log(p_1{}^{\tau} p_2^{1-\tau} + q_1{}^{\tau} q_2^{1-\tau}).$$

2. Binominio skirstinio statistinės hipotezės

2.1 Paprastųjų hipotezių tikrinimas panaudojant $\{\widehat{\mathbf{I}}(*:\mathbf{H_2}) - \hat{\mathbf{I}}(*:\mathbf{H_1}) \geq \mathbf{c}\}$

Tikrinamos binominio skirstinio paprastosios hipotezės $H_1: p = p_1,\ q = q_1;\ H_2: p = p_2,\ q = q_2$.

Tikrinant dvi paprastąsias hipotezes H_1 ir H_2 gali būti naudojama formulė

$$P(\hat{I}(*:H_2) - \hat{I}(*:H_1) \geq c \mid H_2) \leq \alpha$$

su tam parinkta konstanta c. T. y., $\hat{I}(*:H_2) - \hat{I}(*:H_1) \geq c$ nurodo kritinę sritį – sritį atmesti H_2 kai ji teisinga (imtis priklauso populiacijai H_2).

Procedūra panaudojama su binominiu skirstiniu, kai $f(x) = C_n^x p^x q^{n-x}$.

Tikriname hipotezes $H_1: p = p_1, q = q_1; H_2: p = p_2, q = q_2$.

Už parametrą $\theta = np^*$ įvertiname per n$\hat{p}$, kur n$\hat{p} = y, \hat{q} = 1 - \hat{p}$, y = T(x) – skaičius „sėkmių“ imtyje Q_n, t. y. $\hat{p} = \frac{y}{n}$. Dėl Binominio skirstinio Kulbako-Leiblero informacijos pavidalo turima:

$$\hat{I}(*:H_1) = n\left(\hat{p}\log\frac{\hat{p}}{p_1} + \hat{q}\log\frac{\hat{q}}{q_1}\right),$$

$$\hat{I}(*:H_2) = n\left(\hat{p}\log\frac{\hat{p}}{p_2} + \hat{q}\log\frac{\hat{q}}{q_2}\right).$$

Pagal taisyklę hipotezė H_2 atmetama, jei

$\{\widehat{I}(*:H_2) - \hat{I}(*:H_1) \geq c\}$, t. y., jei

$$n\left(\hat{p}\log\frac{\hat{p}}{p_2} + \hat{q}\log\frac{\hat{q}}{q_2}\right) - n\left(\hat{p}\log\frac{\hat{p}}{p_1} + \hat{q}\log\frac{\hat{q}}{q_1}\right) \geq c,$$

arba suprastinus, kai $\hat{p}\log\frac{p_1 q_2}{p_2 q_1} \geq c$.

Iš čia, kai $p_1 > p_2$, H_2 atmetama kai $\hat{p} > c$.

Jei $p_1 < p_2$, H_2 atmetama kai $\hat{p} \leq c$.

Iš tikimybių teorijos žinoma ([8], [15]), kad atsitiktinis dydis $\hat{p} = \frac{y}{n}$, kai n yra pakankamai didelis (n > 50) ir esant patenkintoms sąlygoms: n$\hat{p} > 5$, n$\hat{q}$ >5, gerai aproksimuojamas normaliuoju skirstiniu N(p, $\sqrt{p(1-p)/n}$). Iš čia seka, kad esant teisingai hipotezei H_2, statistika

$$Z = \frac{\hat{p} - p_2}{\sqrt{p_2(1-p_2)/n}} \sim N(0,1).$$

Vadinasi, esant užduotam reikšmingumo lygmeniui α, H_2 atmetamo (kritinė sritis W), kai $p_1 > p_2$, yra

$$W = \{ Z_{imties} > u_{1-\alpha} \},$$

kur $u_{1-\alpha}$ –N(0,1) skirstinio 1- α kvantilis. T. y.,

$$\frac{\hat{p}-p_2}{\sqrt{p_2(1-p_2)/n}} > u_{1-\alpha}.$$

Iš čia

$$\hat{p}-p_2 > u_{1-\alpha}\sqrt{p_2(1-p_2)/n} \text{ ir } \hat{p} > c, \text{ kai } c = p_2 + u_{1-\alpha}\sqrt{p_2(1-p_2)/n}.$$

Jei $p_1 < p_2$, esant užduotam reikšmingumo lygmeniui α, H_2 atmetamo (kritinė sritis W), yra

$$W=\{ Z_{imties} < u_\alpha \},$$

kur u_α –N(0,1) skirstinio α kvantilis. T. y.,

$$\frac{\hat{p}-p_2}{\sqrt{p_2(1-p_2)/n}} < u_\alpha.$$

Iš čia

$$\hat{p}-p_2 < u_\alpha\sqrt{p_2(1-p_2)/n} \text{ ir } \hat{p} < c, \text{ kai } c = p_2 + u_\alpha\sqrt{p_2(1-p_2)/n}.$$

Pavyzdžiai.

1. n= 100, y= 60, $H_2(H_0)$: p= 0,5, H_1: p= 0,7. α= 0,05.

   Sprendimas.

   Pagal sąlygą $p_1 = 0{,}7,\ p_2 = 0{,}5$. T. y., $p_1 > p_2$, vadinasi, H_2 hipotezė atmetama, kai $\hat{p} > c$, čia $c = p_2 + u_{1-\alpha}\sqrt{p_2(1-p_2)/n}$.

   Mūsų atveju $\hat{p} = \frac{y}{n} = \frac{60}{100} = 0{,}6,\ u_{1-\alpha} = u_{1-0{,}05} = u_{0{,}95} = 1{,}645,$

$$c= 0{,}5 + 1{,}645\sqrt{0{,}5\cdot 0{,}5/100} = 0{,}5 + 1{,}645\cdot 0{,}05 = 0{,}58225.$$

   Gauta $\hat{p} = 0{,}6 > c = 0{,}58225$. Vadinasi, H_2 hipotezė atmetama. Duomenys neprieštarauja hipotezės H_1: p= 0,7 teisingumui.

2. n= 100, y= 75, $H_2(H_0)$: p= 0,7, H_1: p=0,8. α= 0,1.

   Sprendimas.

   Pagal sąlygą $p_1 = 0{,}8,\ p_2 = 0{,}7$ T. y., $p_1 > p_2$, vadinasi, H_2 hipotezė atmetama, kai $\hat{p} > c$, čia $c = p_2 + u_{1-\alpha}\sqrt{p_2(1-p_2)/n}$.

   Šiuo atveju $\hat{p} = \frac{y}{n} = \frac{70}{100} = 0{,}7,\ u_{1-\alpha} = u_{1-0{,}1} = u_{0{,}90} = 1{,}282,$

$$c= 0{,}7 + 1{,}282\sqrt{0{,}7\cdot 0{,}3/100} = 0{,}7 + 1{,}282\cdot 0{,}0458 = 0{,}7587.$$

   Gauta $\hat{p} = 0{,}7 < c = 0{,}7587$. Vadinasi, $H_2(H_0)$: p= 0,7 hipotezė priimama ir duomenys prieštarauja hipotezės H_1: p= 0,8 teisingumui.

3. n= 64, y= 30, $H_2(H_0)$: p= 0,6, H_1: p=0,5. α= 0,01.

   Sprendimas.

   Pagal sąlygą $p_1 = 0{,}5,\ p_2 = 0{,}6$. T. y., $p_1 < p_2$, vadinasi H_2 hipotezė atmetama, kai

$\hat{p} \le c$, čia $c = p_2 + u_\alpha\sqrt{p_2(1-p_2)/n}$.

Šiuo atveju $\hat{p} = \frac{y}{n} = \frac{30}{64} = 0{,}46875,\ u_\alpha = u_{0{,}01} = -u_{1-0{,}01} = -2{,}326,$

$\text{c}= 0{,}6 - 2{,}326\sqrt{0{,}6 \cdot 0{,}4/64} = 0{,}6 - 2{,}326 \cdot 0{,}06124 = 0{,}45756.$

Gavome $\hat{p} = 0{,}46875 > \text{c} = 0{,}45756$. Vadinasi, H_2 hipotezė priimama. Duomenys neprieštarauja hipotezės H_2: p= 0,6 teisingumui.

4. n= 64, y= 16, $H_2(H_0)$: p= 0,3, H_1: p=0,2. α= 0,05.

Sprendimas.

Pagal sąlygą $p_1 = 0{,}2,\ p_2 = 0{,}3$. T. y., $p_1 < p_2$, vadinasi H_2 hipotezė atmetama, kai $\hat{p} \le c$, čia $c = p_2 + u_\alpha\sqrt{p_2(1-p_2)/n}$.

Šiuo atveju $\hat{p} = \frac{y}{n} = \frac{16}{64} = 0{,}25,\ u_\alpha = u_{0{,}05} = -u_{1-0{,}05} = -1{,}645,$

$\text{c}= 0{,}3 - 1{,}645\sqrt{0{,}3 \cdot 0{,}7/64} = 0{,}3 - 1{,}645 \cdot 0{,}05728 = 0{,}20577.$

Gauta $\hat{p} = 0{,}25 > \text{c} = 0{,}20577$. Vadinasi, H_2 hipotezė priimama. Duomenys neprieštarauja hipotezės H_2: p= 0,3 teisingumui.

2.2 Sudėtingosios binominės hipotezės

Tarkime, kad yra binominės populiacijos imtis, kai atliekama n nepriklausomų bandymų Bernulio schemoje, kur „pasisekimų“ skaičius yra x, o „nesėkmių“ skaičius n – x.

1 atv. Tikrinamos dvi hipotezės: H_1: $p_1 > p$, H_2: p= p. Jos tikrinamos taikant anksčiau išdėstytą Kulbako-Leiblero informacinę teoriją.

Nulinę hipotezę H_2 atitinka binominio skirstinio diskretus tankis su parametru p:

$f_2(x) = C_n^x p^x q^{n-x}$, kai $n > 1$, q=1-p.

Kaip jau buvo rasta, hipotezei H_1 priklauso jungtinis binominis skirstinys

$f^*(x) = C_n^x (p^*)^x (q^*)^{n-x}$, su parametrais $p^* = \frac{pe^\tau}{pe^\tau+q}$, $q^* = \frac{q}{pe^\tau+q}$,

kai $T(x) = x$, o $M_2(\tau) = (pe^\tau + q)^n$.

Pagal hipotezę : H_1: $p_1 > p$, o p_1 atitinka p^*, tai turime $p^* > p$ arba $\frac{pe^\tau}{pe^\tau+q} > p$ kas įmanoma tik tuo atveju, kai $\tau > 0$.

Kai nagrinėjama Kulbako-Leiblero informacija, turint n nepriklausomų stebėjimų Q_n, pati informacija žymima pavidalu ($I(*\colon 2)$, o jai apibrėžiami dydžiai keičiasi taip:

už θ imamas įvertis $\hat{\theta}(x)$, o už τ imamas $\hat{\tau}(x) = \tau(\hat{\theta}(x))$,

$T(x) = \hat{\theta}(x) = \frac{d}{d\tau} log M_2(\tau)$, kai $\tau = \hat{\tau}(x) = \tau(\hat{\theta}(x))$.

$I(*\colon 2;\ Q_n) = \hat{\theta}(x)\tau(\hat{\theta}(x)) - logM_2\big(\tau(\hat{\theta}(x))\big)$.

Šiuo atveju už jungtinio skirstinio parametrą θ imamas geriausias nepaslinktas p^* įvertis $\hat{\mathrm{p}}^* = \frac{\mathrm{x}}{n}$, t.y. $\hat{\theta} = \mathrm{n}\hat{\mathrm{p}}^* = \mathrm{x}$. Tada, kadangi $T(x) = x$ ir $M_2(\tau) = (pe^{\tau} + q)^n$, tai

$$\hat{\mathrm{I}}(\mathrm{p}^*\colon \mathrm{p}) = n\hat{\mathrm{p}}^* \log \frac{\hat{\mathrm{p}}^*}{p} + n\hat{q}^* \log \frac{\hat{q}^*}{q} = x \log \frac{\mathrm{x}}{np} + (n-x)\log \frac{\mathrm{n}-\mathrm{x}}{nq}.$$

Iš kitos pusės

$$I(*\colon 2;\ Q_n) = \hat{\theta}(x)\tau(\hat{\theta}(x)) - logM_2\big(\tau(\hat{\theta}(x))\big) = \hat{\tau}\mathrm{x} - \mathrm{n}log(\mathrm{pe}^{\hat{\tau}} + \mathrm{q}) = \hat{\tau}\mathrm{x} + \mathrm{n}\log \frac{\mathrm{n}-\mathrm{x}}{nq}.$$

Kadangi $\hat{\mathrm{I}}(\mathrm{p}^*\colon \mathrm{p}) = I(*\colon 2;\ Q_n)$. Tai

$$x \log \frac{\mathrm{x}}{np} + (n-x)\log \frac{\mathrm{n}-\mathrm{x}}{nq} = \hat{\tau}\mathrm{x} + \mathrm{n}\log \frac{\mathrm{n}-\mathrm{x}}{nq}.$$

Iš čia $\hat{\tau} = log\frac{xq}{p(n-x)} = log\frac{q\hat{\mathrm{p}}^*}{p\hat{q}^*}$.

Kai $x = np$, $\hat{\tau} = 0$. Kai $x > pn$, $\frac{\mathrm{x}}{n} = \hat{\mathrm{p}}^* > p$. Iš čia $-\hat{\mathrm{p}}^* < -p$, ir čia $1 - \hat{\mathrm{p}}^* < 1 - p$, t. y. $\hat{q}^* < q$. Vadinasi, kai $x > pn$ tai $\hat{\tau} = log\frac{q\hat{\mathrm{p}}^*}{p\hat{q}^*} > 0$ ir yra priimtinas. Jei $\mathrm{x} < \mathrm{np}$, $\hat{\tau} < 0$ ir yra nepriimtinas. Taigi pagal statistinius duomenis

$$\hat{\mathrm{I}}(\mathrm{H}_1\colon \mathrm{H}_2; \mathrm{O}_\mathrm{n}) = \begin{cases} \mathrm{x}\log\frac{\mathrm{x}}{\mathrm{np}} + (\mathrm{n}-\mathrm{x})\log\frac{\mathrm{n}-\mathrm{x}}{\mathrm{nq}}, \mathrm{x} > \mathrm{np} \\ 0, \mathrm{x} \le \mathrm{np}. \end{cases}$$

Jeigu H_2 nulinė hipotezė galioja, tai $2\hat{\mathrm{I}}(\mathrm{H}_1\colon \mathrm{H}_2; \mathrm{O}_\mathrm{n})$ asimptotiškai χ^2- pasiskirsto vienu laisvės laipsniu. Reikšmės χ^2 reikšmė, atitinkanti reikšmingumo lygį *α*, turėtų būti paimta iš įprastų χ^2 lygio *2α* lentelių, nes neatsižvelgiama į vertę $x < np$, kurioms $\hat{\mathrm{I}}(\mathrm{H}_1\colon \mathrm{H}_2; \mathrm{O}_\mathrm{n})$ turi tą pačią reikšmę, kaip ir kai kurioms *x > np* reikšmėms.

Vadinasi, hipotezių tikrinimo kritinė sritis, t. y. H_2 atmetimo sritis, nusakoma formule

W={ $2\hat{\mathrm{I}}(\mathrm{H}_1\colon \mathrm{H}_2; \mathrm{O}_\mathrm{n}) > \chi^2_{1-2\alpha}(1)$ }. $\chi^2_{1-2\alpha}(1)$ – yra Chi-kvadrat su vienu laisvės laipsniu 1-*2α* reikšmingumo lygmens kvantilis.

Pavyzdžiai.

1. $\mathrm{n} = 100$, $\mathrm{x} = 60$, $H_2\colon \mathrm{p} = 0{,}5$, $H_1\colon \mathrm{p} > 0{,}5$. $\alpha = 0{,}05$.

Sprendimas.

Pagal sąlygą, $\mathrm{x} = 60 > np = 100 \cdot 0{,}5 = 50$. Tai

$\hat{\mathrm{I}}(\mathrm{H}_1\colon \mathrm{H}_2; \mathrm{O}_\mathrm{n}) = \mathrm{x}\log\frac{\mathrm{x}}{\mathrm{np}} + (\mathrm{n}-\mathrm{x})\log\frac{\mathrm{n}-\mathrm{x}}{\mathrm{nq}} = 60 \cdot log\frac{60}{50} + (100-60)log\frac{100-60}{50} = 2{,}01355.$

$\chi^2_{1-2\alpha}(1) = \chi^2_{1-2x0{,}05}(1) = \chi^2_{0{,}9}(1) = 2{,}71.$

Kadangi $2\hat{\mathrm{I}}(\mathrm{H}_1\colon \mathrm{H}_2; \mathrm{O}_\mathrm{n}) = 4{,}0271 > \chi^2_{1-2\alpha}(1) = 2{,}71$, tai duomenys neprieštarauja hipotezės $H_1\colon \mathrm{p} > 0{,}5$ teisingumui.

2. $\mathrm{n} = 100$, $\mathrm{x} = 10$, $H_2\colon \mathrm{p} = 0{,}1$, $H_1\colon \mathrm{p} > 0{,}1$. $\alpha = 0{,}1$.

Sprendimas.

Pagal sąlygą, x= $10 \le np = 100 \cdot 0{,}1 = 10$. Tai

$\hat{I}(H_1: H_2; O_n) = 0$.

$\chi^2_{1-2\alpha}(1) = \chi^2_{1-2x0,1}(1) = \chi^2_{0,8}(1) = 1{,}64$.

Kadangi $2\hat{I}(H_1: H_2; O_n) = 0 < \chi^2_{1-2\alpha}(1) = 1{,}64$, tai duomenys neprieštarauja hipotezės H_2: p= 0,1 teisingumui.

3. n= 100, x= 10, H_2: p= 0,05, H_1: p> 0,05. α= 0,001; 0,1; 0,05

Sprendimas.

Pagal sąlygą, x= $10 > np = 100 \cdot 0{,}05 = 5$. Tai

$\hat{I}(H_1: H_2; O_n) = x \log\frac{x}{np} + (n - x)\log\frac{n-x}{nq} = 10 \cdot log\frac{10}{5} + (100 - 10)log\frac{100-10}{95}$ =2,06542.

$\chi^2_{1-2\alpha}(1) = \chi^2_{1-2x0,001}(1) = \chi^2_{0,998}(1) = 10{,}8$.

Kadangi $2\hat{I}(H_1: H_2; O_n) = 4{,}13084 < \chi^2_{1-2\alpha}(1) = 10{,}8$, tai duomenys neprieštarauja hipotezės H_2: p= 0,05 teisingumui.

Kai α= 0,1 $\chi^2_{1-2\alpha}(1) = \chi^2_{1-2x0,1}(1) = \chi^2_{0,8}(1) = 1{,}64$.

Tada $2\hat{I}(H_1: H_2; O_n) = 4{,}13084 > \chi^2_{1-2\alpha}(1) = 1{,}64$, tai duomenys neprieštarauja hipotezės H_1: p> 0,05 teisingumui.

Kai α= 0,05 $\chi^2_{1-2\alpha}(1) = \chi^2_{1-2x0,05}(1) = \chi^2_{0,9}(1) = 2{,}71$.

Tada $2\hat{I}(H_1: H_2; O_n) = 4{,}13084 > \chi^2_{1-2\alpha}(1) = 2{,}71$, tai duomenys neprieštarauja hipotezės H_1: p> 0,05 teisingumui.

4. n= 100, x= 60, H_2: p= 0,5, H_1: p> 0,5. α= 0,05.

Sprendimas.

Pagal sąlygą, x= $60 > np = 100 \cdot 0{,}5 = 50$.

Tai $\hat{I}(H_1: H_2; O_n) = x\, log\frac{x}{np} + (n - x)log\frac{n-x}{nq} = 60 \cdot log\frac{60}{50} + (100 -$

$60)log\frac{100-60}{50}$ =*2,01355.*

$\chi^2_{1-2\alpha}(1) = \chi^2_{1-2x0,05}(1) = \chi^2_{0,9}(1) = 2{,}71$.

Kadangi $2\hat{I}(H_1: H_2; O_n) = 4{,}0271 > \chi^2_{1-2\alpha}(1) = 2{,}71$, tai duomenys neprieštarauja hipotezės H_1: p> 0,5 teisingumui.

Pastaba. Kadangi uždavinio sąlygoje yra binominio skirstinio parametras p= 0,5, tai galima buvo taikyti skyrelio „Vienpusės“ polinominės hipotezės formulės, kurios sutampa su čia pateiktomis imant c=2.

2 atv. Tikrinamos dvi hipotezės apie binominio skirstinio parametrą p: H_1: $p_1 > p_0$, H_2: p≤ p_0.

Analogiškai ankstesniajam tyrimui nulinę hipotezę H_2 atitinka binominio skirstinio diskretus tankis su parametru p: $f_2(x) = C_n^x p^x q^{n-x}$, kai $n > 1,\ q = 1 - p$. Hipotezei H_1 priklauso jungtinis binominis skirstinys $f^*(x) = C_n^x (p^*)^x (q^*)^{n-x}$ su parametrais $p^* = \frac{pe^\tau}{pe^\tau + q}$, $q^* = \frac{q}{pe^\tau + q}$, kai $T(x) = x,$ o $M_2(\tau) = (pe^\tau + q)^n$. Už jungtinio skirstinio parametrą θ imamas geriausias nepaslinktas p^* įvertis $\hat{p}^* = \frac{x}{n}$, t.y. $\hat{\theta} = n\hat{p}^* =$ x. Tada

$$\hat{\mathrm{I}}(\mathrm{p}^*: \mathrm{p}) = n\hat{\mathrm{p}}^* \log \frac{\hat{\mathrm{p}}^*}{p} + n\hat{q}^* \log \frac{\hat{q}^*}{q} = x \log \frac{\mathrm{x}}{np} + (n - x) \log \frac{\mathrm{n} - \mathrm{x}}{nq}.$$

Tiesiogiai tiriant galima įsitikinti, kad galiojant hipotezei H_2

$$inf_{p \le p_0} \hat{I}(p^*: p) = inf_{p \le p_0} \left(x \log \frac{x}{np} + (n - x) \log \frac{n-x}{nq} \right) = x \log \frac{x}{np_0} + (n - x) \log \frac{n-x}{nq_0},$$

kai $x > np_0$. Vadinasi, pagal statistinius duomenis

$$\hat{I}(H_1: H_2; O_n) = \begin{cases} x \log \frac{x}{np_0} + (n - x) \log \frac{n - x}{nq_0}, x > np_0 \\ 0\ , x \le np_0. \end{cases}$$

Kai teisinga hipotezė H_2, asimptotiškai

$$P\left\{2\hat{I}(H_1: H_2; O_n) > \chi^2_{1-2\alpha}(1)\right\} \le \alpha.$$

Vadinasi, hipotezių tikrinimo kritinė sritis, t. y. H_2 atmetimo sritis, nusakoma formule

$$W = \{ 2\hat{I}(H_1: H_2; O_n) > \chi^2_{1-2\alpha}(1) \},$$

kur $\chi^2_{1-2\alpha}(1)$ − yra Chi-kvadrat su vienu laisvės laipsniu 1-*2α* reikšmingumo lygmens kvantilis.

Pavyzdžiai.

1. n= 100, x= 55, H_2: p≤ 0,5, H_1: p> 0,5. α= 0,05, $p_0 = 0{,}5$.

   Sprendimas.

   Pagal sąlygą, x= $55 > np_0 = 100 \cdot 0{,}5 = 50$. Tai

$$\hat{I}(H_1: H_2; O_n) = x \log \frac{x}{np_0} + (n - x) \log \frac{n-x}{nq_0} = 55 \log \frac{55}{50} + (100 - 55) \log \frac{100-55}{50} = 0{,}5008.$$

$$\chi^2_{1-2\alpha}(1) = \chi^2_{1-2x0{,}05}(1) = \chi^2_{0{,}9}(1) = 2{,}71.$$

   Kadangi $2\hat{\mathrm{I}}(\mathrm{H}_1: \mathrm{H}_2; \mathrm{O}_\mathrm{n}) = 1{,}0017 < \chi^2_{1-2\alpha}(1) = 2{,}71$, tai duomenys neprieštarauja hipotezės H_2: p≤ 0,5 teisingumui.

2. n= 100, x= 65, H_2: p≤ 0,5, H_1: p> 0,5. α= 0,1, $p_0 = 0{,}5$.

   Sprendimas.

   Pagal sąlygą, *x*= $65 > np_0 = 100 \cdot 0{,}5 = 50$. *Tai*

$$\hat{I}(H_1: H_2; O_n) = x \log \frac{x}{np_0} + (n - x) \log \frac{n-x}{nq_0} = 65 \cdot \log \frac{65}{50} + (100 - 65) \log \frac{100-65}{50} = 4{,}5701.$$

   Kai α= 0,1 $\chi^2_{1-2\alpha}(1) = \chi^2_{1-2x0{,}1}(1) = \chi^2_{0{,}8}(1) = 1{,}64$.

   Kadangi $2\hat{I}(H_1: H_2; O_n) = 9{,}1402 > \chi^2_{1-2\alpha}(1) = 1{,}64$, tai duomenys neprieštarauja hipotezės H_1: p> 0,5 teisingumui.

3 atv. Tikrinamos dvi hipotezės apie binominio skirstinio parametrą p: H_1: $p_1 < p_0$, H_2: p≥ p_0.

Analogiškai gaunama

$$\hat{I}(H_1: H_2; O_n) = \begin{cases} x\, log\ \dfrac{x}{np_0} + (n-x)\, log\ \dfrac{n-x}{nq_0}, x < np_0 \\ 0\ , x \geq np_0. \end{cases}$$

Kai teisinga hipotezė H_2, asimptotiškai

$$P\{2\hat{I}(H_1: H_2; O_n)\ > \chi^2_{1-2\alpha}(1)\ \} \leq \alpha.$$

Vadinasi, hipotezių tikrinimo kritinė sritis, t. y. H_2 atmetimo sritis, nusakoma formule

W={ $2\hat{I}(H_1: H_2; O_n) > \chi^2_{1-2\alpha}(1)$ }. $\chi^2_{1-2\alpha}(1)$ − yra Chi-kvadrat su vienu laisvės laipsniu 1-*2α* reikšmingumo lygmens kvantilis.

Pavyzdžiai.

1. n= 100, x= 45, H_2: p≥ 0,5, H_1: p< 0,5. α= 0,05, $p_0 = 0{,}5$.

Sprendimas.

Pagal sąlygą, $x = 45 < np_0 = 100 \cdot 0{,}5 = 50$. Tai

$\hat{I}(H_1: H_2; O_n) = x\, log \frac{x}{np_0} + (n-x) log \frac{n-x}{nq_0} = 45 log \frac{55}{50} + (100-45) log \frac{100-45}{50} = 9{,}5310.$

$\chi^2_{1-2\alpha}(1) = \chi^2_{1-2x0{,}05}(1) = \chi^2_{0{,}9}(1) = 2{,}71.$

Kadangi $2\hat{I}(H_1: H_2; O_n) = 19{,}062 > \chi^2_{1-2\alpha}(1) = 2{,}71$, tai duomenys neprieštarauja hipotezės H_1: p< 0,5 teisingumui.

2. $n = 100$, $x = 35$, H_2: $p \geq 0{,}4$, H_1: $p < 0{,}4$. $\alpha = 0{,}1$, $p_0 = 0{,}4$.

Sprendimas.

Pagal sąlygą, $x = 35 < np_0 = 100 \cdot 0{,}4 = 40$. Tai

$\hat{I}(H_1: H_2; O_n) = x\, log \frac{x}{np_0} + (n-x) log \frac{n-x}{nq_0} = 35 log \frac{35}{40} + (100-35) log \frac{100-35}{60} = 12{,}3801.$

Kai $\alpha = 0{,}1$ $\chi^2_{1-2\alpha}(1) = \chi^2_{1-2x0{,}1}(1) = \chi^2_{0{,}8}(1) = 1{,}64.$

Kadangi $2\hat{I}(H_1: H_2; O_n) = 24{,}7602 > \chi^2_{1-2\alpha}(1) = 1{,}64$, tai duomenys neprieštarauja hipotezės H_1: p< 0,4 teisingumui.

4 atv. Tikrinamos dvi hipotezės apie binominio skirstinio parametrą p: H_1: $p_1 \neq p_0$, H_2: p= p_0.

Tai praktiškai 1 ir 3 atvejų apjungimas į vieną, kai pagal imties duomenis Kulbako-Leiblero informacija yra pavidalo

$$\hat{I}(H_1: H_2; O_n) = x\, log \frac{x}{np_0} + (n-x) log \frac{n-x}{nq_0}.$$

Kai teisinga hipotezė H_2, $2\hat{I}(H_1: H_2; O_n)$ asimptotiškai turi χ^2 − skirstinį su vienu laisvės laipsniu ir todėl

$$P\{2\hat{I}(H_1: H_2; O_n)\ > \chi^2_{1-\alpha}(1)\ \} \leq \alpha.$$

Vadinasi, hipotezių tikrinimo kritinė sritis, t. y. H_2 atmetimo sritis, nusakoma formule

$$W=\{2\hat{I}(H_1: H_2; O_n) > \chi^2_{1-\alpha}(1)\ \},$$

kur $\chi^2_{1-\alpha}(1)$ − yra Chi-kvadrat su vienu laisvės laipsniu 1-α reikšmingumo lygmens kvantilis.

1. $n= 100$, $x= 60$, $H_2: p= 0{,}5$, $H_1: p\neq 0{,}5$. $\alpha= 0{,}05$.

Sprendimas.

$\hat{I}(H_1: H_2; O_n) = x\, log\frac{x}{np} + (n-x) log\frac{n-x}{nq} = 60 log\frac{60}{50} + (100-60) log\frac{100-60}{50}$ =*2,01355.*

$\chi^2_{1-2\alpha}(1) = \chi^2_{1-0{,}05}(1) = \chi^2_{0{,}95}(1) = 3{,}84.$

Kadangi $2\hat{I}(H_1: H_2; O_n) = 4{,}0271 > \chi^2_{1-2\alpha}(1) = 3{,}84$ tai duomenys neprieštarauja hipotezės H_1: p$\neq 0{,}5$ teisingumui.

2. $n= 100$, $x= 8$, $H_2: p= 0{,}1$, $H_1: p\neq 0{,}1$. $\alpha= 0{,}1$.

Sprendimas.

Pagal sąlygą, $x= 10 \leq np = 100 \cdot 0{,}1 = 10$. Tai

$$\hat{I}(H_1: H_2; O_n) = x\, log\frac{x}{np} + (n-x) log\frac{n-x}{nq} = 8 log\frac{8}{10} + (100-8) log\frac{100-8}{90} = 0{,}236911$$

$\chi^2_{1-\alpha}(1) = \chi^2_{1-0{,}1}(1) = \chi^2_{0{,}9}(1) = 2{,}71.$

Kadangi $2\hat{I}(H_1: H_2; O_n) = 0{,}236911 < \chi^2_{1-\alpha}(1) = 2{,}71$, tai duomenys neprieštarauja hipotezės $H_2: p= 0{,}1$ teisingumui.

3. Polinominio skirstinio statistinės hipotezės

3.1 Paprastųjų hipotezių tikrinimas su Kulbako-Leiblero informacija

Tolimesniame darbe naudojami duomenys iš 1 ir 2 priedų.

Yra atsitiktinė imtis: $x_1, x_2, \cdots, x_c, x_1 + x_2 + \cdots + x_c = N$, su c kategorijų polinominiu skirstiniu ir tikrinama nulinė hipotezė

$H_2: (p) = (p_1, p_2, \cdots, p_c),\ p_1 + p_2 + \cdots + p_c = 1,\ H_1$: imtis yra iš bet kurios galimos c kategorijų polinominiu skirstiniu.

Naudojama Kulbako-Leiblero informacija:

$$\hat{I}(H_1: H_2; O_N) = x_1\, log\, \frac{x_1}{Np_1} + \cdots + x_c \frac{x_c}{Np_c}.$$

Jei nulinė hipotezė H_2 teisinga, tai $2\hat{I}(*, 2; O_N)$ turi asimptotiškai χ^2(c-1) skirstinį. Todėl

$$P\{2\hat{I}(H_1: H_2; O_N)\ > \chi^2_{1-\alpha}(c-1)\ \} \leq \alpha.$$

Vadinasi, hipotezių tikrinimo kritinė sritis, t. y. H_2 atmetimo sritis, nusakoma formule

$$W = \{\, 2\hat{I}(H_1: H_2; O_N) > \chi^2_{1-\alpha}(c-1)\ \},$$

kur $\chi^2_{1-\alpha}(c-1)$ − yra Chi-kvadrat su $c-1$ laisvės laipsniu $1-\alpha$ reikšmingumo lygmens kvantilis.

Pavyzdžiai.

1) Patikrinama, ar bedarbystė statistiškai reikšmingai pakito 2016–2017 m. Tuo tikslu nulinei hipotezei imami bedarbių 2016 m. statistiniai dažniai (žr. priedą nr. 2): $H_2: p_1 = 0{,}2402,\ p_2 = 0{,}0697,\ p_3 = 0{,}1739,\ p_4 = 0{,}0973,\ p_5 = 0{,}0607, p_6 = 0{,}0897,\ p_7 = 0{,}1014,\ p_8 = 0{,}0455,\ p_9 = 0{,}0531,\ p_{10} = 0{,}0685.\ \alpha = 0,01.$

Iš 2017 m. imama $N = N(2017) = 139600$ ir x_i: (34600; 9000; 26100; 13100; 7900; 12300; 14000; 6000; 7100; 9500) ir įrašoma į formulę:

$$\hat{I}(H_1 : H_2; O_N) = x_1 \, log \frac{x_1}{Np_1} + \cdots + x_c \frac{x_c}{Np_c}$$

$$= 34600 \, log \frac{34600}{139600 \cdot 0{,}2402}$$

$$+ 9000 \, log \frac{9000}{139600 \cdot 0{,}0697}$$

$$+ 26100 \, log \frac{26100}{139600 \cdot 0{,}1739} + 13100 \, log \frac{13100}{139600 \cdot 0{,}0973}$$

$$+ 7900 \, log \frac{7900}{139600 \cdot 0{,}0607}$$

$$+ 12300 \, log \frac{12300}{139600 \cdot 0{,}0897} + 14000 \, log \frac{14000}{139600 \cdot 0{,}1014}$$

$$+ 6000 \, log \frac{6000}{139600 \cdot 0{,}0455} + 7100 \, log \frac{7100}{139600 \cdot 0{,}0531}$$

$$+ 9500 \, log \frac{9500}{139600 \cdot 0{,}0685} = 81{,}5544$$

Į formulę $\chi^2_{1-\alpha}(\mathrm{k}-1)$ įrašius $\mathrm{k} = 10$ ir $\alpha = 0,01$ gaunama $\chi^2_{0,99}(9) = 21{,}7$. Kadangi $2\hat{I}(\mathrm{H}_1 : \mathrm{H}_2; \mathrm{O}_\mathrm{n}) = 163{,}1088 > \chi^2_{1-\alpha}(9) = 21{,}7$, tai duomenys prieštarauja hipotezės H_2 teisingumui. Galima teigti, kad per metus įvyko esminis registruotų bedarbių skaičiaus pasikeitimas Lietuvos apskrityse.

2)Tikrinama, ar bedarbystė statistiškai reikšmingai pakito 2015–2016 m.

$\mathrm{H}_2 : \mathrm{p}_1 = 0{,}2364,\ \mathrm{p}_2 = 0{,}0702,\ \mathrm{p}_3 = 0{,}1719,\ \mathrm{p}_4 = 0{,}1018,\ \mathrm{p}_5 = 0{,}0613, \mathrm{p}_6 = 0{,}0929,\ \mathrm{p}_7 = 0{,}0986,\ \mathrm{p}_8 = 0{,}0461,\ \mathrm{p}_9 = 0{,}0544,\ \mathrm{p}_{10} = 0{,}0664.\ \alpha = 0,01.$

Iš 2016 m. imama $\mathrm{N} = \mathrm{N}(2016) = 144900$ ir $\mathrm{x_i}$: (34800; 10100; 25200; 14100; 8800; 13000; 14700; 6600; 7700; 9900) ir įrašoma į formulę:

$$\hat{I}(H_1: H_2; O_N) = x_1\, log\frac{x_1}{Np_1} + \cdots + x_c\frac{x_c}{Np_c}$$
$$= 34800\, log\frac{34800}{144900 \cdot 0{,}2364}$$
$$+ 10100\, log\frac{10100}{144900 \cdot 0{,}0702}$$
$$+ 25200\, log\frac{25200}{144900 \cdot 0{,}1719} + 14100\, log\frac{14100}{144900 \cdot 0{,}1018}$$
$$+ 8800\, log\frac{8800}{144900 \cdot 0{,}0613}$$
$$+ 13000\, log\frac{13000}{144900 \cdot 0{,}0929} + 14700\, log\frac{14700}{144900 \cdot 0{,}0986}$$
$$+ 6600\, log\frac{6600}{144900 \cdot 0{,}0461} + 7700\, log\frac{7700}{144900 \cdot 0{,}0544}$$
$$+ 9900\, log\frac{9900}{144900 \cdot 0{,}0664} = 18{,}1292$$

Į formulę $\chi^2_{1-\alpha}(\mathrm{k}-1)$ įrašius $\mathrm{k} = 10$ ir $\alpha = 0,01$ gaunama $\chi^2_{0,99}(9) = 21{,}7$. Kadangi $2\hat{I}(H_1: H_2; O_n) = 36{,}2584 > \chi^2_{1-\alpha}(9) = 21{,}7$, tai duomenys prieštarauja hipotezės H_2 teisingumui, t.y., 2015-2016 metais įvyko esminis registruotų bedarbių skaičiaus pokytis.

Kadangi skaičiai 36.2584 ir 21.7 yra pakankamai artimi, yra tikslinga paskaičiuoti hipotezę su mažesniu reikšmingumo lygmeniu α. Imama $\alpha = 0,001$:

Į formulę $\chi^2_{1-\alpha}(\mathrm{k}-1)$ įrašius $\mathrm{k} = 10$ ir $\alpha = 0,001$ gaunama $\chi^2_{0,999}(9) = 27{,}9$. Kadangi $2\hat{\mathrm{I}}(\mathrm{H}_1: \mathrm{H}_2; \mathrm{O}_\mathrm{n}) = 36{,}2584 > \chi^2_{1-\alpha}(9) = 27{,}9$, tai duomenys vis tiek prieštarauja hipotezės H_2 teisingumui, t.y., 2015-2016 metais įvyko esminis registruotų bedarbių skaičiaus pokytis.

3.2 Sudėtingųjų hipotezių tikrinimas su Kulbako-Leiblero informacija

Yra atsitiktinė imtis: $x_1, x_2, \cdots, x_c, x_1 + x_2 + \cdots + x_c = N$, su c kategorijų polinominiu skirstiniu. Tikrinamos hipotezės:

$$H_1: p_1 > \frac{1}{c},\ p_1 + p_2 + \cdots + p_c = 1,\ H_2: p_1 = p_2 = \cdots = p_c = \frac{1}{c}.$$

Hipotezėms tikrinti sudaromos statistikos – Kulbako-Leiblero informacijos:

$$\hat{I}(H_1: H_2; Q_N) = \sum_{i=1}^{c} x_i \log\frac{cx_i}{N}, kai\ \ x_1 > \frac{N}{c};$$

$$\hat{I}(H_1: H_2; Q_N) = \sum_{i=2}^{c} x_i \log\frac{(c-1)x_i}{N - x_1}, kai\quad x_1 \le \frac{N}{c}.$$

Hipotezių kritinė sritis apibrėžiama taip:

$$W = \{2\hat{I}(H_1: H_2; Q_N) > \chi^2_{1-2\alpha}(c-1)\},$$

kur $\chi^2_{1-\alpha}(c-1)$ – yra Chi-kvadrat su vienu laisvės laipsniu 1-α reikšmingumo lygmens kvantilis.

Pavyzdžiai.

1). Turime 2004 m. vyrų padarytų nusikaltimų duomenis:

Mėnuo	1	2	3	4	5	6	7	8	9	10	11	12
Nusikalt. skaičius.; x_i	275	306	364	321	347	334	334	348	336	302	343	411=x_1

$N = 4021,\ c = 12,\ \alpha = 0{,}05$. Tikrinamos statistinės hipotezės:

$$H_1: p_1 > \frac{1}{c},\ p_1 + p_2 + \cdots + p_c = 1,\ H_2: p_1 = p_2 = \cdots = p_c = \frac{1}{c}.$$

Sprendimas. $c = 12, N = 4021$. Todėl

$\frac{N}{c} = \frac{4021}{12} = 335{,}08$. Kadangi $411 = x_1 > \frac{N}{c} = 335{,}08,$ tai imama

$$\hat{I}(H_1: H_2; Q_N) = \sum_{i=2}^{c} x_i \log \frac{cx_i}{N} =$$

$$= 2(x_1 \log \frac{cx_1}{N} + \cdots + x_c \log \frac{cx_c}{N}) = 2(411 \log \frac{12 \cdot 411}{4021} + \cdots + 275 \log \frac{12 \cdot 275}{4021}) = 16{,}0345;$$

$\chi^2_{1-2\alpha}(11) = \chi^2_{0,90}(11) = 17{,}3.$

Vadinasi, $W = \{2\hat{I}(H_1: H_2; Q_N) > \chi^2_{1-2\alpha}(c-1)\} = \{16{,}0345 > 17{,}3\}$ neteisinga, todėl nulinė hipotezė H_2 priimama.

Kadangi 16 ir 17 artimi, tai naudinga imti dar kitus reikšmingumo lygmenis:

$$\alpha = 0{,}1;\ \chi^2_{1-2\alpha}(11) = \chi^2_{0,80}(11) = 14{,}6.$$

Dabar $W = \{2\hat{I}(H_1: H_2; Q_N) > \chi^2_{1-2\alpha}(c-1)\} = \{16{,}0345 > 14{,}6\}$ teisinga, todėl nulinė hipotezė H_2 atmetama.

$\alpha = 0{,}005;\ \chi^2_{1-2\alpha}(11) = \chi^2_{0,99}(11) = 24{,}7.$

Dabar $W = \{2\hat{I}(H_1: H_2; Q_N) > \chi^2_{1-2\alpha}(c-1)\} = \{16{,}0345 > 24{,}7\}$ teisinga, todėl nulinė hipotezė H_2 priimama.

2). Dabar tiems patiems duomenims tikrinamos šios hipotezės:

$$H_1:\ p_1 = p > \frac{1}{c}, p_2 = p_3 = \cdots = p_c = \frac{1-p}{c-1},$$

$$H_2: p_1 = p_2 = \cdots = p_c = \frac{1}{c}$$

Sprendimas. $c = 12, N = 4021$. Todėl

$\frac{N}{c} = \frac{4021}{12} = 335{,}08$. Kadangi $411 = x_1 > \frac{N}{c} = 335{,}08,$ tai imama

$$\hat{I}(H_1: H_2; Q_N) = x_1\, log\frac{cx_1}{N} + (N - x_1)\, log\frac{(N - x_1)c}{N(c-1)}, x_1 > \frac{N}{c}$$

Hipotezių kritinė sritis apibrėžiama taip:

$$W = \{2\hat{I}(H_1: H_2; Q_N) > \chi^2_{1-2\boldsymbol{\alpha}}(1)\}.$$

$$2\hat{I}(H_1: H_2; Q_N) = 2(x_1 \log\frac{cx_1}{N} + (N - x_1)\log\frac{(N-x_1)c}{N(c-1)}) = 2(411\log\frac{12\cdot 411}{4021} + (4021 -$$

$$411)\log\frac{(4021-411)\cdot 12}{4021(12-1)}) = 7{,}646.$$

$\alpha = 0{,}05$

$\chi^2_{1-2\alpha}(1) = \chi^2_{0{,}90}(1) = 2{,}71$. Vadinasi,

$W = \{2\hat{I}(H_1: H_2; Q_N) > \chi^2_{1-2\boldsymbol{\alpha}}(c-1)\}$={$7{,}646 > 2{,}71$} teisinga, todėl nulinė hipotezė H_2 atmetama.

$\alpha = 0{,}1;\ \chi^2_{1-2\alpha}(1) = \chi^2_{0{,}80}(1) = 1{,}64$

Dabar $W = \{2\hat{I}(H_1: H_2; Q_N) > \chi^2_{1-2\boldsymbol{\alpha}}(c-1)\}$={$7{,}646 > 1{,}64$} teisinga, todėl nulinė hipotezė H_2 atmetama.

$\alpha = 0{,}005;\ \chi^2_{1-2\alpha}(1) = \chi^2_{0{,}99}(1) = 6{,}63.$

Dabar $W = \{2\hat{I}(H_1: H_2; Q_N) > \chi^2_{1-2\boldsymbol{\alpha}}(c-1)\}$={$7{,}646 > 6{,}63$} teisinga, todėl nulinė hipotezė H_2 atmetama.

3). Trečiajam pavyzdžiui reikės dviejų kintamųjų, kurie pažymėti lentelėje:

Mėnuo	1	2	3	4	5	6	7	8	9	10	11	12
Nusikalt. skaičius.; x_i	275	306	364	321	347	334	334	348	336	302	343=x_1	411=x_2

$N = 4021,\ c = 12,\ \alpha = 0{,}05;\ \alpha = 0{,}1;\ \alpha = 0{,}005$. Tikrinamos statistinės hipotezės:

$$H_1: p_1, p_2, \cdots, p_c, p_1 + p_2 + \cdots + p_c = 1,$$

$$H_2: p_1 = p, \qquad p_2 = \cdots = p_c = \frac{1-p}{c-1}.$$

Hipotezių kritinė sritis apibrėžiama taip:

$$W = \{2\hat{I}(H_1: H_2; Q_N) > \chi^2_{1-2\boldsymbol{\alpha}}(c-2)\}.$$

Sprendimas. $c = 12, N = 4021$. Todėl

$\frac{N}{c} = \frac{4021}{12} = 335{,}08$. Kadangi 343=$x_1 > \frac{N}{c} = 335{,}08$, tai imama

$$2\hat{I}(H_1: H_2; Q_N) = 2(x_2 \log\frac{(c-1)x_2}{N-x_1} + \cdots + x_c \log\frac{(c-1)x_c}{N-x_1}) = 2(411\log\frac{(12-1)\cdot 411}{4021-343} + \cdots +$$

$$302\log\frac{(12-1)\cdot 302}{4021-343}) = 16{,}253$$

$\alpha = 0{,}05$

$\chi^2_{1-2\alpha}(10) = \chi^2_{0{,}90}(10) = 16$. Vadinasi,

$W = \{2\hat{I}(H_1: H_2; Q_N) > \chi^2_{1-\boldsymbol{2\alpha}}(c-1)\}$=$\{16{,}253 > 16\}$ teisinga, todėl nulinė hipotezė H_2 atmetama.

$\alpha = 0{,}1;\ \chi^2_{1-2\alpha}(10) = \chi^2_{0{,}80}(10) = 13{,}4.$

Dabar $W = \{2\hat{I}(H_1: H_2; Q_N) > \chi^2_{1-\boldsymbol{2\alpha}}(c-1)\}$=$\{16{,}253 > 13{,}4\}$ teisinga, todėl nulinė hipotezė H_2 atmetama.

$\alpha = 0{,}005;\ \chi^2_{1-2\alpha}(10) = \chi^2_{0{,}99}(10) = 23{,}2.$

Dabar $W = \{2\hat{I}(H_1: H_2; Q_N) > \chi^2_{1-\boldsymbol{2\alpha}}(c-1)\}$=$\{16{,}253 > 23{,}2\}$ neteisinga, todėl nulinė hipotezė H_2 priimama.

IŠVADOS

1. Nustatyta, kaip taikoma Kulbako-Leiblero informacija dviem paprastosioms binominio skirstinio statistinėms hipotezėms tikrinti. Su konkrečiais duomenimis iliustruojami gauti rezultatai.
2. Ištirta, kaip taikant jungtinę Kulbako-Leiblero informaciją yra tikrinamos "vienpusės" ir "dvipusės" sudėtingosios binominio skirstinio statistinės hipotezės, gautas procedūras iliustruojant pavyzdžiais su konkrečiais duomenimis.
3. Parodyta, kaip taikant Kulbako-Leiblero informaciją yra tikrinamos paprastosios ir sudėtingosios polinominio skirstinio statistinės hipotezės, gautas procedūras iliustruojant pavyzdžiais su konkrečiais duomenimis.

KULBAKO-LEIBLERO INFORMACIJOS TAIKYMAS TIKRINANT STATISTINES HIPOTEZES

SANTRAUKA

Informacijos teorija yra matematikos teorijos šaka, apimanti tikimybių ir matematinės statistikos kryptis. Šio darbo tikslas yra ištirti, kaip Kulbako-Leiblero informacija susieta su statistinių hipotezių tikrinimu. Tikslui pasiekti buvo keliami keli uždaviniai. Pirmiausia – įsisavinti informacinę teoriją, kur taikoma Kulbako-Leiblero informacija statistinėms hipotezėms tikrinti. Tai atlikus buvo nustatyti polinominio skirstinio Kulbako-Leiblero informacijos taikymo procedūriniai žingsniai statistinėms hipotezėms tikrinti ir parodyta, kaip taikant Kulbako-Leiblero informaciją yra tikrinamos paprastosios ir sudėtingosios polinominio skirstinio statistinės hipotezės, gautas procedūras iliustruojant pavyzdžiais su konkrečiais duomenimis.

Kadangi binominis ir polinominis skirstiniai yra susiję, buvo nuspręsta parodyti, kaip pritaikyti polinominiam skirstiniui skirtus Kulbako-Leiblero informacijos procedūrinius žingsnius binominio skirstinio paprastosioms ir sudėtingosioms statistinėms hipotezėmis tikrinti. Gauti rezultatai iliustruoti pavyzdžiais su konkrečiais duomenimis.

THE APPLICATION OF KULBACK-LEIBLER INFORMATION FOR THE TESTING OF STATISTICAL HYPOTHESES

SUMMARY

Information theory is a branch of mathematical theory that encompasses the directions of probability and mathematical statistics. The aim of this work is to investigate how Kullbak-Leibler information is related to the testing of statistical hypotheses. Several goals were set to achieve the goal. The first goal - to learn about the information theory, where Kullbak-Leibler information is applied in the test of statistical hypotheses. After doing so, identify the procedural steps for applying the polynomial distribution of Kullbak-Leibler information to test statistical hypotheses and show how the application of Kullbak-Leibler information tests simple and complex polynomial distribution statistical hypotheses by illustrating the procedures with real data.

Since the binomial and polynomial distributions are related, it was decided to show how to apply the procedural steps of Kullbak-Leibler information for a polynomial distribution to test simple and complex statistical hypotheses of a binomial distribution. The obtained results are illustrated with specific data.

LITERATŪRA

1. F. Liese and I. Vajda, Convex Statistical Distances, Teubner, Leipzig, (1987).
2. F. Liese and I. Vajda, f-divergences: sufficiency, deficiency and testing of hypotheses, Advances in Inequalities from Probability Theory and Statistics (Neil S. Barnett and Sever S. Dragomir, ed.), Nova Publishers, Toronto, (2009), pp. 131-173.
3. H. Chernoff, A Measure of Asymptotic Efficiency for Tests of a Hypothesis Based on a Sum of Observations, Ann. Math. Stat. 23 (1952), 493-507.
4. I. Vajda, Generalization of discrimination-rate theorems of Chernoff and Stein, Kybernetika 26(1990), no. 4, 273-288.
5. I. Vajda, Distances and discrimination rates for stochastic processes, Stochastic Process. Appl. 35(1990), 47-57.
6. Yu. N. Lin'kov, Asymptotical Methods of Statistics for Stochastic Processes, Naukova Dumka, Kiev, (1993).
7. Yu. N. Lin'kov, Lectures in mathematical statistics. 2., Istoki, Donetsk, (2001).
8. J. Kruopis, Matematinė statistika. Vilnius, 1993.
9. Kullback S. Information theory and ststistics. USA (1978). http://index-of.co.uk/Information-Theory/Information%20theory%20and%20statistics%20-%20Solomon%20Kullback.pdf
10. V. Kanišauskas and L. Dronova-Platbarzdė, Asymptotical Separation of Two Simple Hypotesess Using the Most Powerful Criteria, Journal of Young Scientists 44 (2015), no. 2, 105-109.
11. V. Kanišauskas, Tikimybių teorija ir matematinės statistikos pagrindai. ŠU, Šiauliai (2000).
12. V. Kanišauskas, Asymptotically minimax separation two simple hypotheses, Lith. Math. J. 38 (1998), no. 2, 169-184.
13. V. Kanišauskas, Asymptotically minimax testing of simple hypotheses, Lith. Math. J. 40 (2000), no. 3, 313-320.
14. W. Stummer and I. Vajda, On Bregman Distances and Divergences of Probability Measures, IEEE Trans. Inform. Theory 58 (2012), no. 3, 1277-1288.
15. В. Е. Гмурман, Руководство к решению задач по теории вероятностей и математической статистике, Учеб. Пособие для втузов, Высш. Школа, Москва (1970).

PRIEDAI

PRIEDAS NR. 1.

Darbo biržoje registruoti bedarbiai (tūkst.) 2005-2017m.

Apskritys/metai	2005	2006	2007	2008	2009	2010	2011	2012
Lietuvos Respublika	100,8	73,2	67,3	73,4	203,1	312,1	247,2	216,9
Vilniaus	20,9	16,1	15,6	17,7	52,3	81,9	62,2	52,9
Alytaus	7,6	5,2	4,4	5,3	12,9	18,6	15,4	14
Kauno	15	11,5	11,1	12,4	35,3	55,9	45,5	39,7
Klaipėdos	10,7	7	7,3	7,7	21,7	34,1	26	21,4
Marijampolės	5,8	3,7	2,8	3,1	9,3	14,9	12,6	11,9
Panevėžio	11	8,4	7,4	7,2	19,8	28,4	22,4	19,8
Šiaulių	11,1	8,1	7	7,7	21	30,5	23,9	21,3
Tauragės	5,5	3,8	3,2	3,3	7,6	12	9,5	9,1
Telšių	6,9	4,5	3,9	4,6	12,6	19	14,8	13,2
Utenos	6,3	5	4,3	4,4	10,6	16,8	14,9	13,6

Apskritys/metai	2013	2014	2015	2016	2017
Lietuvos Respublika	201,3	173	158,2	144,9	139,6
Vilniaus	47,9	40,5	37,4	34,8	34,6
Alytaus	13,9	12,5	11,1	10,1	9
Kauno	36,3	30,6	27,2	25,2	26,1
Klaipėdos	19,4	16,6	16,1	14,1	13,1
Marijampolės	11,7	10,4	9,7	8,8	7,9
Panevėžio	17,9	16,3	14,7	13	12,3
Šiaulių	20,4	17	15,6	14,7	14
Tauragės	8,8	8	7,3	6,6	6
Telšių	12,1	10	8,6	7,7	7,1
Utenos	12,9	11,1	10,5	9,9	9,5

PRIEDAS NR. 2.

Darbo biržoje registruoti bedarbiai (proc.) 2005-2017m.

	2005	2006	2007	2008	2009
Lietuvos Respublika	100,8	73,2	67,3	73,4	203,1
Vilniaus apskritis	20,73%	21,99%	23,18%	24,11%	25,75%
Alytaus apskritis	7,54%	7,10%	6,54%	7,22%	6,35%
Kauno apskritis	14,88%	15,71%	16,49%	16,89%	17,38%
Klaipėdos apskritis	10,62%	9,56%	10,85%	10,49%	10,68%
Marijampolės apskritis	5,75%	5,05%	4,16%	4,22%	4,58%
Panevėžio apskritis	10,91%	11,48%	11,00%	9,81%	9,75%
Šiaulių apskritis	11,01%	11,07%	10,40%	10,49%	10,34%
Tauragės apskritis	5,46%	5,19%	4,75%	4,50%	3,74%
Telšių apskritis	6,85%	6,15%	5,79%	6,27%	6,20%
Utenos apskritis	6,25%	6,83%	6,39%	5,99%	5,22%

	2010	2011	2012	2013
Lietuvos Respublika	312,1	247,2	216,9	201,3
Vilniaus apskritis	26,24%	25,16%	24,39%	23,80%
Alytaus apskritis	5,96%	6,23%	6,45%	6,91%
Kauno apskritis	17,91%	18,41%	18,30%	18,03%
Klaipėdos apskritis	10,93%	10,52%	9,87%	9,64%
Marijampolės apskritis	4,77%	5,10%	5,49%	5,81%
Panevėžio apskritis	9,10%	9,06%	9,13%	8,89%
Šiaulių apskritis	9,77%	9,67%	9,82%	10,13%
Tauragės apskritis	3,84%	3,84%	4,20%	4,37%
Telšių apskritis	6,09%	5,99%	6,09%	6,01%
Utenos apskritis	5,38%	6,03%	6,27%	6,41%

	2014	2015	2016	2017
Lietuvos Respublika	173	158,2	144,9	139,6
Vilniaus apskritis	23,41%	23,64%	24,02%	24,79%
Alytaus apskritis	7,23%	7,02%	6,97%	6,45%
Kauno apskritis	17,69%	17,19%	17,39%	18,70%
Klaipėdos apskritis	9,60%	10,18%	9,73%	9,38%
Marijampolės apskritis	6,01%	6,13%	6,07%	5,66%
Panevėžio apskritis	9,42%	9,29%	8,97%	8,81%
Šiaulių apskritis	9,83%	9,86%	10,14%	10,03%
Tauragės apskritis	4,62%	4,61%	4,55%	4,30%
Telšių apskritis	5,78%	5,44%	5,31%	5,09%
Utenos apskritis	6,42%	6,64%	6,83%	6,81%