VILNIAUS UNIVERSITETAS
ŠIAULIŲ AKADEMIJA

MATEMATIKOS MAGISTRANTŪROS STUDIJŲ PROGRAMA
Didžiųjų duomenų analitikos specializacija

SIMONA VELIČKAITĖ

Pagrindinių studijų baigiamasis darbas

Skirtingų prozos kūrėjų identiteto nustatymas lingvostatistiniais metodais

Darbo vadovas (-ė): doc. dr. Karolina Kanišauskienė

Šiauliai, 2021

Studijuojančiojo, teikiančio baigiamąjį darbą, GARANTIJA

WARRANTY of Final Thesis

Vardas, pavardė *Name, Surname*	**Simona Veličkaitė**
Padalinys *Faculty*	**Šiaulių akademija** ***Šiauliai Academy***
Studijų programa *Study Programme*	**Matematika** ***Mathematics***
Darbo pavadinimas *Thesis topic*	**Skirtingų prozos kūrėjų identiteto nustatymas lingvostatistiniais metodais** ***Identification of different initial creators by linguistic statistical methods***
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, Simona Veličkaitė, pateikdamas (-a) šį darbą, patvirtinu (*pažymėti*) ☒
I, Simona Veličkaitė, by submitting this paper confirm (check)

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

Temos aktualumas. Knygos – vienas geriausių intelektualaus laisvalaikio praleidimo būdų. Jos skatina tobulėti, atrasti ir išmokti ką nors naujo, suteikia peno apmąstymams bei duoda atsakymus į taip rūpimus klausimus. Skaitymo nauda iš tiesų didžiulė, labai gaila, kad tai supranta ne visi. Jaunimas dabar mieliau renkasi žiūrėti filmus, knygų ekranizacijas, nors iš tiesų filmas retai perteikia tai, ką autorius norėjo pasakyti. Knygų skaitymas tikrai praturtina žmones ir daro juos intelektuališkesnius [10]. Dažniausiai žmonės renkasi prozos kūrinius, neeiliuotą grožinę literatūrą [16].

Identitetas – tapatybė, žmogaus arba daikto savybių visuma, pagal kurią jis yra atpažįstamas, atskiriamas nuo kitų [7]. Šiais laikais dauguma žmonių yra labai supanašėję, laikosi tam tikrų visuomenės taisyklių, bijo išsiskirti ir būti kitokiais nei visi. Bijodami pasmerkimo, kartais net pašaipų, daugelis kopijuoja aprangos stilių, vaikosi madų kaip leisti laisvalaikį, ką valgyti ir kaip gyventi. „Mūsų unikalumas yra tai, ką tik mes asmeniškai galime atsinešti į šį pasaulį. Tai, ką mes galime ir turime parodyti ir pasidalinti su kitais“ [15]. Todėl neturime bijoti būti išskirtiniais, nes tai daro mūsų asmenybes įdomias. Kaip ir žmonės kasdienybėje, taip ir literatūros kūrinių autoriai turi savitų bruožų, išskirtinumų. Šiame darbe, tiriant autorių rašyseną, skaičiuojant žodžius ir raides, siekiama nustatyti autorių panašumus ir skirtumus, unikalius ypatumus.

Šiame darbe matematinės statistikos metodais bus ištirtos penkių skirtingų autorių knygos. Autoriai lyginami tarpusavyje skaičiuojant žodžius ir raides.

Tyrimo objektas – penkios skirtingų autorių knygos.

Tyrimo tikslas – nustatyti skirtingų prozos kūrėjų identitetą.

Tyrimo uždaviniai:

1. Išsiaiškinti, kokie nagrinėjamų kūrinių lingvistiniai požymiai tinka statistinei analizei.
2. Ištirti, kokie matematinės statistikos metodai tinka užsibrėžtam tikslui pasiekti.
3. Taikant pasirinktus matematinės statistikos metodus išsiaiškinti, kuo autoriai skiriasi vienas nuo kito.

Darbo struktūra. Magistro darbą sudaro įvadas, trys skyriai, išvados, literatūros sąrašas, santrauka bei priedai. Teorinėje dalyje trumpai aprašytos statistikos sąvokos, metodai ir modeliai, kurie reikalingi praktinėje dalyje. Praktinėje dalyje atliekami skaičiavimai ir pateikiami rezultatai.

1. TEORINĖ DALIS

1.1. Bazinės matematinės statistikos sąvokos

Šiame poskyryje pateikiamos bazinės matematinės statistikos sąvokos, formulės, apibrėžimai. Jam parengti naudotasi [3], [8] literatūros šaltiniais.

Populiacija – visų objektų, kurių požymiai tiriami, aibė. Imtis – populiacijos dalis, naudojama statistiniam tyrimui.

Variacinė eilutė – išdėstyta nemažėjimo tvarka kiekybinio kintamojo duomenų eilutė:

$$x_{(1)} \leq x_{(2)} \leq x_{(3)} \leq \cdots \leq x_{(n)},$$

$x_{min} = x_{(1)}$, $x_{max} = x_{(n)}$.

Vidurkis – tai taškas, kuris vidutiniškai artimiausias visiems imties elementams. Žymimas $\bar{x}$ ir apskaičiuojamas pagal šią formulę:

$$\bar{x} = \frac{1}{n}\sum_{j=1}^{n} x_j.$$

Moda – tai dažniausiai duomenų aibėje pasikartojusi reikšmė. Žymima *Mo*.

Mediana – tai skaičius, padalijantis variacinę eilutę į dvi maždaug lygias dalis. Žymima *Md* ir apskaičiuojama pagal formulę:

$$Md = \begin{cases} x_{\left(\frac{n+1}{2}\right)}, & \text{kai } n - \text{nelyginis}, \\ \dfrac{x_{\left(\frac{n}{2}\right)} + x_{\left(\frac{n}{2}+1\right)}}{2}, & \text{kai } n - \text{lyginis}. \end{cases}$$

Imties dispersija parodo duomenų sklaidą apie vidurkį. Žymima s^2 ir apskaičiuojama pagal formulę:

$$s^2 = \frac{1}{n-1}\sum_{j=1}^{n}(x_j - \bar{x})^2.$$

Standartinis nuokrypis (žymima s) yra dažniausiai taikomas sklaidos matas. Kadangi dispersija matuojama kvadratiniais vienetais, tai ši charakteristika

$$s = \sqrt{s^2}$$

leidžia lengviau interpretuoti gautas reikšmes.

1.2. Sisteminis ėmimas

Poskyriui parengti naudotasi [11] literatūros šaltiniu.

Sisteminis ėmimas – tai toks ėmimas, kai iš kuria nors tvarka sudaryto populiacijos elementų sąrašo, atsitiktinai pasirinkus pradžią, kiekvienas q-asis elementas yra išrenkamas į imtį. Sudarant sisteminę imtį pirmasis imties elementas išrenkamas atsitiktinai iš pirmųjų q sąrašo elementų, o likusieji imties elementai vienareikšmiškai nustatomi vis pridedant po q prie pirmuoju išrinkto elemento numerio iki pat sąrašo pabaigos.

Kartais, jei žinoma populiacijos struktūra, sisteminis ėmimas gali būti net efektyvesnis už paprastąjį atsitiktinį ėmimą, kai bet kuris n skirtingų elementų rinkinys iš N dydžio baigtinės populiacijos turi vienodą tikimybę būti išrinktas.

1.2.1. Sisteminės imties išrinkimo būdai

Sisteminės imties išrinkimo būdų (išrinkimo schemų) yra daug. Skirtingi būdai gali būti taikomi, atsižvelgiant į tai, ar žinomas populiacijos dydis, ar ne. Panagrinėsime atvejį, kai

Populiacijos dydis N yra žinomas.

Tarkime, kad populiacijos dydis N dalus pasirinktam imties dydžiui n, t. y. $N = qn$, čia q – sveikasis skaičius. n dydžio sisteminei imčiai išrinkti nustatomas ėmimo žingsnis $q = N/n$. Iš skaičių rinkinio $\{1,2, \dots, q\}$ su vienodomis tikimybėmis, lygiomis $\frac{1}{q}$, atsitiktinai išrenkamas vienas skaičius. Sakykime, kad tai bus skaičius q_0. Į imtį imami elementai su numeriais

$$q_0, q_0 + q, q_0 + 2q, \dots, q_0 + (n-1)q,$$

taip paimama po vieną elementą iš kiekvieno q ilgio ėmimo intervalo $[1 + (k-1)q, 1 + kq)$.

Jei $N \neq qn$, tai populiacijos dydį galima padaryti dalų imties dydžiui, pašalinant iš populiacijos atitinkamą skaičių atsitiktinai išrinktų elementų arba pridedant keletą elementų, kurių kintamojo reikšmės lygios nuliui, arba pridedant reikiamą skaičių atsitiktinai išrinktų populiacijos elementų. Taip pat galima pasirinkti trupmeninį ėmimo žingsnį $q = \frac{N}{n}$ arba, sujungus ėmimo sąrašo galą su pradžia, rinkti sisteminę imtį žiedu, trūkstamus imties elementus imant iš sąrašo pradžios.

1.3. Skirstiniai

Poskyriui parengti naudotasi [3], [8], [13] literatūros šaltiniais.

Atsitiktinės imties funkcija $f(X_1, X_2, \dots, X_n)$ vadinama statistika.

Kadangi statistika yra atsitiktinių dydžių funkcija, todėl ji taip pat yra atsitiktinis dydis (vienmatis arba daugiamatis). Taigi galima kalbėti apie statistikos, arba vadinamąjį statistikos imties skirstinį.

Sakoma, kad atsitiktinis dydis X turi **normalųjį**, arba **Gauso**, skirstinį su parametrais $\theta = (a, \sigma), -\infty < a < \infty, \sigma > 0$, jei jis turi tankį

$$p(\theta, x) = \frac{1}{\sigma\sqrt{2\pi}} exp\left\{-\frac{(x-a)^2}{2\sigma^2}\right\}, x \in R.$$

Šiuo atveju žymima $X \sim N(a, \sigma^2)$. Normalusis skirstinys su parametrais $(0, 1)$ (žym. $X \sim N(0,1)$) ir tankiu

$$p(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}, \qquad x \in R,$$

vadinamas standartiniu.

Tegu $X_1, X_2, \dots, X_n$ yra nepriklausomi atsitiktiniai dydžiai, kuriems $X_i \sim N(0,1)$, $i = 1, \dots, n$. Tada atsitiktinio dydžio

$$\chi_n^2 = X_1^2 + X_2^2 + \cdots + X_n^2$$

tikimybinis skirstinys vadinamas $\boldsymbol{\chi^2}$**skirstiniu** (chi-kvadrato skirstiniu) su n laisvės laipsnių. Trumpai žymima $\chi_n^2 \sim C\ (n)$.

Sakoma, kad dydis

$$t_n = \frac{Y}{\sqrt{\frac{1}{n}\chi_n^2}}$$

turi **Stjudento skirstinį** su n laisvės laipsnių (žym. $t_n \sim S(n)$), jei $Y \sim N(0,1)$ ir $\chi_n^2 \sim C\ (n)$ yra nepriklausomi atsitiktiniai dydžiai.

Turime nepriklausomų bandymų seką, tardami, kad kiekviename bandyme gali įvykti nesutaikomi įvykiai $A_1, A_2, \dots, A_k$. Įvykio A_i tikimybė kiekviename bandyme lygi p_i ($p_1 + \cdots + p_k = 1$) ir nepriklauso nuo kitų bandymų rezultatų.

Per n bandymų įvykusių įvykių A_i skaičių žymėsime X_i $(i = 1,2, \dots, k)$. Tada daugiamačio atsitiktinio dydžio $X = (X_1, \dots, X_{k-1})$ tikimybinis skirstinys vadinamas **polinominiu** ir žymimas $X \sim P(n, p_1, \dots, p_k)$ arba $X \sim P_{k-1}(n, p)$. Galimos atsitiktinio dydžio X reikšmės yra visi galimi rinkiniai $(n_1, \dots, n_{k-1})$, kuriuose n_i – sveikieji neneigiami skaičiai ir $0 \le n_i \le n$, $n_1 + \cdots + n_{k-1} \le n$. Pagal kombinatorikos formules gauname ($n_k = n - n_1 - \cdots - n_{k-1}$):

$$P\{X_1 = n_1, \dots, X_{k-1} = n_{k-1}\} = \frac{n!}{n_1! \dots n_k!} p_1^{n_1} \dots p_k^{n_k}.$$

1.3.1. Skirstinio formos charakteristikos

Šiam skyreliui parengti naudoti [5], [6] literatūros šaltiniai.

Normaliojo skirstinio tankio funkcijos grafikui apibūdinti naudojamos dvi dažnių skirstinio formos charakteristikos: *eksceso* ir *asimetrijos* koeficientai.

Pirmiausia apibrėšime centrinio empirinio momento sąvoką.

Centriniu empiriniu j-tosios eilės *momentu* (žymimu m_j) vadinamas dydis

$$m_j = \frac{1}{n-1}\sum_{k=1}^{n}(x_k - \bar{x})^j.$$

Imties asimetrijos koeficientas (*skewness*)

$$A = \frac{m_3}{s^3}$$

yra histogramos simetrijos matas. Jeigu $A > 0$, tai histogramos asimetrija teigiama (dešinioji), jeigu $A < 0$, tai asimetrija neigiama (kairioji) (žr. 1 pav.). Histograma simetriška, kai $A = 0$. Jei $A > 0$ ($A < 0$), tai $x > Md$ ($x < Md$).

1 pav. *Dešinioji ir kairioji asimetrija* [5]

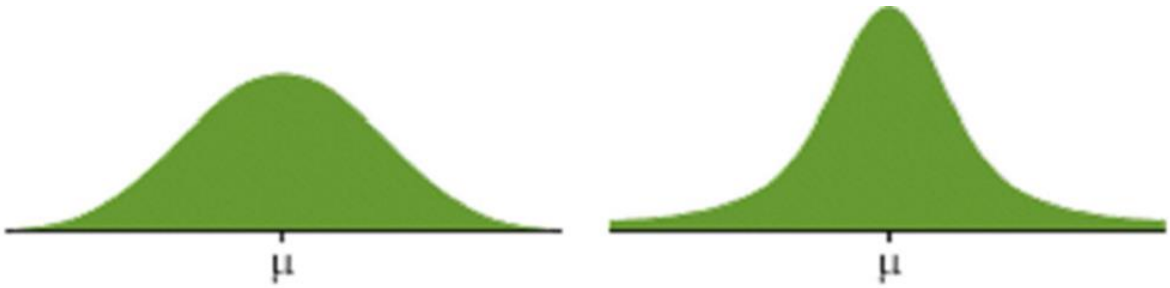
Asimetrijos rodiklio standartinė paklaida apytiksliai gali būti apskaičiuojama pagal formulę:

$$S_A = \sqrt{\frac{6}{n}}.$$

Imties eksceso koeficientas (*kurtosis*)

$$E = \frac{m_4}{s^4} - 3$$

yra histogramos lėkštumo matas. Jeigu $E = 0$, tai sklaida apie vidurkį tokia pati kaip ir normaliosios kreivės. Jeigu $E > 0$ – histograma smaila, t. y. duomenų sklaida apie vidurį yra mažesnė nei normaliosios kreivės, o jeigu $E < 0$, tai histograma lėkšta, t. y. duomenų sklaida apie vidurį didesnė nei normaliosios kreivės atveju (žr. 2 pav.).



2 pav. *Eksceso koeficientai:* $E < 0$ *ir* $E > 0$ [5]

Eksceso rodiklio standartinė paklaida apytiksliai gali būti apskaičiuojama pagal tokią formulę:

$$S_E = 2 \cdot \sqrt{\frac{6}{n}}.$$

Asimetrijos ir eksceso paklaidos rodo, kiek dėl atsitiktinių faktorių poveikio imties asimetrijos ir eksceso rodikliai gali vidutiniškai svyruoti apie jų tikrąsias (aibės) reikšmes.

Turint asimetrijos ir eksceso rodiklių reikšmes ir paklaidas, su tam tikru patikimumu galima įvertinti tyrimo duomenų pasiskirstymo dėsnio skirtumą nuo normaliojo.

Esant bet kokiam atvejų skaičiui, jeigu trigubos asimetrijos arba eksceso rodiklių standartinės paklaidos yra mažesnės arba lygios apskaičiuotoms asimetrijos ir eksceso koeficientų reikšmėms ($3S_A \leq |A|$ arba $3S_E \leq |E|$), galima su dideliu patikimumu tvirtinti, kad empirinis skirstinys tikrai skiriasi nuo normaliojo. Jeigu $3S_A > |A|$ ir $3S_E > |E|$, tikėtina, kad empirinis skirstinys reikšmingai nesiskiria nuo normaliojo, nes asimetrija ir ekscesas neviršija trigubos standartinės paklaidos ribų.

1.4. Hipotezės ir jų tikrinimas

Šiam poskyriui parengti naudoti [6], [8], [14] literatūros šaltiniai.

Hipotezė – teiginys arba prielaida apie vienos ar kelių populiacijų parametrus. Hipotezės skirstomos į parametrines (apie skirstinio parametrus) ir neparametrines (apie patį skirstinį).

Tikrinama hipotezė vadinama pagrindine arba nuline ir žymima H_0. Jai priešinga hipotezė vadinama alternatyviąja ir žymima H_1.

Pagrindinėje (nulinėje) hipotezėje (H_0) teigiama, kad tarp lyginamųjų populiacijų parametrų arba grupių statistikos reikšmingo skirtumo nėra. Alternatyviojoje hipotezėje (H_1) teigiama, kad yra reikšmingas skirtumas tarp lyginamųjų populiacijų parametrų arba grupių statistikos.

Hipotezė, sudaryta iš vienos galimybės (taško), vadinama paprastąja. Tokiu būdu mes sprendžiame dviejų paprastųjų hipotezių tikrinimo uždavinį.

Hipotezių H_0 ir H_1 tikrinimo statistiniu kriterijumi vadinama kiekviena statistika $\delta = \delta(X)$, įgyjanti dvi reikšmes 0 ir 1. Kai $\delta(X)$ įgyja reikšmę 0, tada hipotezė H_0 priimama (H_1 atmetama), o kai įgyja reikšmę 1, tada hipotezė H_1 priimama, o H_0 atmetama.

Sritis W, kurioje priimama hipotezė H_1 (atmetama H_0), vadinama kritine sritimi.

Skaičius α_0 (atitinkamai α_1), vadinamas kriterijaus $\delta = \delta(X)$ 1-osios (atitinkamai 2-osios) rūšies klaidos tikimybe, žymi tikimybę atmesti hipotezę H_0 (atitinkamai H_1), kai ji teisinga.

Kriterijus $\delta = \delta(X)$ bus tuo geresnis, kuo mažesnės jo abiejų rūšių klaidų tikimybės. Dažniausiai parenkamas toks reikšmingumo lygmuo α, kurio neturėtų viršyti pirmosios rūšies klaidos tikimybė. Skaičius α paprastai parenkamas artimas nuliui: 0,1; 0,05; 0,025; 0,01.

Praktiniuose uždaviniuose dažniausiai suformuluojama tik nulinė hipotezė H_0, o alternatyvą H_1 reikia laikyti jos priešingybe. Tokiuose uždaviniuose hipotezės H_0 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) $W = \left\{X^n : K(X^n) < t_{\frac{\alpha}{2}} \; arba \; K(X^n) > t_{1-\frac{\alpha}{2}}\right\}$,

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_0, yra gerai žinomas.

Jei 3) atveju statistikos $K(X^n)$ skirstinio tankio funkcija simetriška 0_y ašies atžvilgiu, tada kritinė sritis W įgyja pavidalą:

3′) $W = \left\{X^n : |K(X^n)| > t_{1-\frac{\alpha}{2}}\right\}$.

1.4.1. Hipotezė apie dviejų nepriklausomų imčių vidurkių lygybę

Skyreliui parengti naudotasi [3] literatūros šaltiniu.

Vidurkių lygybei tikrinti naudojamas Stjudento kriterijus. Prieš tikrinant hipotezę apie dviejų nepriklausomų imčių vidurkių lygybę reikia išsiaiškinti, ar nagrinėjamų populiacijų dispersijos yra lygios ar nelygios, tai yra patikrinti hipotezę apie dviejų dispersijų lygybę.

Tarkime, atsitiktinės imtys $(X_1, X_2, \dots, X_n)$ ir $(Y_1, Y_2, \dots, Y_m)$ gautos stebint du nepriklausomus normaliuosius atsitiktinius dydžius, kurių dispersijos σ_X^2 ir σ_Y^2 nežinomos. Tikrinama hipotezė $\begin{cases} H_0 : \sigma_X^2 = \sigma_Y^2 \\ H_1 : \sigma_X^2 \neq \sigma_Y^2 \end{cases}$. Dispersijų įverčiai yra s_X^2 ir s_Y^2. Dispersijų lygybės kriterijus grindžiamas tuo, kad jei H_0 teisinga, tai santykis

$$F = \frac{s_X^2}{s_Y^2} \tag{1}$$

turi Fišerio skirstinį su $(n-1)$ ir $(m-1)$ laisvės laipsnių.

Hipotezė H_0 atmetama, jeigu $F > F_{\frac{\alpha}{2}}(n-1, m-1)$ arba $F < F_{1-\frac{\alpha}{2}}(n-1, m-1)$; čia $F_{\frac{\alpha}{2}}(n-1, m-1)$ yra Fišerio skirstinio su $(n-1)$ir $(m-1)$ laisvės laipsnių $\frac{\alpha}{2}$ lygmens kritinė reikšmė. Hipotezė H_0 neatmetama, jeigu $F_{1-\frac{\alpha}{2}}(n-1, m-1) \leq F \leq F_{\frac{\alpha}{2}}(n-1, m-1)$.

Tarkime, atsitiktinės imtys $(X_1, X_2, \dots, X_n)$ ir $(Y_1, Y_2, \dots, Y_m)$ gautos stebint du nepriklausomus normaliuosius atsitiktinius dydžius $X \sim N(\mu_X, \sigma^2)$ ir $Y \sim N(\mu_Y, \sigma^2)$, su lygiomis nežinomomis dispersijomis, kurių vidurkiai μ_X ir μ_Y nežinomi. Tikrinama statistinė hipotezė $\begin{cases} H_0: \mu_X = \mu_Y \\ H_1: \mu_X \neq \mu_Y \end{cases}$.

Kritinė sritis sudaroma remiantis tuo, kad statistika

$$t = \frac{\bar{X} - \bar{Y}}{\sqrt{s_p^2 \left(\frac{1}{n} + \frac{1}{m}\right)}}$$

turi Stjudento skirstinį su $(n + m - 2)$ laisvės laipsnių, kai $\mu_X = \mu_Y$.

Hipotezė H_0 atmetama, jeigu $|t| > t_{\frac{\alpha}{2}}(n + m - 2)$; čia $t_{\frac{\alpha}{2}}(n + m - 2)$ yra Stjudento skirstinio su $(n-1)$ laisvės laipsnių $\frac{\alpha}{2}$ lygmens kritinė reikšmė. Hipotezė H_0 neatmetama, jeigu $|t| \leq t_{\frac{\alpha}{2}}(n + m - 2)$.

Tarkime, dvi atsitiktinės imtys $(X_1, X_2, \dots, X_n)$ ir $(Y_1, Y_2, \dots, Y_m)$ gautos stebint du nepriklausomus normaliuosius atsitiktinius dydžius $X \sim \mathcal{N}(\mu_X, \sigma_X^2)$ ir $Y \sim \mathcal{N}(\mu_Y, \sigma_Y^2)$ su nelygiomis dispersijomis. Vidurkiai μ_X, μ_Y ir dispersijos σ_X^2, σ_Y^2 nežinomi.

Statistinei hipotezei $\begin{cases} H_0: \mu_X = \mu_Y \\ H_1: \mu_X \neq \mu_Y \end{cases}$ tikrinti apskaičiuojama kriterijaus statistika:

$$t = \frac{\bar{x} - \bar{y}}{\sqrt{\frac{s_x^2}{n} + \frac{s_y^2}{m}}};$$

čia $\bar{x}$, $\bar{y}$ yra imčių vidurkiai, s_x^2, s_y^2 – imčių dispersijos, o n, m – imčių didumai.

Hipotezė H_0 atmetama, jeigu $|t| > t_{\frac{\alpha}{2}}(k)$; čia $t_{\frac{\alpha}{2}}(k)$ yra Stjudento skirstinio su k laisvės laipsnių $\frac{\alpha}{2}$ lygmens kritinė reikšmė. Laisvės laipsnių skaičius k yra mažiausias sveikasis skaičius, tenkinantis sąlygą

$$k \leq \frac{\left(\frac{s_x^2}{n} + \frac{s_y^2}{m}\right)^2}{\frac{s_x^4}{n^3} + \frac{s_y^4}{m^3}}.$$

1.4.2. Hipotezė apie dviejų koreliacijos koeficientų lygybę

Skyreliui parengti naudotasi [3] literatūros šaltiniu.

Stebime dvi nepriklausomas intervalinių kintamųjų poras (X_1, Y_1) ir (X_2, Y_2). Atsitiktines imtis sudaro poros (X_{11}, Y_{11}), $(X_{12}, Y_{12}), \dots, (X_{1n}, Y_{1n})$ ir $(X_{21}, Y_{21}), (X_{22}, Y_{22}), \dots, (X_{2m}, Y_{2m})$. Reikia nustatyti, ar koreliacija (X_1) su (Y_1) (pažymime ją ϱ_1) skiriasi nuo koreliacijos (X_2) su (Y_2) (pažymime ją ϱ_2). Kadangi empirinių koreliacijos koeficientų R_1 ir R_2 skirtumo skirstinys yra asimetriškas, prieš taikydami normaliąją aproksimaciją, naudojamės Fišerio logaritmine transformacija:

$$\mathrm{Z}_1 = \frac{1}{2} ln \frac{1 + R_1}{1 - R_1},$$

$$\mathrm{Z}_2 = \frac{1}{2} ln \frac{1 + R_2}{1 - R_2}.$$

Kai teisinga hipotezė $H_0: \varrho_1 = \varrho_2$ ir $n > 3$, $m > 3$, statistika

$$\frac{\mathrm{Z}_1 - \mathrm{Z}_2}{\sqrt{\frac{1}{n-3} + \frac{1}{m-3}}} \sim \mathcal{N}(0,1).$$

Remiantis šia formule sudaromos kritinės sritys.

Tikrinama statistinė hipotezė:

$$\begin{cases} H_0: \varrho_1 = \varrho_2, \\ H_1: \varrho_1 \neq \varrho_2. \end{cases}$$

Kriterijaus statistika

$$Z = \frac{z_1 - z_2}{\sqrt{\frac{1}{n-3} + \frac{1}{m-3}}};$$

čia $z_i = \frac{1}{2} ln \frac{1+r_i}{1-r_i}$, $i = 1,2$, o r_1, r_2 yra Pirsono koreliacijos koeficientų R_1 ir R_2 realizacijos.

Hipotezė H_0 atmetama, jeigu $|Z| > z_{\frac{\alpha}{2}}$; čia $z_{\frac{\alpha}{2}}$ yra standartinio normaliojo skirstinio $\frac{\alpha}{2}$ lygmens kritinė reikšmė. Hipotezė H_0 neatmetama, jeigu $|Z| \leq z_{\frac{\alpha}{2}}$.

1.4.3. Hipotezė apie dviejų proporcijų lygybę

Skyreliui parengti naudotasi [3] literatūros šaltiniu.

Stebime du nepriklausomus binominius kintamuosius. Pirmoje n elementų imtyje yra k_1 vienetų (likę – nuliai), antrojoje m elementų imtyje yra k_2 vienetų (likę – nuliai).

Tikriname statistinę hipotezę:

$$\begin{cases} H_0: p_1 = p_2, \\ H_1: p_1 \neq p_2. \end{cases}$$

Apskaičiuojame kriterijaus statistiką:

$$Z = \frac{\widehat{p_1} - \widehat{p_2}}{\sqrt{\bar{p}(1-\bar{p})\left(\frac{1}{n}+\frac{1}{m}\right)}}; \tag{2}$$

čia $\widehat{p_1} = \frac{k_1}{n}, \widehat{p_2} = \frac{k_2}{m}, \bar{p} = \frac{k_1+k_2}{n+m}$.

Hipotezė H_0 atmetama, jeigu $|Z| > z_{\frac{\alpha}{2}}$; čia $z_{\frac{\alpha}{2}}$ yra standartinio normaliojo skirstinio $\frac{\alpha}{2}$ lygmens kritinė reikšmė. Hipotezė H_0 neatmetama, jeigu $|Z| \leq z_{\frac{\alpha}{2}}$.

1.4.4. Polinominio skirstinio taikymas

Skyreliui parengti naudotasi [9], [1] literatūros šaltiniais.

Tarkime, kad turime paprastąją atsitiktinę imtį $X^n = (X_1, X_2, \dots, X_n)$, čia $X_1, X_2, \dots, X_n$ yra nepriklausomi atsitiktiniai elementai, simbolizuojantys raides. Bendrą raidžių skaičių žymime n. Tegu A_j simbolizuoja j-ąjį autorių, kai $j = 1,2,3,4,5$. Y_j pažymėkime A_j raidžių skaičių.

$$Y_j = \sum_{i=1}^{n} \mathbb{I}(X_i \in A_j), \quad j = 1,2,3,4,5.$$

Vietoje pradinės imties $X^n = (X_1, X_2, \dots, X_n)$ gauname imtį $Y^k = (Y_1, Y_2, \dots, Y_k)$. Atsitiktinis vektorius $Y^k = (Y_1, Y_2, \dots, Y_k)$ turi polinominį skirstinį $X^k \sim P = (p_1, \dots, p_k, n)$, t. y.

$$P\{Y_1 = n_1, Y_2 = n_2, \dots, Y_k = n_k\} = \frac{n!}{n_1! \dots n_k!} p_1^{n_1} \dots p_k^{n_k},$$

čia $n_1 + \dots + n_k = n$ ir $p_1 + p_2 + \dots + p_k = 1$, p_i $(i = 1, \dots k)$ yra tikimybės.

Yra žinoma, kad parametrų p_i maksimalaus tikėtinumo įverčių $\widehat{p_i}$ išraiška yra tokia: $\frac{Y_i}{n}$.

Tikrinsime statistinę hipotezę $H_0: p_i = p_{i0},\ i = 1, \dots, k$ taikydami chi-kvadrato kriterijų. Parinkę reikšmingumo lygmenį α norimu tikslumu sužinosime, ar tam tikrų raidžių (raidžių grupių) vartojimas žymiai skiriasi skirtingų autorių imtyse. Jei hipotezė H_0 priimama (su tikimybe $1 - \alpha$), tai raidžių vartojama vienodai visose imtyse. Jei H_0 atmetame, tada galime sakyti, kad raidžių vartojimas skirtingose imtyse yra nevienodas. Hipotezės $H_0: p_i = p_{i0},\ i = 1, \dots, k$ tikrinimui yra sudaroma Pirsono statistika

$$\chi^2 = \sum_{i=1}^{k} \frac{(Y_i - n \cdot p_{i0})^2}{n \cdot p_{i0}} = n \cdot \sum_{i=1}^{k} \frac{(\widehat{p_i} - p_{i0})^2}{p_{i0}}, \tag{3}$$

kuri asimptotiškai turi χ^2 skirstinį su $k - 1$ laisvės laipsnių, kai hipotezė H_0 teisinga. Jei

$$\chi^2 \leq {\chi^2}_{1-\alpha}(k-1),$$

čia ${\chi^2}_{1-\alpha}(k-1)$ yra χ^2 skirstinio su $k - 1$ laisvės laipsnių $1 - \alpha$ eilės kvantilis, tada hipotezę priimame.

1.4.5. Požymių nepriklausomumo tikrinimas

Skyreliui parengti naudotasi [3] literatūros šaltiniu.

Norint išsiaiškinti, ar stebimi intervaliniai kintamieji yra nepriklausomi ar priklausomi, tiesinės priklausomybės matas yra Pirsono koreliacijos koeficientas. Jei kintamieji yra kokybiniai, taikome χ^2 kriterijų. Kategoriniai kintamieji dažnai vadinami požymiais.

Tarkime, turime porinių stebėjimų imtį $(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$, gautą stebint kategorinių kintamųjų porą (X, Y). Tikrinant hipotezę, kad kintamieji X ir Y yra nepriklausomi, duomenų aibę užrašome porine dažnių lentele (žr. 1 lent.). Čia $n_{i\cdot} = \sum_{j=1}^{c} o_{ij}$ yra imties narių, kurių požymio X reikšmė yra x_i, skaičius; $n_{\cdot j} = \sum_{i=1}^{r} o_{ij}$ – imties narių, kurių požymio Y reikšmė yra y_j, skaičius; $n = \sum_{i=1}^{r} \sum_{j=1}^{c} o_{ij}$ – imties didumas.

1 lentelė. *Porinė dažnių lentelė*

	y_1	y_2	...	y_c	$\sum$
x_1	o_{11}	o_{12}	...	o_{1c}	$n_{1\cdot}$
x_2	o_{21}	o_{22}	...	o_{2c}	$n_{2\cdot}$
...	...	...	...	...	...
x_r	o_{r1}	o_{r2}	...	o_{rc}	$n_{r\cdot}$
$\sum$	$n_{\cdot 1}$	$n_{\cdot 2}$	...	$n_{\cdot c}$	n

Suformuluojame tikimybinį uždavinio modelį.

Tarkime, $p_{ij} = P(X = x_i, Y = y_j)$ – populiacijos dalis, kuriai matuojamų požymių pora (X, Y) įgyja reikšmę (x_i, y_j); $p_i = P(X = x_i)$ – populiacijos dalis, kuriai požymis X įgyja reikšmę x_i; $q_j = P(Y = y_i)$ – populiacijos dalis, kuriai požymis Y įgyja reikšmę y_j. Aišku, kad

$$\sum_{i=1}^{r} p_i = \sum_{j=1}^{c} q_j = 1.$$

Tikriname statistinę hipotezę:

$$\begin{cases} H_0: p_{ij} = p_i q_j, & \text{su visais i} = 1, \dots, \text{r}, \text{j} = 1, \dots, \text{c}; \\ H_1: p_{ij} \neq p_i q_j, & \text{bent vienai porai (i, j)}. \end{cases}$$

Jei hipotezė teisinga, tai p_{ij} yra nežinomų parametrų $p_{i,} q_j, i = 1, \dots, r, j = 1, \dots, c$, funkcijos. Atsižvelgiant į tai, kad kiekvieną įvykį apibūdina du indeksai i ir j, kriterijaus statistiką galima užrašyti taip:

$$\chi^2 = \sum_{i=1}^{r} \sum_{j=1}^{c} \frac{\left(O_{ij} - n\widehat{p_{ij}}\right)^2}{n\widehat{p_{ij}}}.$$

Nežinomų parametrų įverčiai, gauti χ^2 minimumo metodu, yra:

$$\widehat{p_i} = \frac{n_{i\cdot}}{n}, i = 1, \dots, r;$$

$$\widehat{q_j} = \frac{n_{\cdot j}}{n}, j = 1, \dots, c.$$

Tikėtinieji dažniai E_{ij} apskaičiuojami pagal formulę

$$E_{ij} = n\widehat{p_{ij}} = n \cdot \widehat{p_i} \cdot \widehat{q_j} = n \cdot \frac{n_{i\cdot}}{n} \cdot \frac{n_{\cdot j}}{n} = \frac{n_{i\cdot} n_{\cdot j}}{n}.$$

Iš čia išplaukia, kad esant teisingai hipotezei H_0 statistika

$$\chi^2 = \sum_{i=1}^{r} \sum_{j=1}^{c} \frac{\left(O_{ij} - \frac{n_{i\cdot} n_{\cdot j}}{n}\right)^2}{\frac{n_{i\cdot} n_{\cdot j}}{n}} \quad (4)$$

yra asimptotiškai pasiskirsčiusi pagal χ^2 dėsnį su dėsnį su $rc - [(r-1) + (c-1)] - 1 = rc - r - c + 1 = (r-1)(c-1)$ laisvės laipsnių. Hipotezė H_0 apie kintamųjų nepriklausomumą yra atmetama, kai apskaičiuotos statistikos χ^2 reikšmė yra didesnė už χ^2 skirstinio su $(r-1)(c-1)$ laisvės laipsnių α lygmens kritinę reikšmę; čia α – pasirinktas reikšmingumo lygmuo.

1.5. Vienfaktorinė dispersinė analizė

Poskyriui parengti naudotasi [4] literatūros šaltiniu.

Vienfaktorinei dispersinei analizei žymėti plačiai naudojama santrumpa ANOVA (*ANalysis Of VAriance*). Vienfaktorinė dispersinė analizė naudojama tada, kai populiacijas viena nuo kitos tyrėjas skiria tik pagal vieną požymį. Kategorinis kintamasis (populiacijos požymis), pagal kurį skiriamos populiacijos viena nuo kitos, vadinamas nepriklausomuoju kintamuoju, arba faktoriumi. Dispersinės analizės tikslas yra nuspręsti, ar priklausomo kintamojo, išmatuoto skirtingose populiacijose, vidurkiai skiriasi.

Taikydami ANOVA, lyginame kelių nepriklausomų imčių vidurkius.

Tarkime turime k nepriklausomų populiacijų. Priklausomas kintamasis, matuojamas i-ojoje populiacijoje, vadinamas populiacijos kintamuoju. Populiacijų kintamuosius pažymime $X_1, X_2, \dots, X_k$. Iš kiekvienos populiacijos parenkama paprastoji atsitiktinė imtis: $X_{i1}, X_{i2}, \dots, X_{in_i}$; čia $i = 1,2, \dots k$ – populiacijos numeris, o n_i – i-osios imties didumas. ANOVA duomenys pateikiami 2 lentele.

2 lentelė. *ANOVA duomenys*

X_1	X_2	X_3	...	X_k
X_{11}	X_{21}	X_{31}	...	X_{k1}
X_{12}	X_{22}	X_{32}	...	X_{k2}
X_{13}	X_{23}	X_{33}	...	X_{k3}
...	...	...	...	...
X_{1n_1}	X_{2n_2}	X_{3n_3}	...	X_{kn_k}

Struktūrinis ANOVA modelis i-osios imties j-ajam stebėjimui X_{ij} užrašomas taip:

$$X_{ij} = \mu_i + e_{ij} = \mu + \tau_i + e_{ij}; \quad (5)$$

čia $\mu_i = \boldsymbol{E}X_i = \boldsymbol{E}X_{ij}$ yra i-osios populiacijos kintamojo vidurkis; e_{ij} – atsitiktinė paklaida; μ – bendrasis visų populiacijų vidurkis; $\tau_i = \mu_i - \mu$ yra i-osios populiacijos vidurkio ir bendrojo vidurkio skirtumas (kartais τ_i dar vadinamas i-osios populiacijos efektu). Modelis nusako, kokią potencialą reikšmę gali įgyti i-osios imties j-asis elementas, todėl (5) formulėje X_{ij} yra atsitiktinis dydis. Beje μ, μ_i, τ_i yra skaičiai, o e_{ij} – atsitiktinis dydis.

1.5.1. Dispersinės analizės prielaidos

Taikant ANOVA, tiriama k populiacijų. Pirmoje populiacijoje stebimas kintamasis X_1, antrojoje – $X_2, \dots, k$-ojoje – X_k . Kad būtų galima taikyti vienfaktorinę dispersinę analizę, reikia, kad būtų:

1. kintamieji pasiskirstę pagal normalųjį dėsnį;
2. kintamųjų dispersijos lygios;
3. kintamieji nepriklausomi.

Vienfatorinė dispersinė analizė ganėtinai atspari nedideliems ir net vidutiniams nuokrypiams nuo pirmųjų dviejų prielaidų, tačiau tik tada, kai imčių didumai panašūs.

Iš pirmųjų dviejų reikalavimų išplaukia, kad

$$X_1 \sim N(\mu_1, \sigma^2), X_2 \sim N(\mu_2, \sigma^2), \dots, X_k \sim N(\mu_k, \sigma^2).$$

Taigi kintamieji $X_1, X_2, \dots, X_k$ turi visiškai identiškos formos tankius, kurie gali skirtis tik padėties parametrais (vidurkiais). ANOVA nelygių dispersijų atveju modifikacijų neturi.

Ar galima ANOVA taikyti, kai ne visų populiacijų dispersijos lygios, labai priklauso nuo imčių didumo. Jeigu didesnė imtis yra iš populiacijos, kurios dispersija didesnė, kriterijus tampa per daug konservatyvus – apskaičiuotoji p-reikšmė yra didesnė nei tikroji. Todėl rečiau, nei turėtų būti, pripažįstamas populiacijų vidurkių skirtumas, didėja antrosios rūšies klaidos tikimybė ir mažėja kriterijaus galia. Jeigu iš populiacijos, kurios dispersija didesnė, parinkta maža imtis, kriterijus tampa per daug liberalus – apskaičiuotoji p-reikšmė yra mažesnė nei tikroji. Todėl dažniau skelbiamas populiacijų vidurkių skirtumas, nors iš tikrųjų jo ir nėra; didėja pirmosios rūšies klaidos

tikimybė. Kuo mažiau skiriasi imčių didumai, tuo mažiau ANOVA jautri populiacijų dispersijų skirtumams. Žinoma, net ir vienodo didumo imčių dispersijos negali per daug skirtis. Reikalaujama, kad didžiausioji iš imčių būtų didesnė už mažiausiąją ne daugiau kaip tris kartus.

Visi populiacijų kintamieji $X_1, X_2, \ldots, X_k$ yra nepriklausomi. Šios prielaidos pažeidimas gali turėti didelės įtakos ANOVA rezultatams, todėl taikydami vienfaktorinę dispersinę analizę turime būti tikri, kad duomenys gauti stebint nepriklausomus populiacijų kintamuosius.

Taikant vienfaktorinę dispersinę analizę, labai svarbu, kad visų imčių didumai skirtųsi kuo mažiau. Kuo vienodesni imčių didumai, tuo ANOVA atsparesnė kintamųjų normalumo bei lygių dispersijų reikalavimų pažeidimams. Žinoma rezultatai tiesiogiai priklauso ir nuo pačių imčių didumų. Jei kiekvienoje imtyje stebėjimų labai daug, tai net ir nedideli imčių vidurkių skirtumai laikomi statistiškai reikšmingais.

1.5.2. Kriterijaus apie vidurkių lygybę sudarymas

Tarkime, kad populiacijose stebimi nepriklausomi kintamieji:

$$X_1 \sim N(\mu_1, \sigma^2), X_2 \sim N(\mu_2, \sigma^2), \ldots, X_k \sim N(\mu_k, \sigma^2).$$

Nulinė vienfaktorinės dispersinės analizės hipotezė teigia, kad visi populiacijų kintamųjų vidurkiai lygūs, t. y. $H_0: \mu_1 = \mu_2 = \cdots = \mu_k$.

Vienfaktorinė dispersinė analizė grindžiama dviejų dispersijos σ^2 įverčių palyginimu. Tarkime, kad visų imčių didumų suma lygi $N = n_1 + n_2 + \cdots + n_k$. Tegul bendrasis visų imčių vidurkis yra $\bar{X}$, o $\overline{X_i}$ yra i-osios imties vidurkis. Imkime pirmosios imties pirmąjį elementą X_{11}. Jo ir bendrojo vidurkio skirtumą galima išskaidyti į dvi komponentes:

$$X_{11} - \bar{X} = (X_{11} - \overline{X_1}) + (\overline{X_1} - \bar{X}). \quad (6)$$

Palyginę šią formulę su (5), matome, kad (6) visiškai analogiška savo sandara struktūriniam vienfaktorinės dispersinės analizės modeliui. Iš tikrųjų $\bar{X}$ atitinka μ; $X_{11} - \overline{X_1}$ atitinka e_{11}, o $\overline{X_1} - \bar{X}$ atitinka τ_1.

Pakėlus abi (6) lygybės puses kvadratu

$$(X_{11} - \bar{X})^2 = (X_{11} - \overline{X_1})^2 + (\overline{X_1} - \bar{X})^2 + 2(X_{11} - \overline{X_1})(\overline{X_1} - \bar{X})$$

ir analogiškai išreiškus visų pirmos imties elementų ir bendrojo vidurkio skirtumus bei gautus rezultatus sudėjus, gaunama:

$$(X_{11} - \bar{X})^2 + (X_{12} - \bar{X})^2 + \cdots + \left(X_{1n_1} - \bar{X}\right)^2 = (X_{11} - \overline{X_1})^2 + \cdots + \left(X_{1n_1} - \overline{X_1}\right)^2 +$$
$$+ n_1(\overline{X_1} - \bar{X})^2 + 2(\overline{X_1} - \bar{X}) \cdot \left((X_{11} - \overline{X_1}) + (X_{12} - \overline{X_1}) + \cdots + \left(X_{1n_1} - \overline{X_1}\right)\right). \quad (7)$$

Paskutinis (7) lygybės dėmuo lygus nuliui, todėl

$$\sum_{j=1}^{n_1} \left(X_{1j} - \bar{X}\right)^2 = \sum_{j=1}^{n_1} \left(X_{1j} - \overline{X_1}\right)^2 + n_1(\overline{X_1} - \bar{X})^2.$$

Analogiškai susitvarkius su kitomis imtimis ir rezultatus susumavus, gaunama

$$SST = SSW + SSB; \tag{8}$$

čia $SST = \sum_{j=1}^{k}\sum_{j=1}^{n_i}\left(X_{ij} - \bar{X}\right)^2$, $SSW = \sum_{j=1}^{k}\sum_{j=1}^{n_i}\left(X_{ij} - \bar{X}_\iota\right)^2$, $SSB = \sum_{j=1}^{k} n_i\,(\bar{X}_\iota - \bar{X})^2$.

SST apima visų duomenų ir bendrojo vidurkio skirtumus ir vadinama visa kvadratų suma. Pirmoji SST komponentė SSW įvertina kiekvienos imties duomenų sklaidą (apie imties vidurkį) ir vadinama vidine kvadratų suma, o antroji komponentė SSB įvertina imčių vidurkių ir bendrojo vidurkio skirtumus ir vadinama grupių kvadratų suma.

Pasinaudojus (8) skaidiniu galima sudaryti du nežinomos dispersijos σ^2 įverčius. Tarkime, kad visų populiacijų vidurkiai lygūs: $\mu_1 = \mu_2 = \cdots = \mu$. Tuomet:

$$\boldsymbol{E}SSW = (N-k)\sigma^2, \boldsymbol{E}SST = (N-1)\sigma^2,$$

$$\boldsymbol{E}SSB = \boldsymbol{E}SST - \boldsymbol{E}SSW = (N-k)\sigma^2 - (N-1)\sigma^2 = (k-1)\sigma^2.$$

Sunormavus kvadratų sumas, gaunami du dispersijos įverčiai:

$$MSW = \frac{SSW}{N-k} = \frac{(n_1-1)S_1^2 + (n_2-1)S_2^2 + \cdots + (n_k-1)S_k^2}{n_1 + n_2 + \cdots + n_k - k},$$

$$MSB = \frac{SSB}{k-1} = \frac{n_1(\overline{X_1} - \bar{X})^2 + \cdots + n_k(\overline{X_k} - \bar{X})^2}{k-1};$$

čia S_i^2 yra i-osios imties dispersija.

ANOVA hipotezei tikrinti naudojama statistika

$$F = \frac{MSB}{MSW}.$$

Jei F reikšmė didelė, tai tikėtina, kad vidurkiai skiriasi, jei artima vienetui – ne.

Jeigu nulinė ANOVA hipotezė teisinga (visų populiacijų vidurkiai lygūs), tai F turi Fišerio skirstinį su $(k-1)$ ir $(N-k)$ laisvės laipsnių. Remiantis šiuo faktu, kiekvienam reikšmingumo lygmeniui sudaromos kritinės sritys.

Tegul reikšmingumo lygmuo lygus α. Hipotezė H_0 atmetama, jeigu $F > F_\alpha(k-1, N-k)$; čia $N = n_1 + n_2 + \cdots + n_k$, $F_\alpha(k-1, N-k)$ yra Fišerio skirstinio su $k-1$ ir $N-k$ laisvės laipsnių α lygmens kritinė reikšmė. Hipotezė H_0 neatmetama, jeigu $F \le F_\alpha(k-1, N-k)$.

Dydžius SSB ir SSW patogiau skaičiuoti pagal tokias formules:

$$T_i = X_{i1} + X_{i2} + .. + X_{in_i},\ T = X_{11} + \cdots + X_{kn_k} = T_1 + \cdots + T_k,$$

$$N = n_1 + n_2 + \cdots + n_k, \qquad SST = \sum_{i=1}^{k}\sum_{j=1}^{n_i} X_{ij}^2 - \frac{T^2}{N},$$

$$SSB = \sum_{i=1}^{k}\frac{T_i^2}{n_i} - \frac{T^2}{N}, \qquad SSW = \sum_{i=1}^{k}\sum_{j=1}^{n_i} X_{ij}^2 - \sum_{i=1}^{k}\frac{T_i^2}{n_i}$$

1.5.3. Bonferonio kriterijus

Kaip įrodė K.E. Bonferonis, jeigu $0 \le \alpha < 1, C > 0$, tai $1-(1-\alpha)^C \le C\alpha$. Todėl eksperimento reikšmingumo lygmuo α_E, t. y. tikimybė lyginant visas įmanomas poras nors kartą neteisingai nustatyti statistiškai reikšmingą dviejų imčių vidurkių skirtumą, neviršija $\frac{k(k-1)\alpha}{2}$; čia k yra imčių skaičius, o α – reikšmingumo lygmuo lyginant vieną porą imčių. Bonferonio kriterijus skamba taip: pasirenkamas eksperimento reikšmingumo lygmuo α_E ir visos imčių poros lyginamos taikant Stjudento kriterijų, esant reikšmingumo lygmeniui $\alpha = \frac{\alpha_E}{C}$; čia $C = \frac{k(k-1)}{2}$.

Bonferonio kriterijų galima taikyti ir naudojantis ANOVA lentele. Paprastai tai daroma lyginant i-ąją imtį su j-ąja ir apskaičiuojamas dydis

$$BSD_{ij} = t_{\frac{\alpha}{2}}(N-k)\sqrt{MSW\left(\frac{1}{n_i}+\frac{1}{n_j}\right)}; \tag{9}$$

čia $N = n_1 + n_2 + \cdots + n_k$, k – bendrasis imčių skaičius, $\alpha = \frac{2\alpha_E}{k(k-1)}$, α_E – pasirinktasis eksperimento reikšmingumo lygmuo, $t_{\frac{\alpha}{2}}(N-k)$ – Stjudento skirstinio su $(N-k)$ laisvės laipsnių $\frac{\alpha}{2}$ lygmens kritinė reikšmė, MSW yra imties dispersijos įvertis, pateikiamas ANOVA rezultatų lentelėje.

Vidurkiai $\bar{x}_i$ ir $\bar{x}_j$ statistiškai reikšmingai skiriasi, jeigu $\left|\bar{x}_i - \bar{x}_j\right| > BSD_{ij}$. Jeigu visų imčių didumai vienodi, tai $BSD_{ij} = BSD$, t. y. visiems lyginimams naudojame tą patį dydį.

Kai imčių daug, Bonferonio kriterijus tampa nebeefektyvus, nes labai sumažėja α. Beveik niekuomet nebefiksuojamas statistiškai reikšmingas vidurkių skirtumas, nors tikrieji populiacijų vidurkiai ir skiriasi. Šis kriterijus netaikytinas, kai imčių yra daug.

1.5.4. Koreliacijos koeficientas ICC

Tarkime, kad imčių didumai vienodi ($n_i = n$). Tegul populiacijos ANOVA tyrimui atsitiktinai parenkamos iš didesnės populiacijų aibės, t. y. nagrinėjame ne fiksuotą, o atsitiktinį ANOVA modelį. Tuomet vidurkiai $\mu_1, \mu_2, \dots, \mu_k$ tėra tam tikra galimų vidurkių imtis ir reikia naudotis tokiu dispersijos įverčiu

$$\sigma_\tau^2 = \frac{1}{k-1}\sum_{i=1}^{k}(\mu_i - \mu)^2.$$

Pasinaudojant 1.5.2. skyrelyje pateiktomis formulėmis gaunama koreliacijos koeficiento ICC išraiška:

$$ICC = \frac{\frac{MSB-MSW}{n}}{MSW + \frac{MSB-MSW}{n}} = \frac{MSB-MSW}{MSB + \frac{n-1}{MSW}} = \frac{F-1}{F+(n-1)}.$$

Jeigu imčių didumai skiriasi, šio koreliacijos koeficiento išraiškoje, užuot naudojus n, imamas imčių didumų harmoninis vidurkis $\bar{n} = \frac{k}{\frac{1}{n_1}+\cdots+\frac{1}{n_k}}$:

$$ICC = \frac{MSB - MSW}{MSB + \frac{\bar{n}-1}{MSW}} = \frac{F-1}{F+(\bar{n}-1)}. \quad (10)$$

Kuo ICC didesnis, tuo kiekvienos imties duomenys panašesni ir tuo labiau duomenų skirtumus lemia priklausymas konkrečiai populiacijai. ICC gali būti ir neigiamas (jei $F < 1$). Minimali reikšmė $ICC = \frac{-1}{n-1}$ rodo, kad visose imtyse duomenų įvairovė didžiausia (imtys labai heterogeniškos).

2. PROGRAMINĖ ĮRANGA

2.1. R programavimo kalba

Poskyriui parengti naudotasi [12], [19] literatūros šaltiniais.

R – tai integruota programavimo kalba ir darbo aplinka, skirta duomenims tvarkyti, skaičiuoti, modeliuoti, grafiškai vaizduoti. Daugelis vartotojų renkasi šią programą dėl jos puikių grafinių galimybių. Ji efektyviai tvarko duomenis, skaičiuoti naudoja masyvus ir matricas. Taip pat apima algoritmų apdorojimą, tiesinę regresiją, laiko eilutes, statistines išvadas. Tai – atvirojo kodo S programavimo kalbos dialektas.

Duomenų analizė su R atliekama keliais žingsniais: programuojama, transformuojama, ištiriami duomenys ir patikslinimos hipotezės, modeliuojama ir pateikiami rezultatai.

R daugiausiai naudojama mokslo įstaigose, taip pat atliekant statistinius tyrimus sveikatos priežiūros įmonėse, vyriausybėje. Šiuo programiniu paketu naudojasi ir daugelis didelių kompanijų tokių kaip „Uber“, „Airbnb“, „Facebook“.

R turi versijas, skirtas Windows, Macintosh, Unix, Linux operacinėms sistemoms. Gautus rezultatus galima pateikti skirtingais formatais: dokumentiniais (HTML, PDF arba Latex, Word) ir prezentacija (HTML, PDF beamer).

R pradėjo kurti Robert Gentleman ir Ross Ihaka 1993 m. Naujojoje Zelandijoje, Auckland universiteto Statistikos departamente. Nuo jų vardų pirmųjų raidžių ir kilo programos pavadinimas.

R programa sudaryta iš įvairios paskirties programų paketų, kurių visuma pateikiama visame pasaulyje išdėstytose veidrodinėse svetainėse *Comprehensive R Archive Network CRAN* (išsamus R archyvų tinklas).

Programinė priemonė R operuoja objektais. Viskas jai yra objektai (duomenų, funkciniai), todėl jos programos kompaktiškos.

R sistemos programinis branduolys nedidelis, bet jis gali būti pildomas papildomais programiniais paketais, kuriuos į savo kompiuterį iš veidrodinės svetainės pagal poreikį atsisiunčia ir susiinstaliuoja vartotojas. R programos paketų rinkinys yra pildomas naujais programų paketais, kuriuos kuria patys sistemos vartotojai. Savanorių programuotojų grupė, pasivadinusi R vystymo branduoliu, gali modifikuoti R programinės įrangos programų archyvą.

2.2. SPSS programinis paketas

Šiam skyriui parengti naudotasi [17] literatūros šaltiniu.

Šiuolaikinė statistika yra neatsiejama nuo kompiuterinės duomenų analizės, padedančios greitai ir efektyviai spręsti įvairius statistikos uždavinius. Programiniai paketai, kuriuose įdiegti modernūs matematinės statistikos metodai, o daugelis operacijų yra formalizuota, įgalina spręsti įvairius taikomuosius uždavinius ne tik gerai įvaldžiusiems tuos metodus, bet ir tik bendrą supratimą apie juos turintiems vartotojams. Tam tik reikia gerai suvokti, kam skirti konkretūs statistiniai metodai, kokia jų taikymo sritis ir kaip interpretuoti gautus rezultatus.

SPSS programinis paketas (*Statistical Package for the Social Sciences*) – vienas labiausiai paplitusių statistinės informacijos apdorojimo programinių paketų, tinkamų ir pradedančiajam, ir patyrusiam vartotojui. Pagrindinis SPSS programinio paketo privalumas – tai didelė šiuolaikinių statistinių analizės metodų pasirinktis bei duomenų analizės rezultatų vizualizavimo priemonių (duomenų pateikimo lentelių, diagramų, skirstinių kreivių) įvairovė, lengvai įvaldoma dialoginė sąsaja. SPSS programinis paketas taikomas sociologijoje, biologijoje, medicinoje, psichologijoje, rinkodaroje, kokybės valdymo procese.

3. DUOMENŲ ANALIZĖ

Tyrimui atlikti buvo pasirinkta dešimt 2017 m. geriausių lietuvių grožinės paauglių literatūros rašytojų ir populiariausi jų kūriniai, iš kurių atrinktos ir šiame tyrime analizuojamos tos knygos, kurios tarpusavyje yra panašiausios savo apimtimi sakiniais:

1. Renata Šerelytė (2016), „Žvaigždžių medžiaga“, Alma litter;
2. Kristina Gudonytė (2015) „Jie grįžta per pilnatį“, Tyto alba;
3. Unė Kaunaitė (2011), „Sudie, rytojau“, Žara;
4. Akvilina Cicėnaitė (2015), „Niujorko respublika“, Alma littera;
5. Neringa Vaitkutė (2015) „Devynetas tamsos nešėjų“, Nieko rimto.

Atlikus atsitiktinę puslapių generaciją sisteminiu ėmimo būdu iš kiekvienos knygos buvo išrinkta po penkis puslapius, iš kurių išrašoma po keturiasdešimt tyrime analizuojamų sakinių.
Sisteminio ėmimo būdo pavyzdys:

Renatos Šerelytės knygos „Žvaigždžių medžiaga“ apimtis – 200 puslapių, todėl $N = 200$. Iš visų knygų renkama po 5 imtis, todėl $n = 5$, atsitiktinai „Microsoft Excel“ sugeneruota reikšmė 12, todėl $q_0 = 12$.
Randame ėmimo žingsnį:

$$q = \frac{N}{n} = \frac{200}{5} = 40.$$

Kadangi $q_0 = 12$, tai pirmosios imties sakiniai pradedami imti iš dvylikto knygos puslapio. $q_0 + q = 52$, todėl antroji imtis pradedama imti iš penkiasdešimt antro knygos puslapio. $q_0 + 2q = 92$, bet 92 puslapis nėra pilnai užpildytas tekstu, sekantis (93) puslapis yra tuščias, todėl trečioji imtis pradedama imti iš devyniasdešimt ketvirto knygos puslapio. $q_0 + 3q = 132$, todėl ketvirtoji imtis pradedama imti iš šimtas trisdešimt antro knygos puslapio. $q_0 + 4q = 172$, todėl penktoji imtis pradeda imti iš šimtas septyniasdešimt antro knygos puslapio.

Tyrimo metu analizuojami tokie duomenys:

- kiekvieno autoriaus pasikartojantys žodžiai;
- kiekvieno autoriaus pasikartojantys skirtingo ilgio žodžiai;
- kiekvieno autoriaus pasikartojantys skirtingo ilgio skirtingi žodžiai;
- skirtingo ilgio sakiniai;
- žodžių ir raidžių skaičius sakinyje.

Pirmiausia, naudojantis programine įranga R buvo pasiruošti tekstai: atskirti žodžiai ir skyrybos ženklai, sakiniai išskaidyti žodžiais ir raidėmis, tuomet suskaičiuoti žodžiai ir raidės,

pasikartojančių raidžių ir žodžių dažniai. Žemiau pateikiama programos dalis, skirta skaičiuoti raides skirtinguose sakiniuose:

```
sakV=as.character(Sakiniai)
t1 = sub("^ +", "", sakV)
t2 = gsub(" +", " ", t1)
sakiniai = sub(" +$", "", t2)
p="a"
s=gsub(p,"",sakiniai[1])
num1a= nchar(sakiniai[1])-nchar(s)
num1a
```

Skaičiavimams atlikti ir hipotezėms tikrinti naudojamos teorinėje dalyje pateiktos formulės bei SPSS programinis paketas.

3 lentelė. *Žodžių skaičius*

	Kiekvieno autoriaus pavartotų žodžių skaičius	Kiekvieno autoriaus pavartotų skirtingų žodžių skaičius
1 autorius	1540	908
2 autorius	1583	954
3 autorius	1248	745
4 autorius	1564	960
5 autorius	1820	1277
Bendrai visi autoriai	7755	3879
Vidurkis	1551	968,8

Iš 3 lentelėje pateiktų duomenų galime matyti, kad daugiausiai žodžių tiriamose imtyse pavartojo 5 autorius, mažiausiai – 3 autorius. Pavartotų skirtingų žodžių skaičius taip pat didžiausias 5 autoriaus ir mažiausias 3 autoriaus imtyse. Bendrai pavartotų skirtingų žodžių, kai neskaičiuojame to paties žodžio, besikartojančio kelis kartus, skaičius yra 1,9992 kartų mažesnis nei iš viso pavartotų žodžių skaičius. Vidutiniškai visi autoriai pavartojo po 1551 žodį, o skirtingų žodžių – po 968,8.

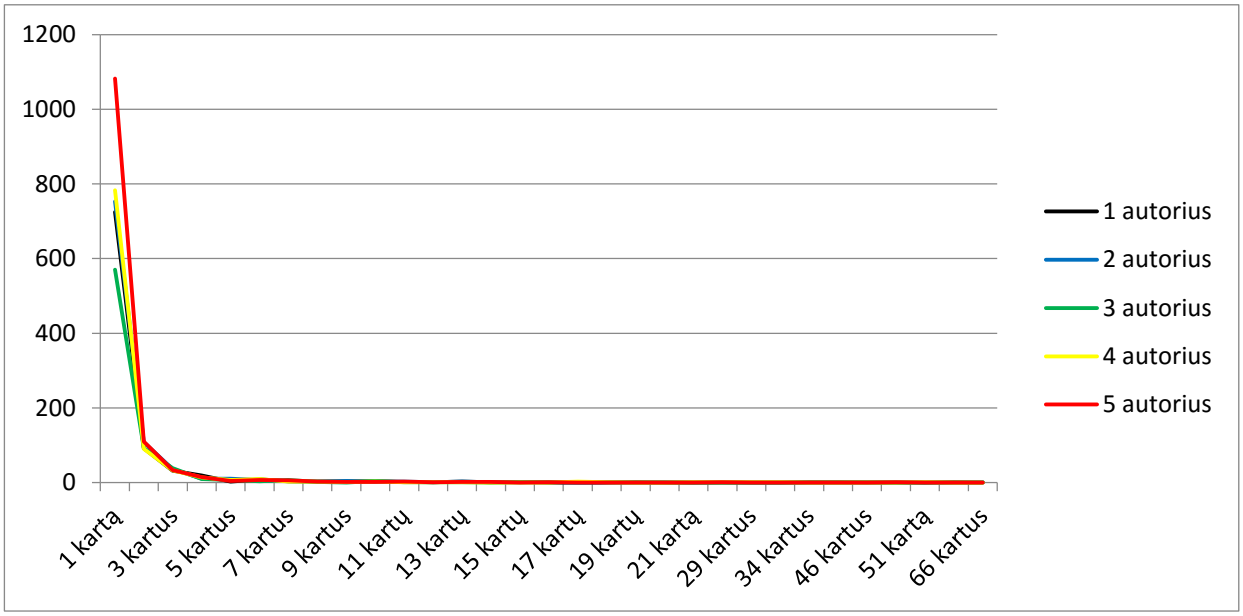
Norint palyginti autorius ir pažiūrėti, kaip tarpusavyje koreliuoja nagrinėjamų autorių vartojami žodžiai ir skirtingo ilgio žodžiai buvo apskaičiuota Pirsono koreliacija. Gauti rezultatai pateikti 4 lentelėje.

4 lentelė. *Vartojamų žodžių koreliacija*

		Autorių_žodžių_skaičius	Autorių_skirtingų_žodžių_skaičius
Autorių_žodžių_skaičius	Pearson Correlation	1	,960**
	Sig. (2-tailed)		,010
	N	5	5
Autorių_skirtingų_žodžių_skaičius	Pearson Correlation	,960**	1
	Sig. (2-tailed)	,010	
	N	5	5

Matome, kad yra labai stipri teigiama koreliacija ($r = 0{,}96$), vadinasi, vartojamų žodžių skaičius ir vartojamų skirtingų žodžių skaičius yra priklausomi, keičiantis pavartotų žodžių skaičiui tolygiai didėja arba mažėja ir pavartotų skirtingų žodžių skaičius. Ji yra statistiškai reikšminga, nes gauta p reikšmė (Sig. (2-tailed)) yra mažesnė už pasirinktą reikšmingumo lygmenį: $p = 0{,}05 > 0{,}01 = \alpha$.

Kiekvieno autoriaus pasikartojančių žodžių duomenys pateikti 1 priede ir 3 paveikslėlyje. Iš pateiktų rezultatų matome, kad daugiausiai autoriai žodžius vartoja tik po vieną kartą, pakankamai dažnai pasitaiko, kai tą patį žodį kartoja 2 kartus, rečiau – 3, 4, 5 kartus ir labai retai, kai kartojasi nuo 6 iki 66 kartų.



3 pav. *Kiekvieno autoriaus pasikartojančių žodžių skaičiaus grafikas*

1 autorius dažniausiai (20 ir daugiau kartų) vartojo tokius žodžius: *Timis* – 34 kartus, *į* – 36 kartus, *ir* – 56 kartus; 2 autorius: *kad* – 20 kartų, *tai* – 22 kartus, *ir* – 66 kartus; 3 autorius: *ir* – 46 kartus; 4 autorius: *buvo* – 21 kartą, *ji* – 22 kartus, *Beatričė* – 29 kartus, *į* – 30 kartų, *ir* – 51 kartą; 5 autorius: *į* – 22 kartus, *ir* – 48 kartus. Bendrai tarp visų autorių daugiausiai kartų pasikartojo žodis *ir*

(267 kartus), taip pat pakankamai nemažai kartojosi žodžiai *į* (109 kartus) ir *kad (*78 kartus). Dažniausiai vartojami žodžiai pateikti 2 priede.

Pasinaudojus 1 priede pateiktais duomenimis apie bendrą pasikartojančių žodžių skaičių buvo apskaičiuoti vidurkiai. 5 lentelėje pateikti rezultatai parodo, kiek žodžių vidutiniškai kartojosi 1, 2, 3, ..., 66 kartus.

5 lentelė. *Pasikartojančių žodžių vidurkiai*

Kartai	Vidurkis	**Kartai**	Vidurkis	**Kartai**	Vidurkis
1 kartą	782,6	**12 kartų**	1	**22 kartus**	0,6
2 kartus	98,4	**13 kartų**	2	**29 kartus**	0,2
3 kartus	34,8	**14 kartų**	0,8	**30 kartų**	0,2
4 kartus	13,8	**15 kartų**	0,8	**34 kartus**	0,2
5 kartus	6,6	**16 kartų**	0,6	**36 kartus**	0,2
6 kartus	7,4	**17 kartų**	0,8	**46 kartus**	0,2
7 kartus	4,8	**18 kartų**	0,6	**48 kartus**	0,2
8 kartus	3	**19 kartų**	0,2	**51 kartą**	0,2
9 kartus	2,6	**20 kartų**	0,2	**56 kartus**	0,2
10 kartų	3,2	**21 kartą**	0,2	**66 kartus**	0,2
11 kartų	2				

Iš 5 lentelėje pateiktų rezultatų matome, kad nagrinėjami autoriai vidutiniškai daugiausiai žodžių pavartojo po vieną kartą po 782,6 skirtingų (savo kūryboje) žodžių, taip pat nemažai žodžių po du kartus – po 98,4 skirtingų žodžių, kai kartų skaičius padidėja iki 14–66 kartų, pavartotų žodžių skaičių vidurkiai nesiekia vieneto ir svyruoja tarp 0,2 ir 0,8.

Apskaičiuotos kiekvieno autoriaus žodžių skaičiaus statistinės charakteristikos pateiktos 6 lentelėje.

6 lentelė. *Kiekvieno autoriaus žodžių skaičiaus statistinės charakteristikos*

	N	Vidurkis	Mediana	Moda
1 autorius	1540	5,655	6	6
2 autorius	1583	5,647	5	3
3 autorius	1248	5,641	5	3
4 autorius	1564	5,850	6	3
5 autorius	1820	6,153	6	6

Analizuodami 6 lentelę matome, kad 1 autorius vartojo vidutiniškai 5,655 raidžių ilgio žodžius, kiti autoriai – panašiai. Galime daryti išvadą, kad visų autorių vidutiniškai vartojamų žodžių ilgis skiriasi nežymiai (skirtumas tarp trumpiausio ir ilgiausio vidutiniškai vartojamo ilgio

žodžio – 0,512). Iš gautų rezultatų taip pat galima pastebėti, kad 1 ir 5 autoriai daugiausiai vartoja 6 raidžių žodžius, 2, 3 ir 4 autoriai – 3 raidžių ilgio žodžius.

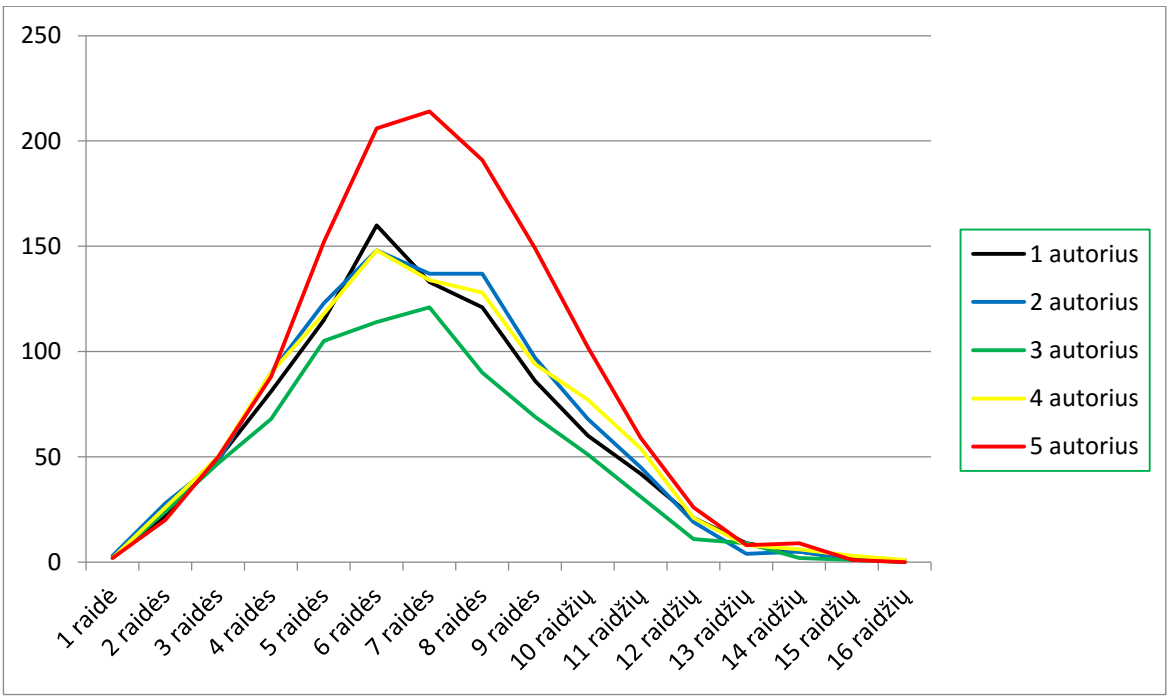
3.1. Žodžių ilgio statistinė analizė

Kiekvieno autoriaus pasikartojančių skirtingo ilgio skirtingų žodžių duomenys pateikti 7 lentelėje.

7 lentelė. *Kiekvieno autoriaus skirtingo ilgio skirtingų žodžių skaičius*

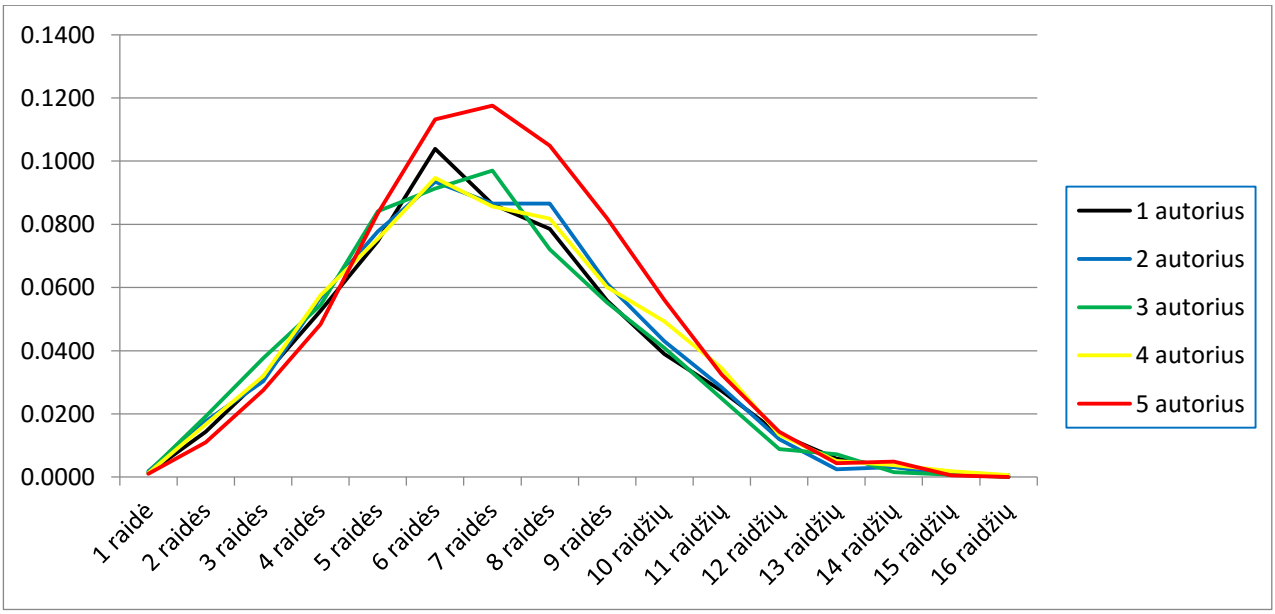
	1 autorius	2 autorius	3 autorius	4 autorius	5 autorius	Vidutiniškai
1 raidė	2	3	2	2	2	2,2
2 raidės	22	28	24	26	20	24
3 raidės	49	48	47	50	50	8,8
4 raidės	81	90	68	90	88	83,4
5 raidės	115	123	105	118	152	122,6
6 raidės	160	148	114	148	206	155,2
7 raidės	133	137	121	134	214	147,8
8 raidės	121	137	90	128	191	133,4
9 raidės	86	97	69	94	149	99
10 raidžių	60	68	51	77	102	71,6
11 raidžių	42	45	31	54	59	46,2
12 raidžių	21	19	11	21	26	19,6
13 raidžių	9	4	9	8	8	7,6
14 raidžių	5	5	2	6	9	5,4
15 raidžių	2	1	1	3	1	1,6
16 raidžių	0	1	0	1	0	0,4

Iš 7 lentelėje pateiktų rezultatų matome, kad 1, 2 ir 4 autorius daugiausiai vartoja 6 raidžių ilgio skirtingus žodžius, 3 ir 5 autorius 7 raidžių ilgio skirtingus žodžius. Mažiausiai skirtingo ilgio žodžių 1 autorius vartoja su 1 ir 15 raidžių, 2 autorius su 15 ir 16 raidžių, 3 ir 5 autorius su 15 raidžių ir 4 autorius su 16 raidžių. Analizuodami bendrai visų autorių vidutiniškai vartojamus skirtingo ilgio skirtingus žodžius galime teigti, kad daugiausiai autoriai pavartojo žodžių su 6 raidėmis, taip pat pakankamai nemažai žodžių su 7 ir 8 raidėmis, mažiausiai 16 – raidžių ilgio žodžių.



4 pav. *Kiekvieno autoriaus skirtingo ilgio skirtingų žodžių skaičiaus grafikas*

Iš grafiko (žr. 4 pav.) galime pastebėti, kad visi autoriai labai panašiai rašo, retai naudoja žodžius su mažai (1–4 raidėmis) ir daug raidžių (10–16 raidžių), daugiausia vartoja žodžius su 5–9 raidėmis, ypatingai 5 autorius.



5 pav. *Kiekvieno autoriaus skirtingo ilgio skirtingų žodžių skaičiaus santykinių dažnių grafikas*

Santykinių dažnių grafikas parodo, kaip dažnai kartojasi žodžiai su tam tikru raidžių skaičiumi. Iš pateikto grafiko (žr. 5 pav.) pastebime, kad santykinai dažnai kartojasi žodžiai su 6–8 raidėmis.

Visų tyrime analizuojamų autorių imčių bendri skirtingo ilgio skirtingų žodžių duomenys pateikti 8 lentelėje.

8 lentelė. *Skirtingo ilgio skirtingų žodžių skaičius*

Raidžių skaičius žodyje	Skirtingų žodžių skaičius	**Raidžių skaičius žodyje**	Skirtingų žodžių skaičius	**Raidžių skaičius žodyje**	Skirtingų žodžių skaičius
1 raidė	3	**7 raidės**	650	**13 raidžių**	38
2 raidės	38	**8 raidės**	596	**14 raidžių**	27
3 raidės	80	**9 raidės**	464	**15 raidžių**	8
4 raidės	248	**10 raidžių**	343	**16 raidžių**	2
5 raidės	444	**11 raidžių**	220		
6 raidės	623	**12 raidžių**	95		

Iš 8 lentelėje pateiktų rezultatų matome, kad daugiausiai skirtingo ilgio skirtingų žodžių yra 7 raidžių ilgio, šiek tiek mažiau – 6 raidžių, o mažiausiai – 16 ir 1 raidės ilgio žodžių. Taip pat galime pastebėti, kad autoriai mažai vartoja trumpų žodžių, kurių ilgis 1–3 raidės, ir ilgų žodžių, kurių ilgis 12–16 raidžių, o daugiausiai vartoja žodžius su 6–9 raidėmis.

9 lentelė. *Skirtingo ilgio skirtingų žodžių statistinės charakteristikos*

	Vidurkis	Mediana	Moda	Dispersija
1 autorius	6,890	7	6	6,222
2 autorius	6,866	7	6	6,165
3 autorius	6,715	7	7	6,172
4 autorius	6,995	7	6	6,666
5 autorius	7,152	7	7	5,518

Pagal 9 lentelėje pateiktus rezultatus galima daryti išvadą, kad visų autorių vidutiniškai vartojamų skirtingų žodžių ilgis skiriasi nežymiai (1 autorius vidutiniškai vartojo 6,890 raidžių ilgio skirtingus žodžius, 2 autorius – 6,866 ir t. t., skirtumas tarp trumpiausio ir ilgiausio vidutiniškai vartojamo žodžio – 0,437). Iš gautų rezultatų taip pat galima teigti, kad 1, 2 ir 4 autoriai daugiausiai vartoja 6 raidžių skirtingus žodžius, 3 ir 5 autoriai – 7 raidžių ilgio skirtingus žodžius.

Norint palyginti autorius ir pažiūrėti, kaip tarpusavyje koreliuoja skirtingų autorių porų vartojami skirtingo ilgio skirtingi žodžiai, buvo apskaičiuota Pirsono koreliacija. Gauti rezultatai pateikti 10 lentelėje.

10 lentelė. *Skirtingo ilgio skirtingų žodžių koreliacija*

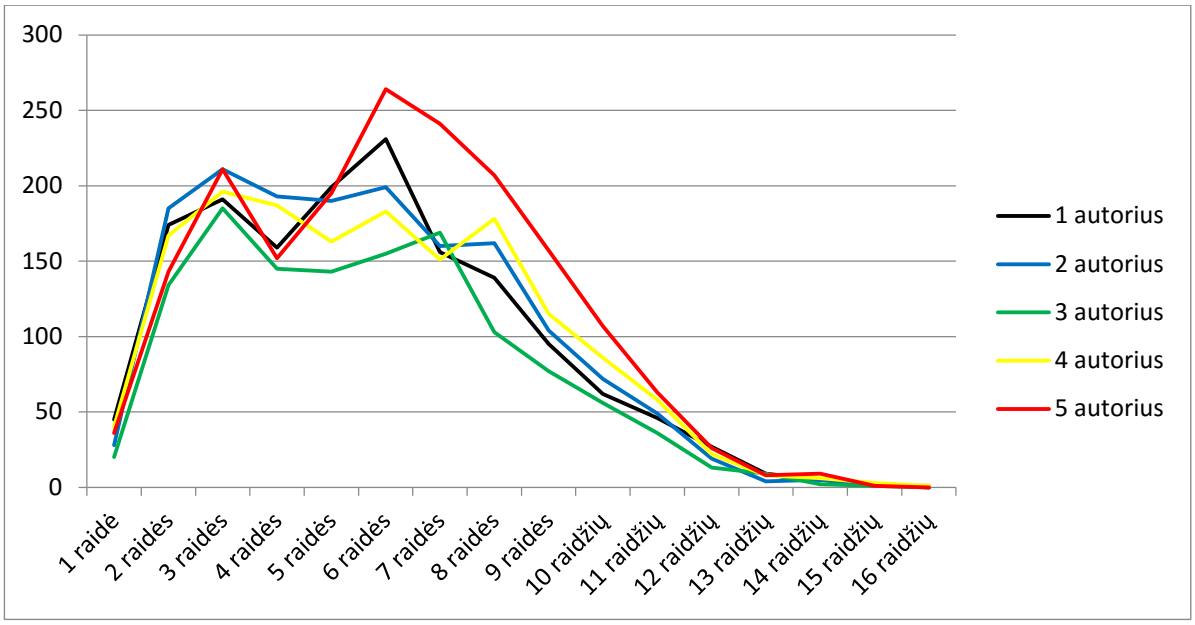
		autorius_1	autorius_2	autorius_3	autorius_4	autorius_5
autorius_1	Pearson Correlation	1	,993**	,986**	,993**	,979**
	Sig. (2-tailed)		,000	,000	,000	,000
	N	16	16	16	16	16
autorius_2	Pearson Correlation	,993**	1	,986**	,997**	,984**
	Sig. (2-tailed)	,000		,000	,000	,000
	N	16	16	16	16	16
autorius_3	Pearson Correlation	,986**	,986**	1	,984**	,969**
	Sig. (2-tailed)	,000	,000		,000	,000
	N	16	16	16	16	16
autorius_4	Pearson Correlation	,993**	,997**	,984**	1	,982**
	Sig. (2-tailed)	,000	,000	,000		,000
	N	16	16	16	16	16
autorius_5	Pearson Correlation	,979**	,984**	,969**	,982**	1
	Sig. (2-tailed)	,000	,000	,000	,000	
	N	16	16	16	16	16

Matome, kad tarp visų autorių yra labai stipri teigiama koreliacija ($r = 0{,}997$, $r = 0{,}993$, $r = 0{,}986$, $r = 0{,}982$, $r = 0{,}979$, $r = 0{,}969$), vadinasi, autoriai rašo panašiai, keičiantis žodžių ilgiui visų autorių vartojamų skirtingų žodžių skaičius didėja arba mažėja tolygiai. Koreliacija yra statistiškai reikšminga, nes $p = 0{,}000 < 0{,}01 = \alpha$.

11 lentelė. *Kiekvieno autoriaus skirtingo ilgio žodžių skaičius*

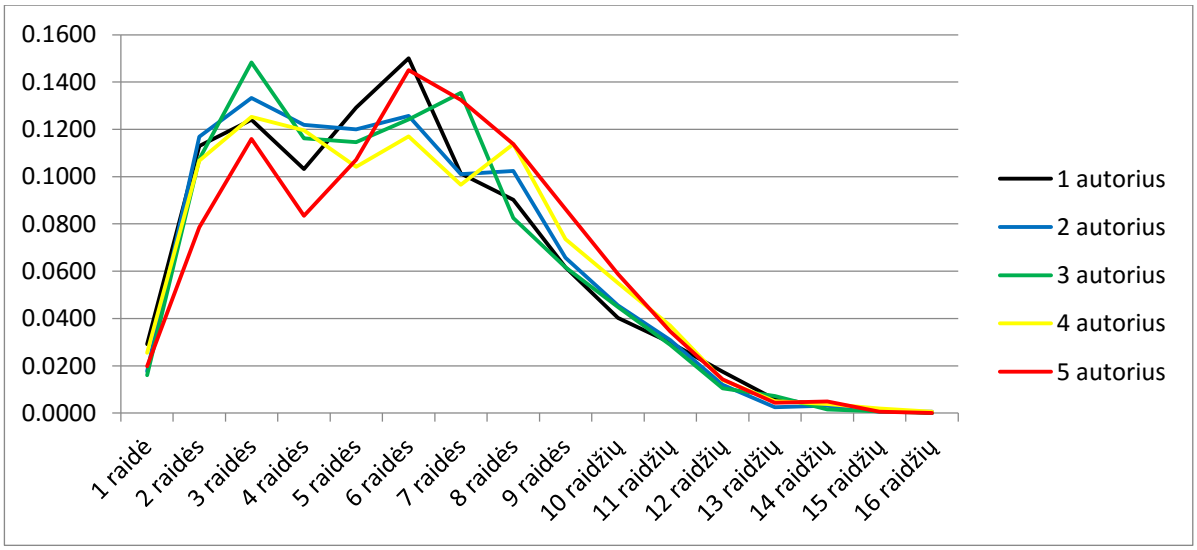
	1 autorius	2 autorius	3 autorius	4 autorius	5 autorius	Vidutiniškai
1 raidė	45	28	20	40	36	33,8
2 raidės	174	185	134	167	143	160,6
3 raidės	191	211	185	196	211	198,8
4 raidės	159	193	145	187	152	167,2
5 raidės	199	190	143	163	195	178
6 raidės	231	199	155	183	264	206,4
7 raidės	156	160	169	151	241	175,4
8 raidės	139	162	103	178	207	157,8
9 raidės	95	104	77	115	157	109,6
10 raidžių	62	72	56	86	107	76,6
11 raidžių	46	49	36	58	63	50,4
12 raidžių	27	19	13	22	26	21,4
13 raidžių	9	4	9	8	8	7,6
14 raidžių	5	5	2	6	9	5,4
15 raidžių	2	1	1	3	1	1,6
16 raidžių	0	1	0	1	0	0,4

11 lentelėje pateikti kiekvieno autoriaus pasikartojančių skirtingo ilgio žodžių duomenys. Matome, 1 ir 5 autorius daugiausiai vartoja 6 raidžių ilgio žodžių, 2, 3 ir 4 autorius – 3 raidžių ilgio žodžių. Mažiausiai 1, 3 ir 5 autorius vartoja žodžių iš 15 raidžių, 4 autorius – iš 16 raidžių, 2 autorius mažiausiai vartoja žodžių iš 15 ir 16 raidžių. Išanalizavus bendrai visų autorių vidutiniškai vartojamus skirtingo ilgio žodžius galima teigti, kad daugiausiai autoriai pavartojo žodžių su 6 raidėmis, taip pat pakankamai nemažai žodžių su 3 raidėmis, mažiausiai – 16 raidžių ilgio žodžių.



6 pav. *Kiekvieno autoriaus skirtingo ilgio žodžių skaičiaus grafikas*

Iš grafiko (žr. 6 pav.) galima pastebėti, kad visi autoriai labai panašiai rašo, retai naudoja žodžius su mažai (1–2) ir daug (10–16) raidžių, daugiausia vartoja žodžius su 3–9 raidėmis. Taip pat galima pastebėti, kad santykinai dažnai kartojasi žodžiai su 3 raidėmis ir su 6–8 raidėmis (žr. 7 pav.).



7 pav. *Kiekvieno autoriaus skirtingo ilgio žodžių skaičiaus santykinių dažnių grafikas*

Visų tyrime analizuojamų autorių imčių bendri pasikartojančių skirtingo ilgio žodžių duomenys pateikti 12 lentelėje. Matome, kad daugiausiai skirtingo ilgio žodžių yra 6 raidžių ilgio, šiek tiek mažiau – 5 ir 7 raidžių, o mažiausiai – 16 ir 15 raidžių ilgio žodžių. Galima teigti, kad autoriai mažai vartoja ilgų žodžių, kurių ilgis 12–16 raidžių, ir labai trumpų žodžių, kurių ilgis 1 raidė, o daugiausiai vartoja žodžius su 2–8 raidėmis.

12 lentelė. *Skirtingo ilgio žodžių skaičius*

Raidžių skaičius žodyje	Kiek kartų pavartota	**Raidžių skaičius žodyje**	Kiek kartų pavartota	**Raidžių skaičius žodyje**	Kiek kartų pavartota
1 raidė	169	**7 raidės**	877	**13 raidžių**	38
2 raidės	803	**8 raidės**	789	**14 raidžių**	27
3 raidės	994	**9 raidės**	548	**15 raidžių**	8
4 raidės	836	**10 raidžių**	383	**16 raidžių**	2
5 raidės	890	**11 raidžių**	252		
6 raidės	1032	**12 raidžių**	107		

13 lentelėje pateikti skirtingų autorių porų vartojamų skirtingo ilgio žodžių Pirsono koreliacijos rezultatai. Matome, kad tarp visų autorių yra labai stipri teigiama koreliacija ($r = 0{,}991,\ r = 0{,}982,\ r = 0{,}978,\ r = 0{,}965,\ r = 0{,}963,\ r = 0{,}961,\ r = 0{,}942,\ r = 0{,}940,\ r = 0{,}933$), vadinasi, nagrinėjami autoriai panašiai vartoja skirtingo ilgio žodžius. Koreliacija yra statistiškai reikšminga ($p = 0{,}000 < 0{,}01 = \alpha$).

13 lentelė. *Skirtingo ilgio žodžių koreliacija*

		autorius_1	autorius_2	autorius_3	autorius_4	autorius_5
autorius_1	Pearson Correlation	1	,982**	,965**	,963**	,942**
	Sig. (2-tailed)		,000	,000	,000	,000
	N	16	16	16	16	16
autorius_2	Pearson Correlation	,982**	1	,978**	,991**	,933**
	Sig. (2-tailed)	,000		,000	,000	,000
	N	16	16	16	16	16
autorius_3	Pearson Correlation	,965**	,978**	1	,961**	,940**
	Sig. (2-tailed)	,000	,000		,000	,000
	N	16	16	16	16	16
autorius_4	Pearson Correlation	,963**	,991**	,961**	1	,942**
	Sig. (2-tailed)	,000	,000	,000		,000
	N	16	16	16	16	16
autorius_5	Pearson Correlation	,942**	,933**	,940**	,942**	1
	Sig. (2-tailed)	,000	,000	,000	,000	
	N	16	16	16	16	16

3.1.1. Normalumo tikrinimas

Norėdami įsitikinti, ar skirtingo ilgio žodžių duomenys yra pasiskirstę pagal normalųjį skirstinį, taikysime Kolmogorovo-Smirnovo ir Šapiro-Vilko testus (naudosimės SPSS programų paketu). Tikrinsime hipotezę $H_0: X \sim N(\bar{x}, s^2)$ su alternatyva $H_1: X \nsim N(\bar{x}, s^2)$, kai $\alpha = 0{,}01$; čia X – 1 autoriaus, 2 autoriaus, 3 autoriaus, 4 autoriaus, 5 autoriaus skirtingo ilgio žodžių skaičius.

14 lentelė. *Normalumo tikrinimas*

	Kolmogorov-Smirnov[a]			Shapiro-Wilk		
	Statistic	df	Sig.	Statistic	df	Sig.
autorius_1	,170	16	,200*	,897	16	,072
autorius_2	,203	16	,076	,842	16	,011
autorius_3	,176	16	,199	,875	16	,033
autorius_4	,193	16	,115	,856	16	,017
autorius_5	,171	16	,200*	,901	16	,082

Iš 14 lentelėje pateiktų rezultatų matome, kad abiejų testų p reikšmės (Sig.) yra didesnės, nei pasirinktas reikšmingumo lygmuo ($\alpha = 0{,}01$). Todėl nulinės hipotezės nėra pagrindo atmesti ir tolesniuose tyrimuose galime taikyti metodus, tinkančius normaliajam skirstiniui.

3.1.2. Hipotezės apie autorių porų vidurkių lygybę tikrinimas

Kiekvieno autoriaus pavartotiems skirtingo ilgio žodžiams tikrinama hipotezė apie autorių porų vidurkių lygybę:

$$\begin{cases} H_0: \text{pirmo ir antro autorių pavartotų skirtingo ilgio žodžių vidurkiai lygūs,} \\ H_1: \text{pirmo ir antro autorių pavartotų skirtingo ilgio žodžių vidurkiai nėra lygūs.} \end{cases}$$

Pagal 11 lentelėje pateiktus duomenis apskaičiuoti vidurkiai ir dispersijos pateikti 15 lentelėje.

15 lentelė. *Kiekvieno autoriaus pavartotų žodžių ilgio vidurkis ir dispersija*

	Maksimalus žodžio ilgis	Vidurkis	Dispersija
1 autorius	15	102,6667	6314,238
2 autorius	16	98,9375	7100,729
3 autorius	15	83,2000	4573,743
4 autorius	16	97,7500	5941,000
5 autorius	15	121,3333	8434,524

Tegu x yra 1 autoriaus duomenų imtis, y – 2 autoriaus duomenų imtis.

$n = 15$, $m = 16$, $\bar{x} = 102{,}6667$, $\bar{y} = 98{,}9375$, $s_x^2 = 6314{,}238$, $s_y^2 = 7100{,}729$.

Reikšmingumo lygmuo $\alpha = 0{,}05$.

Pirmiausia tikrinama hipotezė apie dviejų dispersijų lygybę:

$$\begin{cases} H_0^1: \sigma_X^2 = \sigma_Y^2, \\ H_1^1: \sigma_X^2 \neq \sigma_Y^2. \end{cases}$$

Remiantis (1) formule sudaroma kritinė sritis:

$$F = \frac{6314{,}238}{7100{,}729} = 0{,}889.$$

Hipotezė H_0 neatmetama, jeigu $F_{1-\frac{\alpha}{2}}(n-1, m-1) \leq F \leq F_{\frac{\alpha}{2}}(n-1, m-1)$.

Kadangi Fišerio skirstinio kvantilių lentelėse nebuvo reikalingų reikšmių, jos buvo apskaičiuotos remiantis [8] pateiktomis formulėmis:

$$F_{1-p}(k_1, k_2) = \frac{1}{F_p(k_2, k_1)}, \tag{11}$$

$$F_p(k_1, k_2) \approx \frac{k_2}{k_2 - 2}\sqrt{\frac{2(k_1 + k_2 - 2)}{k_1(k_2 - 4)}}u_p + \frac{k_2}{k_2 - 2}. \tag{12}$$

$k_1 = 14, k_2 = 15, p = \frac{\alpha}{2} = 0{,}025, u_p = u_{0{,}025} = 1{,}959.$

$$F_p(k_1, k_2) \approx \frac{k_2}{k_2 - 2}\sqrt{\frac{2(k_1 + k_2 - 2)}{k_1(k_2 - 4)}}u_p + \frac{k_2}{k_2 - 2} = 0{,}6833 \cdot 1{,}959 + 1{,}1538 = 2{,}4923;$$

$$F_p(k_2, k_1) \approx \frac{k_1}{k_1 - 2}\sqrt{\frac{2(k_2 + k_1 - 2)}{k_2(k_1 - 4)}}u_p + \frac{k_1}{k_1 - 2} = 0{,}7 \cdot 1{,}959 + 1{,}1667 = 2{,}538;$$

$$F_{1-p}(k_1, k_2) = \frac{1}{F_p(k_2, k_1)} = \frac{1}{2{,}538} = 0{,}394.$$

$F_{1-\frac{\alpha}{2}}(n-1, m-1) = 0{,}394;$

$F_{\frac{\alpha}{2}}(n-1, m-1) = 2{,}4923.$

Kadangi $F_{1-\frac{\alpha}{2}}(n-1, m-1) = 0{,}394 < F = 0{,}889 < F_{\frac{\alpha}{2}}(n-1, m-1) = 2{,}4923$, tai H_0^1 priimama. Taigi, galima teigti, kad pirmo ir antro autorių pavartotų skirtingo ilgio žodžių dispersijos statistiškai reikšmingai nesiskiria.

Kadangi dipersijos statistiškai reikšmingai nesiskiria, galima naudoti Stjudento kriterijų ir tikrinti hipotezę apie pirmo ir antro autorių pavartotų skirtingo ilgio žodžių vidurkių lygybę (H_0), kai populiacijų dispersijos lygios.

Sudaroma kritinė sritis

$$t = \frac{\bar{x} - \bar{y}}{\sqrt{(n-1)s_x^2 + (m-1)s_y^2}}\sqrt{\frac{nm(n+m-2)}{n+m}}:$$

$$t = \frac{102{,}6667 - 98{,}9375}{\sqrt{(14 \cdot 6314{,}238) + (15 \cdot 7100{,}7292)}} \sqrt{\frac{15 \cdot 16(15 + 16 - 2)}{15 + 16}} =$$

$$= \frac{3{,}7292}{\sqrt{194910{,}27}} \sqrt{\frac{6960}{31}} = 0{,}0084 \cdot 14{,}9839 = 0{,}1259.$$

Kadangi $|t| = 0{,}1259 < 2{,}042 = t_{0{,}025}(29)$, tai H_0 priimama. Taigi, galima teigti, kad pirmo ir antro autorių pavartotų žodžių vidurkiai statistiškai reikšmingai nesiskiria.

Tos pačios hipotezės tikrinimo SPSS programų paketu rezultatai pateikti 16 lentelėje.

16 lentelė. *Stjudento kriterijus dviem nepriklausomoms imtims*

		Levene's Test for Equality of Variances		t-test for Equality of Means						
									95% Confidence Interval of the Difference	
		F	Sig.	t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	Lower	Upper
žodžių_ skaičius	Equal variances assumed	,294	,592	,127	29	,900	3,72917	29,46413	-56,53174	63,99007
	Equal variances not assumed			,127	28,998	,900	3,72917	29,40654	-56,41414	63,87247

Iš 16 lentelės matyti, kad gauta p reikšmė (Sig.) didesnė už pasirinktą reikšmingumo lygmenį ($p = 0{,}592 > 0{,}05 = \alpha$), vadinasi, grupių dispersijos statistiškai reikšmingai nesiskiria ir tolimesniam tyrimui naudojama lentelės eilutė, atitinkanti lygių dispersijų atvejį (Equal Variances assumed). Gauta p reikšmė (Sig.(2-tailed)) didesnė už pasirinktą reikšmingumo lygmenį ($p = 0{,}900 > 0{,}05 = \alpha$), vadinasi, autorių vidurkiai statistiškai reikšmingai nesiskiria, todėl hipotezė H_0 neatmetama – pirmo ir antro autorių pavartotų skirtingo ilgio žodžių vidurkiai statistiškai reikšmingai nesiskiria.

Iš 17 lentelėje pateiktų rezultatų matome, kad visos gautos p reikšmės (Sig.) didesnės už pasirinktą reikšmingumo lygmenį ($\alpha = 0{,}05$), vadinasi, grupių dispersijos statistiškai reikšmingai nesiskiria. Lygių dispersijų atveju visos gautos p reikšmės (Sig.(2-tailed)) yra didesnės už pasirinktą reikšmingumo lygmenį ($\alpha = 0{,}05$), vadinasi, autorių vidurkiai statistiškai reikšmingai nesiskiria, todėl visos hipotezės H_0 neatmetamos, t. y. visų autorių porų pavartotų skirtingo ilgio žodžių vidurkiai statistiškai reikšmingai nesiskiria.

17 lentelė. *Bendra visų autorių porų rezultatų lentelė*

Autorių pora	p reikšmė (Sig.)	p reikšmė (Sig. (2-tailed))	H_0 hipotezė
1 ir 2 autoriai	0,592	0,900	priimama
1 ir 3 autoriai	0,325	0,476	priimama
1 ir 4 autoriai	0,928	0,862	priimama
1 ir 5 autoriai	0,486	0,556	priimama
2 ir 3 autoriai	0,112	0,572	priimama
2 ir 4 autoriai	0,501	0,967	priimama
2 ir 5 autoriai	0,790	0,485	priimama
3 ir 4 autoriai	0,331	0,582	priimama
3 ir 5 autoriai	0,121	0,206	priimama
4 ir 5 autoriai	0,413	0,444	priimama

3.1.3. Požymių nepriklausomumo tikrinimas

Remiantis anksčiau gautais rezultatais buvo sudarytos trys grupės: 1–4 raidžių, 5–9 raidžių ir 10–16 raidžių ilgio žodžiai.

Po kiek kiekvieno autoriaus žodžių patenka į išskirtas grupes pateikta 18 lentelėje.

18 lentelė. *Sugrupuotų skirtingo ilgio žodžių skaičius*

	1-4 raidžių ilgio	5-9 raidžių ilgio	10-16 raidžių ilgio	
1 autorius	569	820	151	1540
2 autorius	617	815	151	1583
3 autorius	484	647	117	1248
4 autorius	590	790	184	1564
5 autorius	542	1064	214	1820
	2802	4136	817	7755

Tikrinama statistinė hipotezė:

$$\begin{cases} H_0: p_{ij} = p_i q_j, & \mathrm{i} = 1, \dots, 5, \mathrm{j} = 1, \dots, 3; \\ H_1: p_{ij} \neq p_i q_j, & \text{bent vienai porai } (\mathrm{i}, \mathrm{j}); \end{cases}$$

čia i – autoriaus numeris, j – žodžių ilgio grupės numeris.

Pagal (4) formulę apskaičiuojama kriterijaus statistika:

$$\chi^2 = \sum_{i=1}^{5}\sum_{j=1}^{3} \frac{\left(O_{ij} - e_{ij}\right)^2}{e_{ij}} = \sum_{i=1}^{5}\sum_{j=1}^{3} \frac{\left(O_{ij} - \frac{n_{i\cdot} n_{\cdot j}}{n}\right)^2}{\frac{n_{i\cdot} n_{\cdot j}}{n}} = 49{,}2134.$$

Tarpiniai – tikėtinųjų dažnių e_{ij} – rezultatai pateikti 19 lentelėje.

19 lentelė. *Tikėtinieji dažniai*

e_{11}	556,4255	e_{21}	571,9621	e_{31}	450,9215	e_{41}	565,0971	e_{51}	657,5938
e_{12}	821,3333	e_{22}	844,2667	e_{32}	665,6	e_{42}	834,1333	e_{52}	970,6667
e_{13}	162,2411	e_{23}	166,7712	e_{33}	131,4785	e_{43}	164,7696	e_{53}	191,7395

Kai $\alpha = 0{,}05$, $\chi^2_{0{,}05}((5-1)(3-1)) = \chi^2_{0{,}05}(8) = 15{,}507$.

Kadangi $\chi^2 = 49{,}2134 > 15{,}507 = \chi^2_{0{,}05}(8)$, tai hipotezė, apie požymių nepriklausomumą atmetama. Taigi, autoriai ir žodžių ilgiai susiję.

Ta pati hipotezė buvo patikrinta su SPSS programų paketu.

$$\begin{cases} H_0\text{: Žodžių ilgis nepriklauso nuo autoriaus,} \\ H_1\text{: Žodžių ilgis priklauso nuo autoriaus.} \end{cases}$$

20 lentelė. χ^2 *skaičiavimo rezultatai*

	Value	df	Asymp. Sig. (2-sided)
Pearson Chi-Square	49.213[a]	8	.000
Likelihood Ratio	50.094	8	.000
Linear-by-Linear Association	21.574	1	.000
N of Valid Cases	7755		

Pagal gautus rezultatus (žr. 20 lent.) matome, kad $p = 0{,}000$ yra mažesnė nei pasirinktas reikšmingumo lygmuo $\alpha = 0{,}05$, todėl nulinė hipotezė H_0 atmetama. Galima teigti, kad tiriamieji požymiai yra priklausomi, tai reiškia, kad žodžių ilgis priklauso nuo autoriaus.

3.2. Sakinių ilgio statistinė analizė

Kokio ilgio sakinius vartoja kiekvienas autorius nagrinėjamose imtyse, pateikta 21 lentelėje. Galima pastebėti, kad daugiausiai 1 ir 3 autoriaus sakinių – iš 3 žodžių, 2 ir 4 autoriaus – iš 6 žodžių, o 5 autoriaus – iš 7 žodžių. Mažiausiai 1 autorius rašo 20, 21, 22 ir 46 žodžių sakinių, 2 autorius – 16, 25, 26, 28 ir 46 žodžių sakinių, 3 autorius – 16, 17, 22, 23 ir 28 žodžių sakinių, 4 autorius – 18, 23, 24, 28 ir 35 žodžių sakinių, 5 autorius – 21, 29 ir 30 žodžių sakinių. Galima pastebėti, kad didėjant žodžių skaičiui sakiniuose sakinių skaičius mažėja. Išanalizavus bendrai visų autorių vidutiniškai rašomų sakinių ilgį galima teigti, kad daugiausiai autoriai rašo 3 žodžių sakinių, taip pat pakankamai nemažai 4, 5 ir 6 žodžių ilgio sakinių, mažiausiai – labai ilgų sakinių, kuriuose 29 ir 35 žodžiai.

21 lentelė. *Kiekvieno autoriaus vartojamų skirtingo ilgio sakinių skaičius*

Sakinio ilgis	1 autorius	2 autorius	3 autorius	4 autorius	5 autorius	Vidutiniškai
1	3	12	17	1	2	7
2	14	20	26	12	5	15,4
3	32	20	27	20	13	22,4
4	29	15	25	22	18	21,8
5	21	20	20	21	14	19,2
6	14	21	16	23	19	18,6
7	13	14	17	19	20	16,6
8	9	7	9	17	19	12,2
9	6	9	8	11	12	9,2
10	12	17	5	9	8	10,2
11	10	6	2	10	14	8,4
12	3	9	2	5	11	6
13	6	2	6	5	12	6,2
14	3	2	3	5	7	4
15	4	6	3	7	6	5,2
16	4	1	1	0	3	1,8
17	3	2	1	4	4	2,8
18	2	3	5	1	3	2,8
19	2	0	0	2	3	1,4
20	1	2	2	2	0	1,4
21	1	2	2	0	1	1,2
22	1	2	1	0	0	0,8
23	3	0	1	1	2	1,4
24	0	2	0	1	0	0,6
25	0	1	0	0	2	0,6
26	3	1	0	0	0	0,8
28	0	1	1	1	0	0,6
29	0	0	0	0	1	0,2
30	0	2	0	0	1	0,6
35	0	0	0	1	0	0,2
46	1	1	0	0	0	0,4

22 lentelė. *Žodžių skaičiaus sakinyje statistinės charakteristikos*

	Vidurkis	Mediana	Moda
1 autorius	7,7	6	3
2 autorius	7,92	6	6
3 autorius	6,24	5	3
4 autorius	7,82	7	6
5 autorius	9,1	8	7

Analizuodami 22 lentelėje pateiktus rezultatus matome, kad vidutiniškai trumpiausi yra 3 autoriaus sakiniai (6,24 žodžių ilgio), ilgiausi – 5 autoriaus, o 1, 2 ir 4 autorių vidutiniai sakinių ilgiai skiriasi nežymiai. Taip pat iš gautų rezultatų galima teigti, kad daugiausiai 1 ir 3 autoriaus sakinių sudaryti iš 3 žodžių, 2 ir 4 autoriaus – iš 6 žodžių, 5 autoriaus – iš 7 žodžių.

23 lentelė. *Raidžių skaičiaus sakinyje statistinės charakteristikos*

	Vidurkis	Mediana	Moda
1 autorius	43,525	33	23
2 autorius	44,69	34	28
3 autorius	35,2	25	14
4 autorius	45,735	38	33
5 autorius	55,99	49,5	45

Analogiškai tą patį galima pasakyti ir apie vidutinį raidžių skaičių nagrinėjamų autorių sakiniuose (žr. 23 lent.). Taip pat matome, kad trumpiausius sakinius rašantis 3 autorius sakiniuose dažniausiai vartojo 14 raidžių, ilgiausius sakinius rašantis 5 autorius dažniausiai vartojo 45 raides.

Norint palyginti autorius ir pažiūrėti, kaip tarpusavyje koreliuoja skirtingų autorių porų rašomi skirtingo ilgio sakiniai (žodžiais), buvo apskaičiuotas Spirmeno koreliacijos koeficientas (žr. 24 lent.), kadangi duomenys nėra pasiskirstę pagal normalųjį skirstinį (žr. 3.2.1.).

24 lentelė. *Skirtingo ilgio sakinių koreliacija*

			Autorius_1	Autorius_2	Autorius_3	Autorius_4	Autorius_5
Spearman's rho	Autorius_1	Correlation Coefficient	1,000	,786**	,754**	,911**	,771**
		Sig. (2-tailed)	.	,000	,000	,000	,000
		N	25	23	22	20	21
	Autorius_2	Correlation Coefficient	,786**	1,000	,840**	,780**	,602**
		Sig. (2-tailed)	,000	.	,000	,000	,004
		N	23	27	22	20	21
	Autorius_3	Correlation Coefficient	,754**	,840**	1,000	,706**	,517*
		Sig. (2-tailed)	,000	,000	.	,001	,019
		N	22	22	23	20	20
	Autorius_4	Correlation Coefficient	,911**	,780**	,706**	1,000	,868**
		Sig. (2-tailed)	,000	,000	,001	.	,000
		N	20	20	20	23	19
	Autorius_5	Correlation Coefficient	,771**	,602**	,517*	,868**	1,000
		Sig. (2-tailed)	,000	,004	,019	,000	.
		N	21	21	20	19	24

Matome, kad tarp 1 ir 4 autoriaus ($r = 0{,}911$) labai stipri teigiama koreliacija, tarp 1 ir 2 ($r = 0{,}786$), 1 ir 3 ($r = 0{,}754$), 1 ir 5 ($r = 0{,}771$), 2 ir 3 ($r = 0{,}840$), 2 ir 4 ($r = 0{,}780$), 3 ir 4 ($r = 0{,}706$), 4 ir 5 ($r = 0{,}868$) stipri teigiama koreliacija, tarp 2 ir 5 ($r = 0{,}620$) ir 3 ir 5

($r = 0{,}517$) vidutinio stiprumo teigiama koreliacija. Vadinasi, nagrinėjami autoriai panašiai vartoja skirtingo ilgio sakinius. Koreliacija tarp visų autorių yra statistiškai reikšminga, nes gauta p reikšmė (Sig. (2-tailed)) yra mažesnė už pasirinktą reikšmingumo lygmenį $\alpha = 0{,}05$.

3.2.1. Normalumo tikrinimas

Naudojantis SPSS programų paketu ir taikant Kolmogorovo-Smirnovo ir Šapiro-Vilko testus buvo tikrinta, ar skirtingo ilgio sakinių duomenys yra pasiskirstę pagal normalųjį skirstinį.

Tikrinama hipotezė:

$$\begin{cases} H_0: X \sim N(\bar{x}, s^2), \\ H_1: X \nsim N(\bar{x}, s^2), \end{cases}$$

kai $\alpha = 0{,}01$; čia X – 1 autoriaus, 2 autoriaus, 3 autoriaus, 4 autoriaus, 5 autoriaus sakinių ilgis (žodžiais).

25 lentelė. *Normalumo tikrinimas*

	Kolmogorov-Smirnov[a]			Shapiro-Wilk		
	Statistic	df	Sig.	Statistic	df	Sig.
Autorius_1	,240	25	,001	,772	25	,000
Autorius_2	,256	27	,000	,802	27	,000
Autorius_3	,226	23	,004	,798	23	,000
Autorius_4	,202	23	,016	,848	23	,002
Autorius_5	,168	24	,079	,884	24	,010

Iš 25 lentelėje pateiktų rezultatų matome, kad 1, 2 ir 3 autorių abiejų testų p reikšmės (Sig.) yra mažesnės, nei pasirinktas reikšmingumo lygmuo ($\alpha = 0{,}01$), todėl nulinę hipotezę atmetame. Tolesniuose tyrimuose negalime taikyti metodų, tinkančių normaliajam skirstiniui. 4 ir 5 autorių Kolmogorovo-Smirnovo testo p reikšmės (Sig.) yra didesnės, nei pasirinktas reikšmingumo lygmuo ($\alpha = 0{,}01$), bet Shapiro-Wilko testo p reikšmės nėra didesnės, todėl nulinę hipotezę atmetame. Tolesniuose tyrimuose taip pat negalime taikyti metodų, tinkančių normaliajam skirstiniui.

3.2.2. Hipotezės apie autorių porų vidurkių lygybę tikrinimas

Skirtingo ilgio sakiniams buvo tikrinta hipotezė apie autorių porų vidurkių lygybę:

$$\begin{cases} H_0: \text{pirmo ir antro autorių pavartotų sakinių ilgio vidurkiai lygūs}, \\ H_1: \text{pirmo ir antro autorių pavartotų sakinių ilgio vidurkiai nėra lygūs}. \end{cases}$$

Pagal 21 lentelėje pateiktus duomenis apskaičiuoti vidurkiai ir dispersijos pateikti 26 lentelėje.

26 lentelė. *Kiekvieno autoriaus sakinių ilgio vidurkis ir dispersija*

	N	Vidurkis	Dispersija
1 autorius	25	8	73,417
2 autorius	27	7,4074	51,635
3 autorius	23	8,6957	81,585
4 autorius	23	8,6957	62,040
5 autorius	24	8,3333	42,667

Tegu x – 1 autoriaus duomenų imtis, y – 2 autoriaus duomenų imtis.

$n = 25, m = 27, \bar{x} = 8, \bar{y} = 7{,}4074, s_x^2 = 73{,}417, s_y^2 = 51{,}635.$

Reikšmingumo lygmuo $\alpha = 0{,}05$.

Pirmiausia tikrinama hipotezė apie dviejų dispersijų lygybę:

$$\begin{cases} H_0^1: \sigma_X^2 = \sigma_Y^2, \\ H_1^1: \sigma_X^2 \neq \sigma_Y^2. \end{cases}$$

Remiantis (1) formule sudaroma kritinė sritis:

$$F = \frac{73{,}417}{51{,}635} = 1{,}4218.$$

Hipotezė H_0 neatmetama, jeigu $F_{1-\frac{\alpha}{2}}(n-1, m-1) \leq F \leq F_{\frac{\alpha}{2}}(n-1, m-1)$.

Remiantis (11) ir (12) formulėmis buvo apskaičiuotos reikalingos Fišerio skirstinio kvantilių reikšmės.

$k_1 = 24, k_2 = 26, p = \frac{\alpha}{2} = 0{,}025, u_p = u_{0{,}025} = 1{,}959.$

$$F_p(k_1, k_2) \approx \frac{k_2}{k_2 - 2}\sqrt{\frac{2(k_1 + k_2 - 2)}{k_1(k_2 - 4)}}u_p + \frac{k_2}{k_2 - 2} = 0{,}4619 \cdot 1{,}959 + 1{,}0833 = 1{,}9882;$$

$$F_p(k_2, k_1) \approx \frac{k_1}{k_1 - 2}\sqrt{\frac{2(k_2 + k_1 - 2)}{k_2(k_1 - 4)}}u_p + \frac{k_1}{k_1 - 2} = 0{,}4688 \cdot 1{,}959 + 1{,}0909 = 2{,}0093;$$

$$F_{1-p}(k_1, k_2) = \frac{1}{F_p(k_2, k_1)} = \frac{1}{2{,}0093} = 0{,}4977.$$

$$F_{1-\frac{\alpha}{2}}(n-1, m-1) = 0{,}4977;$$

$$F_{\frac{\alpha}{2}}(n-1, m-1) = 1{,}9882.$$

Kadangi $F_{1-\frac{\alpha}{2}}(n-1, m-1) = 0{,}4977 < F = 1{,}4218 < F_{\frac{\alpha}{2}}(n-1, m-1) = 1{,}9882$, tai H_0^1 priimama. Taigi, galima teigti, kad pirmo ir antro autorių rašomų sakinių ilgio dispersijos statistiškai reikšmingai nesiskiria.

Kadangi dipersijos statistiškai reikšmingai nesiskiria, galima naudoti Stjudento kriterijų ir tikrinti hipotezę apie pirmo ir antro autorių pavartotų sakinių ilgio vidurkių lygybę (H_0), kai populiacijų dispersijos lygios.

Sudaroma kritinė sritis

$$t = \frac{\bar{x} - \bar{y}}{\sqrt{(n-1)s_x^2 + (m-1)s_y^2}} \sqrt{\frac{nm(n+m-2)}{n+m}}:$$

$$t = \frac{8 - 7{,}4074}{\sqrt{(25 \cdot 73{,}417) + (27 \cdot 51{,}635)}} \sqrt{\frac{25 \cdot 27(25+27-2)}{25+27}} =$$

$$= \frac{0{,}5926}{\sqrt{3229{,}57}} \sqrt{\frac{33750}{52}} = 0{,}0104 \cdot 25{,}4762 = 0{,}2650.$$

Kadangi $|t| = 0{,}2650 < 2 = t_{0{,}025}(50)$, tai H_0 priimama. Taigi, galima teigti, kad pirmo ir antro autorių rašomų sakinių ilgio vidurkiai statistiškai reikšmingai nesiskiria.

Tos pačios hipotezės tikrinimo SPSS programų paketu rezultatai pateikti 27 lentelėje.

27 lentelė. *Stjudento kriterijus dviem nepriklausomoms imtims*

		Levene's Test for Equality of Variances		t-test for Equality of Means						
									95% Confidence Interval of the Difference	
		F	Sig.	t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	Lower	Upper
Sakinio_ilgis	Equal variances assumed	,112	,739	,271	50	,788	,59259	2,18707	-3,80026	4,98544
	Equal variances not assumed			,269	47,027	,789	,59259	2,20206	-3,83732	5,02251

Iš 27 lentelės matyti, kad gauta p reikšmė (Sig.) didesnė už pasirinktą reikšmingumo lygmenį ($p = 0{,}739 > 0{,}05 = \alpha$), vadinasi, grupių dispersijos statistiškai reikšmingai nesiskiria ir tolimesniam tyrimui naudojama lentelės eilutė, atitinkanti lygių dispersijų atvejį (Equal Variances assumed). Gauta p reikšmė (Sig.(2-tailed)) didesnė už pasirinktą reikšmingumo lygmenį ($p = 0{,}788 > 0{,}05 = \alpha$), vadinasi, autorių vidurkiai statistiškai reikšmingai nesiskiria, todėl hipotezė H_0 neatmetama – pirmo ir antro autorių pavartotų sakinių ilgio vidurkiai statistiškai reikšmingai nesiskiria.

Iš 28 lentelėje pateiktų rezultatų matome, kad visos gautos p reikšmės (Sig.) didesnės už pasirinktą reikšmingumo lygmenį ($\alpha = 0{,}05$), vadinasi, grupių dispersijos statistiškai reikšmingai nesiskiria. Lygių dispersijų atveju visos gautos p reikšmės (Sig.(2-tailed)) yra didesnės už pasirinktą reikšmingumo lygmenį ($\alpha = 0{,}05$), vadinasi, autorių vidurkiai statistiškai reikšmingai nesiskiria, todėl visos hipotezės H_0 neatmetamos, t. y. visų autorių porų pavartotų sakinių ilgio vidurkiai statistiškai reikšmingai nesiskiria.

28 lentelė. *Bendra visų autorių porų rezultatų lentelė*

Autorių pora	p reikšmė (Sig.)	p reikšmė (Sig. (2-tailed))	H_0 hipotezė
1 ir 2 autoriai	0,739	0,788	priimama
1 ir 3 autoriai	0,474	0,785	priimama
1 ir 4 autoriai	0,919	0,772	priimama
1 ir 5 autoriai	0,502	0,879	priimama
2 ir 3 autoriai	0,210	0,577	priimama
2 ir 4 autoriai	0,598	0,548	priimama
2 ir 5 autoriai	0,647	0,634	priimama
3 ir 4 autoriai	0,476	1,000	priimama
3 ir 5 autoriai	0,099	0,875	priimama
4 ir 5 autoriai	0,333	0,864	priimama

3.3. Žodžių ir raidžių skaičiaus sakinyje statistinė analizė

Detaliau nagrinėsime žodžių ir raidžių skaičius sakinyje. Tiriamos vienodo didumo imtys – po 200 kiekvieno autoriaus sakinių.

29 lentelė. *Žodžių ir raidžių skaičiaus sakinyje vidurkiai*

	1 autorius	2 autorius	3 autorius	4 autorius	5 autorius
Sakinių skaičius *N*	200	200	200	200	200
Žodžių skaičiaus sakinyje vidurkiai	7,7	7,92	6,24	7,82	9,1
Raidžių skaičiaus sakinyje vidurkiai	43,525	44,69	35,2	45,735	55,99

Jau 3.2. poskyrio pradžioje buvo išsiaiškinta, kad ilgiausiais sakiniais išsiskiria 5 autorius, trumpiausiais – 3 autorius, o 1, 2 ir 4 autoriai sakiniuose vartoja panašų tiek raidžių, tiek žodžių skaičių (žr. 29 lent.).

3.3.1. Normalumo tikrinimas

Simetriškoje skirstinio kreivėje moda ir mediana sutampa su vidurkiu. Jeigu šie rodikliai nesutampa, kreivė laikoma asimetriška. Asimetriškumui įvertinti naudojamas asimetrijos koeficientas *A*. Skirstinio kreivės vertikaliam iškilimui įvertinti naudojamas eksceso koeficientas *E*. Remiantis šiais rodikliais buvo išanalizuoti nagrinėjamų autorių sakinių ir žodžių ilgiai.

Vienas iš kriterijų duomenų normalumui tikrinti yra toks, kad vidurkis, mediana ir moda būtų beveik lygūs. Iš 22 ir 23 lentelėse pateiktų rezultatų matome, kad šios trys charakteristikos nesutampa visiems autoriams.

SPSS programų paketu surasti asimetrijos ir eksceso koeficientai bei jų standartinės paklaidos pateikti 30 ir 31 lentelėse.

30 lentelė. *Žodžių skaičiaus sakinyje asimetrijos ir eksceso koeficientai*

		Žodž_sk_sak_1 aut	Žodž_sk_sak_2 aut	Žodž_sk_sak_3 aut	Žodž_sk_sak_4 aut	Žodž_sk_sak_5 aut
N	Valid	200	200	200	200	200
	Missing	0	0	0	0	0
Mean		7,7000	7,9150	6,2400	7,8200	9,1000
Skewness		2,206	2,079	1,645	1,814	1,208
Std. Error of Skewness		,172	,172	,172	,172	,172
Kurtosis		8,032	6,559	2,723	5,190	2,035
Std. Error of Kurtosis		,342	,342	,342	,342	,342

Naudojantis 30 lentelėje gautais rezultatais tikrinama, ar žodžių skaičiaus sakinyje empirinis skirstinys skiriasi nuo normaliojo skirstinio.

Kadangi 1 autoriaus atveju $3 \cdot S_A = 0{,}516 < 2{,}206 = A$ ir $3 \cdot S_E = 1{,}026 < 8{,}032 = E$, tai empirinis skirstinys skiriasi nuo normaliojo. 2 autoriaus atveju $3 \cdot S_A = 0{,}516 < 2{,}079 = A$ ir $3 \cdot S_E = 1{,}026 < 6{,}559 = E$, todėl empirinis skirstinys skiriasi nuo normaliojo. 3 autoriaus atveju $3 \cdot S_A = 0{,}516 < 1{,}645 = A$ ir $3 \cdot S_E = 1{,}026 < 2{,}723 = E$, todėl empirinis skirstinys skiriasi nuo normaliojo. 4 autoriaus atveju $3 \cdot S_A = 0{,}516 < 1{,}814 = A$ ir $3 \cdot S_E = 1{,}026 < 5{,}190 = E$, todėl empirinis skirstinys skiriasi nuo normaliojo. 5 autoriaus atveju $3 \cdot S_A = 0{,}516 < 1{,}208 = A$ ir $3 \cdot S_E = 1{,}026 < 2{,}035 = E$, todėl empirinis skirstinys skiriasi nuo normaliojo.

31 lentelė. *Raidžių skaičiaus sakinyje asimetrijos ir eksceso koeficientai*

		Raidž_sk_sak_1 aut	Raidž_sk_sak_2 aut	Raidž_sk_sak_3 aut	Raidž_sk_sak_4 aut	Raidž_sk_sak_5 aut
N	Valid	200	200	200	200	200
	Missing	0	0	0	0	0
Mean		43,5250	44,6900	35,2000	45,7350	55,9900
Skewness		2,534	2,618	1,765	2,349	1,126
Std. Error of Skewness		,172	,172	,172	,172	,172
Kurtosis		10,495	11,949	2,990	9,624	1,661
Std. Error of Kurtosis		,342	,342	,342	,342	,342

Pagal 31 lentelės rezultatus tikrinama, ar žodžių skaičiaus sakinyje empirinis skirstinys skiriasi nuo normaliojo skirstinio.

1 autoriaus atveju: $3 \cdot S_A = 0{,}516 < 2{,}534 = A$ ir $3 \cdot S_E = 1{,}026 < 10{,}495 = E$. 2 autoriaus atveju: $3 \cdot S_A = 0{,}516 < 2{,}618 = A$ ir $3 \cdot S_E = 1{,}026 < 11{,}949 = E$. 3 autoriaus atveju: $3 \cdot S_A = 0{,}516 < 1{,}765 = A$ ir $3 \cdot S_E = 1{,}026 < 2{,}990 = E$. 4 autoriaus atveju: $3 \cdot S_A = 0{,}516 < 2{,}349 = A$ ir $3 \cdot S_E = 1{,}026 < 9{,}624 = E$. 5 autoriaus atveju: $3 \cdot S_A = 0{,}516 < 1{,}126 = A$ ir $3 \cdot S_E = 1{,}026 < 1{,}661 = E$. Vadinasi, visais atvejais empirinis skirstinys skiriasi nuo normaliojo.

Žodžių ir raidžių skaičiaus sakinyje normalumui patikrinti taikyti Kolmogorovo-Smirnovo ir Šapiro-Vilko testai. Naudojantis SPSS programų paketu patikrinta hipotezė:

$$\begin{cases} H_0: X \sim N(\bar{x}, s^2), \\ H_1: X \nsim N(\bar{x}, s^2), \end{cases}$$

kai $\alpha = 0{,}01$; čia X – 1 autoriaus, 2 autoriaus, 3 autoriaus, 4 autoriaus, 5 autoriaus sakinių ilgis (žodžiais ir raidėmis).

32 lentelė. *Žodžių skaičiaus sakinyje normalumo tikrinimas*

	Kolmogorov-Smirnov[a]			Shapiro-Wilk		
	Statistic	df	Sig.	Statistic	df	Sig.
Žodž_sk_sak_1aut	,176	200	,000	,799	200	,000
Žodž_sk_sak_2aut	,166	200	,000	,818	200	,000
Žodž_sk_sak_3aut	,180	200	,000	,829	200	,000
Žodž_sk_sak_4aut	,161	200	,000	,856	200	,000
Žodž_sk_sak_5aut	,134	200	,000	,918	200	,000

Kadangi abiejų testų p reikšmės (Sig.) yra mažesnės (žr. 32 ir 33 lent.), nei pasirinktas reikšmingumo lygmuo ($\alpha = 0{,}01$), todėl nulinė hipotezė atmetama – tiek žodžių, tiek raidžių skaičius sakinyje nėra pasiskirstęs pagal normalųjį skirstinį ir tolesniuose tyrimuose negalima taikyti metodų, tinkančių normaliajam skirstiniui.

33 lentelė. *Raidžių skaičiaus sakinyje normalumo tikrinimas*

	Kolmogorov-Smirnov[a]			Shapiro-Wilk		
	Statistic	df	Sig.	Statistic	df	Sig.
Raidž_sk_sak_1aut	,163	200	,000	,775	200	,000
Raidž_sk_sak_2aut	,142	200	,000	,790	200	,000
Raidž_sk_sak_3aut	,185	200	,000	,806	200	,000
Raidž_sk_sak_4aut	,149	200	,000	,821	200	,000
Raidž_sk_sak_5aut	,111	200	,000	,927	200	,000

3.3.2. Hipotezės apie autorių porų vidurkių lygybę tikrinimas

Žodžių ir raidžių skaičiui sakinyje SPSS programų paketu buvo tikrintos hipotezės apie autorių porų vidurkių lygybę:

$$\begin{cases} H_0: i \text{ ir } j \text{ autorių pavartotų žodžių/raidžių skaičiaus vidurkiai lygūs,} \\ H_1: i \text{ ir } j \text{ autorių pavartotų žodžių/raidžių skaičiaus vidurkiai nėra lygūs;} \end{cases}$$

čia $i, j = \overline{1,5}$.

Gauti žodžių skaičiaus sakinyje rezultatai pateikti 34 lentelėje, raidžių skaičiaus sakinyje rezultatai pateikti 35 lentelėje.

34 lentelė. *Bendra visų autorių porų žodžių skaičiaus sakinyje rezultatų lentelė*

Autorių pora	p reikšmė (Sig.)	p reikšmė (Sig. (2-tailed))	H_0 hipotezė
1 ir 2 autoriai	0,588	0,734	priimama
1 ir 3 autoriai	0,045	0,009	atmetama
1 ir 4 autoriai	0,036	0,830	priimama
1 ir 5 autoriai	0,235	0,014	atmetama
2 ir 3 autoriai	0,014	0,004	atmetama
2 ir 4 autoriai	0,011	0,871	priimama
2 ir 5 autoriai	0,088	0,045	atmetama
3 ir 4 autoriai	0,918	0,002	atmetama
3 ir 5 autoriai	0,341	0,000	atmetama
4 ir 5 autoriai	0,291	0,013	atmetama

Iš 34 lentelėje pateiktų rezultatų matyti, kad 1 ir 2 autorių, 1 ir 5 autorių, 2 ir 5 autorių, 3 ir 4 autorių, 3 ir 5 autorių, 4 ir 5 autorių gautos p reikšmės (Sig.) didesnės už pasirinktą reikšmingumo lygmenį ($\alpha = 0{,}05$), vadinasi, grupių dispersijos statistiškai reikšmingai nesiskiria ir tolimesniam tyrimui galima naudoti lygių dispersijų atvejį. 1 ir 3 autorių, 1 ir 4 autorių, 2 ir 3 autorių, 2 ir 4 autorių gautos p reikšmės (Sig.) mažesnės už pasirinktą reikšmingumo lygmenį ($\alpha = 0{,}05$), vadinasi, grupių dispersijos statistiškai reikšmingai skiriasi ir tolimesniam tyrimui naudojamas nelygių dispersijų atvejis. Pagal kitame stulpelyje pateiktus rezultatus matyti, kad 1 ir 2 autorių, 1 ir 4 autorių, 2 ir 4 autorių gautos p reikšmės (Sig.(2-tailed)) didesnės už pasirinktą reikšmingumo lygmenį ($\alpha = 0{,}05$), vadinasi, autorių vidurkiai statistiškai reikšmingai nesiskiria, todėl hipotezės H_0 (žodžių skaičiui) neatmetamos. 1 ir 3 autorių, 1 ir 5 autorių, 2 ir 3 autorių, 2 ir 5 autorių, 3 ir 4 autorių, 3 ir 5 autorių, 4 ir 5 autorių gautos p reikšmės (Sig.(2-tailed)) mažesnės už pasirinktą reikšmingumo lygmenį ($\alpha = 0{,}05$), vadinasi, autorių žodžių skaičiaus vidurkiai statistiškai reikšmingai skiriasi, todėl hipotezės H_0 atmetamos.

35 lentelė. *Bendra visų autorių porų raidžių skaičiaus sakinyje rezultatų lentelė*

Autorių pora	p reikšmė (Sig.)	p reikšmė (Sig. (2-tailed))	H_0 hipotezė
1 ir 2 autoriai	0,236	0,754	priimama
1 ir 3 autoriai	0,406	0,012	atmetama
1 ir 4 autoriai	0,411	0,510	priimama
1 ir 5 autoriai	0,533	0,000	atmetama
2 ir 3 autoriai	0,042	0,008	atmetama
2 ir 4 autoriai	0,046	0,771	priimama
2 ir 5 autoriai	0,481	0,002	atmetama
3 ir 4 autoriai	0,984	0,001	atmetama
3 ir 5 autoriai	0,109	0,000	atmetama
4 ir 5 autoriai	0,118	0,002	atmetama

Analogiškai (žr. 35 lent.), 1 ir 2 autorių, 1 ir 3 autorių, 1 ir 4 autorių, 1 ir 5 autorių, 2 ir 5 autorių, 3 ir 4 autorių, 3 ir 5 autorių, 4 ir 5 autorių poroms pritaikius lygių dispersijų atvejį, o 2 ir 3 autorių, 2 ir 4 autorių poroms – nelygių dispersijų atvejį, nustatyta, kad 1 ir 2 autorių, 1 ir 4 autorių, 2 ir 4 autorių raidžių skaičiaus vidurkiai statistiškai reikšmingai nesiskiria, todėl hipotezės H_0 neatmetamos, o 1 ir 3 autorių, 1 ir 5 autorių, 2 ir 3 autorių, 2 ir 5 autorių, 3 ir 4 autorių, 3 ir 5 autorių, 4 ir 5 autorių vidurkiai statistiškai reikšmingai skiriasi, todėl hipotezės H_0 atmetamos (raidžių skaičiui).

3.3.3. Hipotezės apie dviejų koreliacijos koeficientų lygybę tikrinimas

Kadangi duomenys nėra pasiskirstę pagal normalųjį skirstinį, norint pažiūrėti, kaip tarpusavyje koreliuoja žodžių ir raidžių skaičius sakinyje, skaičiuojamas Spirmeno ranginės koreliacijos koeficientas.

36 lentelė. *Koreliacija tarp žodžių ir raidžių skaičiaus sakinyje*

	Koreliacijos koeficientas
1 autoriaus	0,960
2 autoriaus	0,975
3 autoriaus	0,940
4 autoriaus	0,958
5 autoriaus	0,960

Iš 36 lentelėje pateiktų rezultatų matyti, kad tarp visų autorių pavartotų žodžių ir raidžių sakinyje yra labai stipri teigiama koreliacija, vadinasi, žodžių ir raidžių skaičiai sakinyje yra priklausomi, keičiantis žodžių skaičiui sakinyje tolygiai didėja arba mažėja ir raidžių skaičius

sakinyje. Visų tyrime analizuojamų autorių koreliacija tarp žodžių ir raidžių sakinyje yra statistiškai reikšminga, nes gautos p reikšmės yra mažesnės už pasirinktą reikšmingumo lygmenį $\alpha = 0{,}01$.

Tikrinama skirtingų autorių porų koreliacijos koeficientų tarp žodžių ir raidžių skaičiaus sakinyje lygybė. Žemiau pateikiama pora hipotezės tikrinimo pavyzdžių.

$$\begin{cases} H_0\text{: 1 ir 2 autoriaus koreliacijos koeficientai lygūs,} \\ H_1\text{: 1 ir 2 autoriaus koreliacijos koeficientai nelygūs.} \end{cases}$$

$\alpha = 0{,}05$.

$r_1 = 0{,}960, z_1 = 1{,}946, n = 200;$

$r_2 = 0{,}975, z_2 = 2{,}185, m = 200.$

$$Z = \frac{z_1 - z_2}{\sqrt{\frac{1}{n-3} + \frac{1}{m-3}}} = \frac{1{,}946 - 2{,}185}{\sqrt{\frac{1}{200-3} + \frac{1}{200-3}}} = \frac{-0{,}239}{\sqrt{\frac{2}{197}}} = \frac{-0{,}239}{0{,}101} = -2{,}366.$$

Kadangi $|z| = 2{,}366 > 1{,}959 = z_{0{,}025}$, tai H_0 atmetama.

$$\begin{cases} H_0\text{: 1 ir 4 autoriaus koreliacijos koeficientai lygūs,} \\ H_1\text{: 1 ir 4 autoriaus koreliacijos koeficientai nelygūs.} \end{cases}$$

$r_1 = 0{,}960, z_1 = 1{,}946, n = 200;$

$r_2 = 0{,}958, z_2 = 1{,}946, m = 200.$

$$Z = \frac{z_1 - z_2}{\sqrt{\frac{1}{n-3} + \frac{1}{m-3}}} = \frac{1{,}946 - 1{,}946}{\sqrt{\frac{1}{200-3} + \frac{1}{200-3}}} = \frac{0}{\sqrt{\frac{2}{197}}} = 0.$$

Kadangi $|z| = 0 \leq 1{,}959 = z_{0{,}005}$, tai H_0 neatmetama.

Atlikti skaičiavimai tarp kitų autorių porų pateikti 3 priede.

Hipotezės apie koreliacijos koeficientų lygybę tarp žodžių ir raidžių skaičiaus sakinyje gauti rezultatai pateikti 37 lentelėje.

37 lentelė. *Bendra visų autorių porų rezultatų lentelė*

Autorių pora	H_0 hipotezė
1 ir 2 autoriai	atmetama
1 ir 3 autoriai	atmetama
1 ir 4 autoriai	priimama
1 ir 5 autoriai	priimama
2 ir 3 autoriai	atmetama
2 ir 4 autoriai	atmetama
2 ir 5 autoriai	atmetama
3 ir 4 autoriai	atmetama
3 ir 5 autoriai	atmetama
4 ir 5 autoriai	priimama

Hipotezė H_0 atmetama daugumai porų.

3.3.4. Požymių nepriklausomumo tikrinimas

Pagal [2], sakiniai literatūroje yra grupuojami intervalais remiantis statistiniu kriterijumi: intervalų ribų pagrindu laikomos empirinio (imties) pasiskirstymo procentilės, integralinio pasiskirstymo „šuoliai“, atitinkantys sakinio ilgius. Šia ribas įvardijant yra išskiriamos penkios sakinių ilgio grupės: labai trumpi sakiniai – iki 8 žodžių imtinai, trumpi – nuo 9 iki 17 žodžių, vidutinio ilgio – nuo 18 iki 30 žodžių, ilgi – nuo 31 iki 46 žodžių, labai ilgi – 47 ir daugiau žodžių. Šiame darbe nagrinėjamos knygos skirtos paaugliams ir dauguma sakinių yra labai trumpi arba trumpi, todėl remiantis sakinių ilgio grupavimu turimi kiekvieno autoriaus sakiniai buvo suskirstyti į 3 grupes: 1–8 žodžių ilgio sakiniai, 9–17 žodžių ilgio sakiniai ir ilgesni nei 17 žodžių ilgio sakiniai.

Po kiek kiekvieno autoriaus sakinių pateko į atitinkamas grupes pateikta 38 lentelėje.

38 lentelė. *Sugrupuotų skirtingo ilgio sakinių skaičius*

	1-8 žodžių ilgio	9-17 žodžių ilgio	ilgesni nei 17 žodžių ilgio	∑
1 autorius	135	51	14	200
2 autorius	129	54	17	200
3 autorius	157	31	12	200
4 autorius	135	56	9	200
5 autorius	110	77	13	200
∑	666	269	65	1000

Tikrinama statistinė hipotezė:

$$\begin{cases} H_0: p_{ij} = p_i q_j, & \text{i} = 1, \dots, 5, \text{j} = 1, \dots, 3; \\ H_1: p_{ij} \neq p_i q_j, & \text{bent vienai porai (i, j);} \end{cases}$$

čia i – autorių skaičius, j – sakinių ilgio grupių skaičius.

Pagal (4) formulę apskaičiuojama kriterijaus statistika

$$\chi^2 = \sum_{i=1}^{5}\sum_{j=1}^{3} \frac{\left(O_{ij} - e_{ij}\right)^2}{e_{ij}} = \sum_{i=1}^{5}\sum_{j=1}^{3} \frac{\left(O_{ij} - \frac{n_i . n_{.j}}{n}\right)^2}{\frac{n_i . n_{.j}}{n}} = 30{,}9932,$$

kai tikėtinieji dažniai yra tokie:

$$e_{11} = \frac{200 \cdot 666}{1000} = 133{,}2, e_{12} = \frac{200 \cdot 269}{1000} = 53{,}8, e_{13} = \frac{200 \cdot 65}{1000} = 13,$$

$$e_{11} = e_{21} = e_{31} = e_{41} = e_{51}, e_{12} = e_{22} = e_{32} = e_{42} = e_{52}, e_{13} = e_{23} = e_{33} = e_{43} = e_{53},$$

$$e_{14} = e_{24} = e_{34} = e_{44} = e_{54}, e_{15} = e_{25} = e_{35} = e_{45} = e_{55}.$$

Kai $\alpha = 0{,}05, \chi^2_{0{,}05}((5-1)(3-1)) = \chi^2_{0{,}05}(8) = 15{,}507.$

Kadangi $\chi^2 = 30{,}9932 > 15{,}507 = \chi^2_{0{,}05}(8)$, tai hipotezė, apie požymių nepriklausomumą atmetama. Taigi, autoriai ir sakinių ilgiai susiję.

Ta pati hipotezė buvo patikrinta su SPSS programų paketu.

$$\begin{cases} H_0: \text{Sakinių ilgis nepriklauso nuo autoriaus,} \\ H_1: \text{Sakinių ilgis priklauso nuo autoriaus.} \end{cases}$$

39 lentelė. *χ^2 skaičiavimo rezultatai*

	Value	df	Asymp. Sig. (2-sided)
Pearson Chi-Square	30.993[a]	8	.000
Likelihood Ratio	31.628	8	.000
Linear-by-Linear Association	1.561	1	.211
N of Valid Cases	1000		

Kadangi $p = 0{,}000$ (žr. 39 lent.) yra mažesnė nei pasirinktas reikšmingumo lygmuo $\alpha = 0{,}05$, nulinė hipotezė H_0 atmetama. Galima teigti, kad tiriamieji požymiai yra priklausomi, tai reiškia, kad sakinių ilgis priklauso nuo autoriaus.

3.3.5. Vienfaktorinės dispersinės analizės taikymas

Norint sužinoti, ar vidutiniai žodžių skaičiai sakinyje statistiškai reikšmingai nesiskiria, tikrinta statistinė hipotezė:

$$\begin{cases} H_0: \mu_1 = \mu_2 = \mu_3 = \mu_4 = \mu_5, \\ H_1: \text{bent du vidurkiai skiriasi.} \end{cases}$$

Prieš taikant vienfaktorinę dispersinę analizę reikia patikrinti, ar kintamieji pasiskirstę pagal normalųjį dėsnį, ar kintamųjų dispersijos lygios, ar kintamieji nepriklausomi. Visi šie skaičiavimai atlikti 3.2. poskyryje.

Naudojantis 1.5.2. skyrelyje pateiktomis formulėmis atliekami skaičiavimai, o gauti tarpiniai rezultatai pateikiami 40 lentelėje.

40 lentelė. *Žodžių skaičiaus sakinyje rezultatai*

	1 autorius	2 autorius	3 autorius	4 autorius	5 autorius
n_i	200	200	200	200	200
T_i	1540	1583	1248	1564	1820
$\bar{x}_i$	7,7	7,915	6,24	7,82	9,1
$\sum x_i^2$	19200	21037	12848	17250	21926

$N = 1000, T = 7755, \bar{X} = 7{,}755, \sum x_{ij}^2 = 92261, k = 5, \alpha = 0{,}05, \sum_{i=1}^{k} \frac{T_i^2}{n_i} = 60967{,}445.$

$SSB = 60967{,}445 - 60140{,}025 = 827{,}42.$

$SSW = 92261 - 60967{,}445 = 31293{,}555.$

$SST = 92261 - 60140{,}025 = 32120{,}975.$

$k - 1 = 5 - 1 = 4.$

$N - k = 1000 - 5 = 995.$

$$MSB = \frac{SSB}{k-1} = \frac{827{,}42}{4} = 206{,}855.$$

$$MSW = \frac{SSW}{N-k} = \frac{31293{,}555}{995} = 31{,}4508.$$

$$F = \frac{MSB}{MSW} = \frac{206{,}855}{31{,}4508} = 6{,}5771.$$

Kadangi $F = 6{,}5771 > 2{,}3809 = F_{0{,}05}(4, 995)$, tai hipotezė H_0 atmetama. Taigi, vidutiniai žodžių skaičiai sakinyje statistiškai reikšmingai skiriasi. Galime teigti, kad žodžių skaičius sakinyje priklauso nuo autoriaus.

3.3.5.1. Bonferonio kriterijaus taikymas

Norint išsiaiškinti, kurių imčių vidurkiai statistiškai reikšmingai skiriasi, buvo taikytas Bonferonio kriterijus. Naudojantis 1.5.2. skyrelyje pateiktomis formulėmis apskaičiuotas Bonferonio kriterijus.

$\bar{X}_1 = 7{,}7;\ \bar{X}_2 = 7{,}915;\ \bar{X}_3 = 6{,}24;\ \bar{X}_4 = 7{,}82;\ \bar{X}_5 = 9{,}1.$

$N = 1000,\ k = 5,\ MSW = 31{,}4508,\ \alpha_E = 0{,}1.$

Kai reikšmingumo lygmuo

$$\alpha = \frac{2\alpha_E}{k(k-1)} = \frac{2 \cdot 0{,}1}{5(5-1)} = \frac{0{,}2}{20} = 0{,}01,$$

pagal (9) formulę apskaičiuojamas dydis

$$BSD = t_{0{,}005}(995)\sqrt{31{,}4508\left(\frac{1}{200} + \frac{1}{200}\right)} = 2{,}581 \cdot 0{,}5608 = 1{,}4474,$$

kuris naudojamas visiems lyginimams, nes visų imčių didumai vienodi.

$|\bar{x}_1 - \bar{x}_2| = |7{,}7 - 7{,}915| = 0{,}215 < BSD;$

$|\bar{x}_1 - \bar{x}_3| = |7{,}7 - 6{,}24| = 1{,}46 > BSD;$

$|\bar{x}_1 - \bar{x}_4| = |7{,}7 - 7{,}82| = 0{,}12 < BSD;$

$|\bar{x}_1 - \bar{x}_5| = |7{,}7 - 9{,}1| = 1{,}4 < BSD;$

$|\bar{x}_2 - \bar{x}_3| = |7{,}915 - 6{,}24| = 1{,}675 > BSD;$

$|\bar{x}_2 - \bar{x}_4| = |7{,}915 - 7{,}82| = 0{,}095 < BSD;$

$|\bar{x}_2 - \bar{x}_5| = |7{,}915 - 9{,}1| = 1{,}185 < BSD;$

$|\bar{x}_3 - \bar{x}_4| = |6{,}24 - 7{,}82| = 1{,}58 > BSD;$

$|\bar{x}_3 - \bar{x}_5| = |6{,}24 - 9{,}1| = 2{,}86 > BSD;$

$|\bar{x}_4 - \bar{x}_5| = |7{,}82 - 9{,}1| = 1{,}28 < BSD.$

Taigi, 1 ir 3 autoriaus, 2 ir 3 autoriaus, 3 ir 4 autoriaus, 3 ir 5 autoriaus vidutinis žodžių skaičius sakinyje statistiškai reikšmingai skiriasi, visų kitų porų vidutinio žodžių skaičiaus sakinyje skirtumas statistiškai nereikšmingas. Kadangi trečio autoriaus žodžių skaičius sakiniuose statistiškai reikšmingai skiriasi nuo visų kitų autorių, galima daryti prielaidą, kad tokius trečio autoriaus rezultatus gali įtakoti akivaizdžiai mažesnis pavartotų žodžių skaičius visose penkiose imtyse (1248 žodžiai, kai kitų autorių – 1540 ir daugiau). Ir analizuojant sakinių ilgį buvo gauta, kad trumpiausi sakiniai yra trečio autoriaus (vidutiniškai 6,24 žodžiai sakinyje, kai kitų autorių vidutiniškai 7,7 žodžių sakinyje ir daugiau). Remiantis raidžių skaičiaus sakinyje analize galima pastebėti, kad trečio autoriaus sakiniuose raidžių vartojama mažiausiai, vidutiniškai vartojamos 35,2 raidės viename sakinyje, kiti autoriai vidutiniškai vartoja 43,525 ir daugiau raidžių sakinyje.

Taikant Bonferonio kriterijų susiduriama su sunkumais, kai reikia rasti tinkamą Studento skirstinio kritinę reikšmę. Šios problemos pavyksta išvengti, kai naudojamas statistinis paketas.

Paprastai tuo atveju pateikiamos kriterijaus p-reikšmės. Jeigu p-reikšmė yra mažesnė už α, tai vidurkiai statistiškai skiriasi. Priešingu atveju – statistiškai reikšmingo skirtumo nėra.

41 lentelėje pateikti rezultatai, gauti skaičiuojant SPSS programų paketu.

41 lentelė. *Bonferonio kriterijaus rezultatai*

(I) Autorius	(J) Autorius	Mean Difference (I-J)	Std. Error	Sig.	90% Confidence Interval	
					Lower Bound	Upper Bound
1 autorius	2 autorius	-.21500	.56081	1.000	-1.6623	1.2323
	3 autorius	1.46000*	.56081	.094	.0127	2.9073
	4 autorius	-.12000	.56081	1.000	-1.5673	1.3273
	5 autorius	-1.40000	.56081	.127	-2.8473	.0473
2 autorius	1 autorius	.21500	.56081	1.000	-1.2323	1.6623
	3 autorius	1.67500*	.56081	.029	.2277	3.1223
	4 autorius	.09500	.56081	1.000	-1.3523	1.5423
	5 autorius	-1.18500	.56081	.348	-2.6323	.2623
3 autorius	1 autorius	-1.46000*	.56081	.094	-2.9073	-.0127
	2 autorius	-1.67500*	.56081	.029	-3.1223	-.2277
	4 autorius	-1.58000*	.56081	.049	-3.0273	-.1327
	5 autorius	-2.86000*	.56081	.000	-4.3073	-1.4127
4 autorius	1 autorius	.12000	.56081	1.000	-1.3273	1.5673
	2 autorius	-.09500	.56081	1.000	-1.5423	1.3523
	3 autorius	1.58000*	.56081	.049	.1327	3.0273
	5 autorius	-1.28000	.56081	.227	-2.7273	.1673
5 autorius	1 autorius	1.40000	.56081	.127	-.0473	2.8473
	2 autorius	1.18500	.56081	.348	-.2623	2.6323
	3 autorius	2.86000*	.56081	.000	1.4127	4.3073
	4 autorius	1.28000	.56081	.227	-.1673	2.7273

Gauti analogiški rezultatai – vidutinis žodžių skaičius sakinyje statistiškai reikšmingai skiriasi tarp 1 ir 3 autoriaus, 2 ir 3 autoriaus, 3 ir 4 autoriaus ir tarp 3 ir 5 autoriaus, kadangi gauta p-reikšmė (Sig.) yra mažesnė nei pasirinktas eksperimento reikšmingumo lygmuo $\alpha_E = 0{,}1$.

3.3.5.2. Koreliacijos koeficiento ICC taikymas

Norint palyginti duomenų skirtumus pačiose imtyse su imčių skirtumais taikytas intraklasinės koreliacijos koeficientas. Remiantis anksčiau gautais kriterijaus statistikos rezultatais taikant vienfaktorinę dispersinę analizę, pagal (10) formulę galima apskaičiuoti koreliacijos koeficientą ICC.

Visų imčių didumai yra vienodi $n = 200$, $F = 6{,}5771$.

$$ICC = \frac{6{,}5771 - 1}{6{,}5771 + (200 - 1)} = \frac{5{,}5771}{205{,}5771} = 0{,}0271.$$

Kadangi gauta ICC reikšmė yra pakankamai maža, tai reiškia, kad kiekvienos imties duomenys yra labai mažai panašūs ir visose imtyse yra didelė duomenų įvairovė. Duomenų skirtumus mažai lemia priklausymas konkrečiai populiacijai, šiuo atveju, konkrečiam autoriui.

Intraklasinės koreliacijos koeficientą apskaičiavus SPSS programų paketu, gauti rezultatai pateikti 42 lentelėje.

42 lentelė. *ICC koeficiento rezultatai*

	Intraclass Correlation[b]	90% Confidence Interval		F Test with True Value 0			
		Lower Bound	Upper Bound	Value	df1	df2	Sig
Single Measures	.032[a]	-.020	.084	1.065	999	999	.158
Average Measures	.061[c]	-.042	.154	1.065	999	999	.158

Intraklasinis koreliacijos koeficientas statistiškai reikšmingai nesiskiria nuo 0, kadangi F kriterijaus p-reikšmė (Sig.) didesnė už reikšmingumo lygmenį $\alpha = 0{,}1$.

3.4. Raidžių dažnio statistinė analizė

Kiekvieno autoriaus naudojamų skirtingų raidžių dažnis pateikiamas 43 lentelėje.

Galima pastebėti, kad dažniausiai visų autorių vartojama raidė ***i***, rečiausiai 1 ir 2 autorius vartoja raidę ***h***, 4 autorius nei karto jos nevartojo, o 3 ir 5 autorius rečiausiai vartoja raidę ***f***. Taip pat galima pastebėti, kad visi autoriai pakankamai dažnai vartoja raides ***a***, ***s***, o retai, be jau minėtų ***f*** ir ***h***, vartoja ***c*** ir ***z***.

[18] atliktas tyrimas lyginant raidžių dažnius dviejuose žodynuose (V. Žilinskienės žodyne ir A. Utkos žodyne) ir Vikipedijoje rodo, kad dažniausiai pasitaikanti raidė yra ***i***, po to ***a***, ***s***, ***o***, ***r***, ***e***, ***t***, rečiausiai pasitaiko raidė ***ę***, prieš ją, šiek tiek dažniau – ***h***, ***z***, ***f***, ***ū***, ***č***, ***į***, ***ą***, ***c.*** Šiame baigiamajame darbe nagrinėjamų penkių autorių kūriniuose dažniausiai ir rečiausiai vartojamos raidės beveik sutampa su [18] straipsnyje aprašytais rezultatais.

43 lentelė. *Kiekvieno autoriaus raidžių dažnis*

	Viso raidžių				
Raidė	**1 autorius**	**2 autorius**	**3 autorius**	**4 autorius**	**5 autorius**
a	981	1067	849	1002	1354
ą	64	54	54	80	101
b	107	148	135	186	149
c	9	10	7	6	6
č	32	28	23	71	55
d	214	215	169	204	295
e	497	500	414	468	594
ę	38	28	26	25	29
ė	272	243	193	329	240
f	12	11	1	4	6
g	139	208	137	193	226
h	2	3	3	0	19
i	1127	1245	905	1191	1413
į	92	71	44	68	82
y	105	121	98	89	115
j	203	235	154	215	218
k	442	453	374	478	578
l	269	273	212	308	420
m	270	252	197	251	321
n	458	425	365	423	555
o	468	423	369	544	560
p	279	277	207	300	380
r	426	435	315	478	505
s	692	690	532	679	951
š	154	131	113	127	207
t	548	535	490	540	579
u	434	393	336	442	730
ų	69	86	46	67	72
ū	43	56	22	40	55
v	170	210	180	238	241
z	9	11	7	11	10
ž	80	101	63	90	132
Viso raidžių	8705	8938	7040	9147	11198

3.4.1. Hipotezės apie dviejų proporcijų lygybę tikrinimas

Norint išsiaiškinti, ar autoriai vienodai vartoja tą pačią raidę, buvo tikrinamos hipotezės apie dviejų proporcijų lygybę skirtingoms autorių poroms.

Pasirinkus dažniausiai vartojamą raidę *i* pagal (2) formulę tikrinama statistinė hipotezė:

$$\begin{cases} H_0: p_1 = p_2, \\ H_1: p_1 \neq p_2. \end{cases}$$

1 ir 2 autorių palyginimas:

$$\widehat{p_1} = \frac{1127}{8705} = 0{,}12947;$$

$$\widehat{p_2} = \frac{1245}{8938} = 0{,}13929;$$

$$\bar{p} = \frac{2372}{17643} = 0{,}13444;$$

$$Z = \frac{\widehat{p_1} - \widehat{p_2}}{\sqrt{\bar{p}(1-\bar{p})\left(\frac{1}{n} + \frac{1}{m}\right)}} = \frac{0{,}12947 - 0{,}13929}{\sqrt{0{,}134(1-0{,}134)\left(\frac{1}{8705} + \frac{1}{8938}\right)}} = \frac{-0{,}00982}{\sqrt{0{,}116(0{,}00011 + 0{,}00011)}}$$

$$= \frac{-0{,}00982}{\sqrt{0{,}000026}} = \frac{-0{,}00982}{0{,}0051} = -1{,}925;$$

$$z_{\frac{\alpha}{2}} = z_{\frac{0{,}05}{2}} = z_{0{,}025} = 1{,}959.$$

Kadangi $|Z| = 1{,}925 \leq 1{,}959 = z_{0{,}025}$, tai H_0 neatmetama.

2 ir 3 autorių palyginimas:

$$\widehat{p_1} = \frac{1245}{8938} = 0{,}13929;$$

$$\widehat{p_2} = \frac{905}{7040} = 0{,}12855;$$

$$\bar{p} = \frac{2150}{15978} = 0{,}13456;$$

$$Z = \frac{\widehat{p_1} - \widehat{p_2}}{\sqrt{\bar{p}(1-\bar{p})\left(\frac{1}{n} + \frac{1}{m}\right)}} = \frac{0{,}13929 - 0{,}12855}{\sqrt{0{,}135(1-0{,}135)\left(\frac{1}{8938} + \frac{1}{7040}\right)}} = \frac{0{,}01074}{\sqrt{0{,}117(0{,}00011 + 0{,}00014)}}$$

$$= \frac{0{,}01074}{\sqrt{0{,}000029}} = \frac{0{,}01074}{0{,}0054} = 1{,}988;$$

$$z_{\frac{\alpha}{2}} = z_{\frac{0{,}05}{2}} = z_{0{,}025} = 1{,}959.$$

Kadangi $|Z| = 1{,}988 > 1{,}959 = z_{0{,}025}$, tai H_0 atmetama.

Skaičiavimai, atlikti lyginant kitas autorių poras, pateikti 4 priede.

Patikrinus hipotezes su visomis autorių poromis gaunama, kad hipotezė H_0 atmetama su 2 ir 3 autoriaus, 2 ir 5 autoriaus poromis (žr. 44 lent.).

44 lentelė. *Raidės i rezultatai*

Autorių pora	H_0 **hipotezė**
1 ir 2 autorius	neatmetama
1 ir 3 autorius	neatmetama
1 ir 4 autorius	neatmetama
1 ir 5 autorius	neatmetama
2 ir 3 autorius	atmetama
2 ir 4 autorius	neatmetama
2 ir 5 autorius	atmetama
3 ir 4 autorius	neatmetama
3 ir 5 autorius	neatmetama
4 ir 5 autorius	neatmetama

Patikrinus proporcijų lygybės hipotezes su visomis abėcėlės raidėmis galima pastebėti (žr. 45 lent.), kad pagal raides *t, h, ė, u* galima identifikuoti penktą autorių (t. y. atskirti nuo kitų keturių), nes visoms poroms, kuriose yra penktas autorius, hipotezė atmetama. Raidė *č* padeda identifikuoti ketvirtą autorių, nes visoms poroms, kuriose yra ketvirtas autorius, hipotezė atmetama. Taip pat galime pastebėti, kad raidės *c, z, d, k, m, p, ž* procentiškai visų autorių kūryboje vartojamos vienodai dažnai ir visiškai netinka autorių identifikacijai.

Gautus rezultatus lyginant su tiriamųjų kūrinių ilgiais (1 autoriaus kūrinio ilgis 200 psl., 2 autoriaus 254 psl., 3 autoriaus 232 psl., 4 autoriaus 224 psl., 5 autoriaus 277 psl.), galima pastebėti, kad raidžių *t, h, f, ę, ė, g, į, l, u* vartojimas statistiškai reikšmingai skiriasi tarp trumpiausio, t. y 1 autoriaus kūrinio, ir ilgiausio, t. y. 5 autoriaus kūrinio.

Gautus rezultatus lyginant su žodžių skaičiumi imtyse (1 autoriaus žodžių skaičius tiriamoje imtyje 1540, 2 autoriaus – 1583 žodžiai, 3 autoriaus – 1248 žodžiai, 4 autoriaus – 1564 žodžiai, 5 autoriaus – 1820 žodžiai), galima pastebėti, kad raidžių *s, t, h, b, ė, y, l, u* vartojimas statistiškai reikšmingai skiriasi tarp mažiausiai žodžių pavartojusio 3 autoriaus ir 5 autoriaus.

45 lentelė. *Raidžių dažniai*

Raidė	Z (1 ir 2 aut.)	Z (1 ir 3 aut.)	Z (1 ir 4 aut.)	Z (1 ir 5 aut.)	Z (2 ir 3 aut.)	Z (2 ir 4 aut.)	Z (2 ir 5 aut.)	Z (3 ir 4 aut.)	Z (3 ir 5 aut.)	Z (4 ir 5 aut.)
i	-1.913	0.170	-0.147	0.688	1.975	1.789	2.731	-0.311	0.467	0.855
a	-1.386	-1.538	0.669	-1.788	-0.235	2.077	-0.333	2.191	-0.064	-2.521
s	0.567	0.915	1.320	-1.381	0.385	0.754	-1.991	0.320	-2.251	-2.795
t	0.856	-1.672	1.093	3.405	-2.496	0.233	2.514	2.730	5.009	2.281
e	0.332	-0.458	1.751	1.245	-0.774	1.427	0.901	2.124	1.658	-0.600
h	-0.418	-0.688	1.450	-3.162	-0.293	1.752	-2.905	1.974	-2.407	-3.941
f	0.272	2.686	2.101	1.962	2.494	1.853	1.687	-1.060	-1.322	-0.315
c	-0.172	0.077	0.871	1.270	0.240	1.047	1.459	0.753	1.130	0.351
z	-0.388	0.077	-0.337	0.319	0.442	0.054	0.738	-0.394	0.218	0.684
ę	1.341	0.659	1.838	2.145	-0.606	0.497	0.720	1.080	1.323	0.198
ą	-1.386	-1.538	0.669	-1.788	-0.235	2.077	-0.333	2.191	-0.064	-2.521
b	-2.374	-3.491	-4.228	-0.630	-1.245	-1.886	1.902	-0.524	3.117	3.919
č	0.620	0.432	-3.603	-1.311	-0.149	-4.219	-1.957	-3.731	-1.657	2.578
d	0.228	0.234	1.007	-0.780	0.020	0.783	-1.027	0.716	-0.976	-1.854
ė	1.601	1.412	-1.748	4.338	-0.088	-3.373	2.652	-3.054	2.583	6.255
g	-3.493	-1.660	-2.537	-2.198	1.646	0.992	1.500	-0.732	-0.340	0.458
į	1.822	2.912	2.221	2.440	1.257	0.392	0.504	-0.901	-0.851	0.092
y	-0.871	-1.028	1.502	1.200	-0.207	2.390	2.141	2.473	2.234	-0.384
j	-1.269	0.606	-0.082	1.874	1.799	1.203	3.245	-0.689	1.120	1.985
k	0.028	-0.661	-0.448	-0.267	-0.692	-0.479	-0.299	0.245	0.446	0.205
l	0.138	0.286	-1.046	-2.528	0.157	-1.193	-2.693	-1.273	-2.658	-1.465
m	1.106	1.116	1.419	0.969	0.080	0.308	-0.200	0.208	-0.270	-0.526
n	1.542	0.215	1.964	0.971	-1.244	0.415	-0.659	1.642	0.686	-1.100
o	1.952	0.375	-1.649	1.187	-1.471	-3.630	-0.878	-1.930	0.719	2.964
p	0.403	0.955	-0.282	-0.737	0.581	-0.691	-1.168	-1.229	-1.688	-0.449
r	0.083	1.235	-1.011	1.273	1.164	-1.102	1.194	-2.195	-0.112	2.369
š	1.598	0.792	2.042	-0.417	-0.714	0.438	-2.101	1.131	-1.219	-2.569
u	1.850	0.616	0.474	-4.573	-1.130	-1.394	-6.520	-0.175	-4.894	-5.137
ų	-1.206	1.020	0.462	1.249	2.141	1.686	2.551	-0.599	0.086	0.771
ū	-1.179	1.766	0.556	0.028	2.828	1.751	1.289	-1.274	-1.812	-0.561
v	-1.814	-2.556	-2.901	-0.981	-0.843	-1.092	0.940	-0.179	1.772	2.105
ž	-1.391	0.159	-0.446	-1.771	1.464	0.961	-0.322	-0.580	-1.815	-1.331

3.4.2. Polinominio skirstinio hipotezių tikrinimas

Atsižvelgiant į jau prieš tai atliktus skaičiavimus ir gautus rezultatus (3.4.1. skyrelyje) prieš tikrinant polinominio skirstinio hipotezes visos raidės buvo sugrupuotos į keturias grupes:

- galima identifikuoti 5-ą autorių (ė, h, t, u),
- visiškai netinka autorių identifikacijai (c, d, k, m, p, z, ž),
- likusios balsės (a, ą, e, ę, i, į, y, o, ų, ū),

- likusios priebalsės (b, č, f, g, j, l, n, r, s, š, v).

Taip suskirstytų raidžių grupių dažniai pateikti 46 lentelėje.

46 lentelė. *Raidžių grupių dažniai*

		ė, h, t, u	c, d, k, m, p, z, ž	a, ą, e, ę, i, į, y, o, ų, ū	b, č, f, g, j, l, n, r, s, š, v	Raidžių skaičius iš viso (n)
Grupės raidžių dažnis	1 autorius	0,1443	0,1497	0,4002	0,3058	8705
	2 autorius	0,1313	0,1476	0,4085	0,3126	8938
	3 autorius	0,1452	0,1455	0,4016	0,3078	7040
	4 autorius	0,1433	0,1465	0,3907	0,3194	9147
	5 autorius	0,1400	0,1538	0,3907	0,3155	11198

Naudojantis Pirsono statistika (3) buvo atlikti skaičiavimai ir tikrintos hipotezės:

$$\begin{cases} H_0: p_i = p_{i0}, \\ H_1: p_i \neq p_{i0}; \end{cases}$$

čia $i = 1,2,3,4,5$.

Norint palyginti pirmą ir antrą autorius išsikelta hipotezė:

$$\begin{cases} H_0: \text{pirmo ir antro autoriaus atitinkamų grupių raidžių pasirodymo dažniai statistiškai reikšmingai nesiskiria,} \\ H_1: \text{pirmo ir antro autoriaus atitinkamų grupių raidžių pasirodymo dažniai statistiškai reikšmingai skiriasi.} \end{cases}$$

$n_2 = 8938$

$$\chi^2 = 8938 \cdot \left(\frac{(0,1313 - 0,1443)^2}{0,1443} + \frac{(0,1476 - 0,1497)^2}{0,1497} + \frac{(0,4085 - 0,4002)^2}{0,4002} + \frac{(0,3126 - 0,3058)^2}{0,3058} \right) = 8938 \cdot 0,0015 = 13,5024.$$

Kai $\alpha = 0,05, k = 4, \chi^2_{0,95}(3) = 7,81$.

Kadangi $\chi^2 = 13,5024 > \chi^2_{0,95}(3) = 7,81$, tai hipotezė H_0 atmetama. Vadinasi, išskirtos raidžių grupės gali padėti atskirti vieną autorių nuo kito.

Kai $\alpha = 0,1, k = 4, \chi^2_{0,9}(3) = 6,25$.

Kadangi $\chi^2 = 13,407 > \chi^2_{0,9}(3) = 6,25$, tai hipotezė H_0 taip pat atmetama.

Norint palyginti išskirtų grupių raidžių pasirodymo dažnius, ta pati polinominio skirstinio hipotezė buvo tikrinta visoms autorių poroms. Gauti rezultatai pateikti 47 lentelėje.

47 lentelė. *Polinominio skirstinio hipotezės tikrinimo rezultatai*

Autorių pora	χ^2	Kai $\alpha = 0{,}05$, hipotezė H_0	Kai $\alpha = 0{,}1$, hipotezė H_0
1 ir 2 autorius	13,5024	atmetama	atmetama
1 ir 3 autorius	1,0040	priimama	priimama
1 ir 4 autorius	8,3136	atmetama	atmetama
1 ir 5 autorius	8,6521	atmetama	atmetama
2 ir 1 autorius	14,0898	atmetama	atmetama
2 ir 3 autorius	11,7930	atmetama	atmetama
2 ir 4 autorius	18,4900	atmetama	atmetama
2 ir 5 autorius	18,3130	atmetama	atmetama
3 ir 1 autorius	1,2705	priimama	priimama
3 ir 2 autorius	13,7670	atmetama	atmetama
3 ir 4 autorius	6,9798	priimama	atmetama
3 ir 5 autorius	12,8202	atmetama	atmetama
4 ir 1 autorius	7,7463	priimama	atmetama
4 ir 2 autorius	17,5370	atmetama	atmetama
4 ir 3 autorius	5,3180	priimama	priimama
4 ir 5 autorius	5,4497	priimama	priimama
5 ir 1 autorius	6,6991	priimama	atmetama
5 ir 2 autorius	14,5185	atmetama	atmetama
5 ir 3 autorius	7,9502	atmetama	atmetama
5 ir 4 autorius	4,3168	priimama	priimama

Kai $\alpha = 0{,}1$, hipotezė H_0 atmetama su 1 ir 2 autoriaus, 1 ir 4 autoriaus, 1 ir 5 autoriaus, 2 ir 1 autoriaus, 2 ir 3 autoriaus, 2 ir 4 autoriaus, 2 ir 5 autoriaus, 3 ir 2 autoriaus, 3 ir 4 autoriaus, 4 ir 1 autoriaus, 4 ir 2 autoriaus, 5 ir 1 autoriaus, 5 ir 2 autoriaus, 5 ir 3 autoriaus poromis, tai reiškia, kad šios autorių poros išskirtas raidžių grupes vartoja nevienodai dažnai. Hipotezė H_0 priimama su 1 ir 3 autoriaus, 3 ir 1 autoriaus, 4 ir 3 autoriaus, 4 ir 5 autoriaus, 5 ir 4 autoriaus poromis, tai reiškia, kad atitinkamų grupių raidžių pasirodymo dažniai statistiškai reikšmingai nesiskiria.

Pagal gautus rezultatus tikrinant hipotezę su reikšmingumo lygmeniu $\alpha = 0{,}05$ galima pastebėti, kad išsiskiria 2 autorius, kadangi tiek visus likusius autorius lyginant su 2 autoriumi, tiek 2 autorių lyginant su likusiais autoriais, hipotezė H_0 yra atmetama, vadinasi, šiam autoriui išskirtos raidžių grupės yra individualios.

IŠVADOS

1. Atlikus požymių nepriklausomumo tikrinimą žodžių ir sakinių ilgiui gauta, kad tiriamieji požymiai yra priklausomi, tai reiškia, kad žodžių ir sakinių ilgis priklauso nuo autoriaus.
2. Vienfaktorinės dispersinės analizės rezultatai taip pat patvirtino, kad žodžių skaičius sakinyje priklauso nuo autoriaus.
3. Atlikus raidžių vartojimo analizę gauta, kad dažniausiai vartojama visų autorių raidė *i*, rečiausiai 1 ir 2 autoriai vartojo raidę *h*, 4 autorius nevartojo, o 3 ir 5 autoriai rečiausiai vartojo raidę *f*.
4. Atlikus skaičiavimus gauta, kad pagal raides *t, h, ė, u* galima identifikuoti penktą autorių (t. y. atskirti nuo kitų keturių). Raidė *č* padeda identifikuoti ketvirtą autorių. Taip pat galima pastebėti, kad raidės *c, z, d, k, m, p, ž* procentiškai visų autorių kūryboje vartojamos vienodai dažnai ir visiškai netinka autorių identifikacijai.
5. Lyginant raidžių dažnius su tiriamųjų kūrinių ilgiais, galima pastebėti, kad raidžių *t, h, f, ę, ė, g, į, l, u* vartojimas statistiškai reikšmingai skiriasi tarp trumpiausio, t. y. 1 autoriaus kūrinio, ir ilgiausio, t. y. 5 autoriaus kūrinio.
6. Lyginant raidžių dažnius su žodžių skaičiumi imtyse, galima pastebėti, kad raidžių *s, t, h, b, ė, y, l, u* vartojimas statistiškai reikšmingai skiriasi tarp mažiausiai žodžių pavartojusio 3 autoriaus ir 5 autoriaus.
7. Patikrinus polinominio skirstinio hipotezę visoms autorių poroms galima pastebėti, kad hipotezė atmetama visoms poroms, kuriose yra antras autorius. Vadinasi, išskirtos raidžių grupės šiam autoriui yra individualios.
8. Pritaikius koreliacijos koeficientą ICC gauta, kad kiekvienos imties duomenys yra labai mažai panašūs ir visose imtyse yra didelė duomenų įvairovė.

LITERATŪRA

1. Bagdonavičius, V.; Kruopis, J. *Matematinė statistika II*. Vilnius: TEV leidykla, 2009.
2. Bitinienė, A. *Lietuvių kalbotyros klausimai XXXIII, Gramatika ir leksikologija*. 1995.
3. Čekanavičius, V; Murauskas, G. *Statistika ir jos taikymai I*. Vilnius: TEV leidykla, 2006.
4. Čekanavičius, V; Murauskas, G. *Statistika ir jos taikymai II*. Vilnius. TEV leidykla, 2008.
5. Dažnių skirstinių formos charakteristikos [žiūrėta 2020-11-26]. Prieiga per internetą: <*https://www.spssanalize.lt/dazniu-skirstiniu-formos-charakteristikos/*>.
6. Gonestas, E; Strielčiūnas, R. R. *Taikomoji statistika*. Lietuvos kūno kultūros akademija, 2003.
7. Identiteto sąvoka [žiūrėta 2020-01-04]. Prieiga per internetą: <*https://www.zodynas.lt/terminu-zodynas/I/identitetas*>.
8. Kanišauskas, V. *Tikimybių teorijos ir matematinės statistikos pagrindai*. Šiauliai: Šiaulių universiteto leidykla, 2000.
9. Kanišauskas, V. *Vidinės migracijos matematinis modelis*. Lietuvos matematikos rinkinys, Lietuvos matematikų draugijos darbai, ser. B, 57 t., 2016, 7-12.
10. Knygų skaitymo nauda [žiūrėta 2020-01-04]. Prieiga per internetą: <*http://www.sviesuva.lt/knygu-skaitymo-nauda/*>.
11. Krapavickaitė, D; Plikusas, A. *Imčių teorijos pagrindai*. VGTU leidykla „Technika“, 2005.
12. Krapavickaitė, D. *R programa ir jos taikymas imčių tyrimams*. VGTU leidykla „Technika“, 2017.
13. Kruopis, J. *Matematinė statistika*. Vilnius: Mokslo ir enciklopedijų leidykla, 1993.
14. Leonavičienė, T. *SPSS programų paketo taikymas statistiniuose tyrimuose*. Vilnius: Vilniaus pedagoginis universitetas, 2007.
15. Paskaita „Kuo esame unikalūs. Kiekvienos asmenybės išskirtinumas“ [žiūrėta 2020-01-04]. Prieiga per internetą: <*https://www.silaineskrastas.lt/renginiai/paskaita-kuo-esame-unikalus-kiekvienos-asmenybes-isskirtinumas/*>.
16. Prozos sąvoka [žiūrėta 2021-01-04]. Prieiga per internetą: <*https://www.vle.lt/straipsnis/proza/*>.
17. Pukėnas, K. *Kokybinių duomenų analizė SPSS programa*. Kaunas: Lietuvos kūno kultūros akademija, 2009.
18. Raidžių dažnių lietuvių ir kitose kalbose, vartojančiose lotyniškus rašmenis, analizė [žiūrėta 2020-11-05]. Prieiga per internetą: <*http://www.ims.mii.lt/ims/asmen/gintas/publ/gg15-raid%c4%97s.pdf*>.
19. What is R Programming Language? Introduction & Basics of R [žiūrėta 2021-05-17]. Prieiga per internetą: <*https://www.guru99.com/r-programming-introduction-basics.html*>.

SANTRAUKA

Magistro darbo objektas yra penkios skirtingų autorių knygos. Darbo tikslas – nustatyti skirtingų prozos kūrėjų identitetą. Tikslui pasiekti iškelti keli uždaviniai. Pirmiausia – išsiaiškinti, kokie nagrinėjamų kūrinių lingvistiniai požymiai tinka statistinei analizei. Šiam uždaviniui pasiekti, naudojantis programine įranga R, buvo paruošti tekstai: atskirti žodžiai ir skyrybos ženklai, sakiniai išskaidyti žodžiais ir raidėmis, tada suskaičiuoti žodžiai ir raidės bei pasikartojančių raidžių ir žodžių dažniai. Gautiems kiekvieno autoriaus pasikartojančių skirtingo ilgio žodžių, kiekvieno autoriaus pasikartojančių skirtingo ilgio skirtingų žodžių ir skirtingo ilgio sakinių duomenims atlikta statistinė analizė. Kitas uždavinys – ištirti, kokie matematinės statistikos metodai tinka užsibrėžtam tikslui pasiekti. Buvo ištirta, kad norint nustatyti skirtingų prozos kūrėjų identitetą galima taikyti (lingvo)statistinę duomenų analizę, koreliacinę analizę, vienfaktorinę dispersinę analizę bei kompiuterinės statistikos metodus. Ir dar vienas darbo uždavinys – taikant pasirinktus matematinės statistikos metodus išsiaiškinti, kuo autoriai skiriasi vienas nuo kito. Pavyko išsiaiškinti, kad pagal raides *t, h, ė, u* galima identifikuoti penktą autorių. Raidė *č* padeda identifikuoti ketvirtą autorių. Taip pat galima pastebėti, kad raidės *c, z, d, k, m, p, ž* procentiškai visų autorių kūryboje vartojamos vienodai dažnai ir visiškai netinka autorių identifikacijai. Patikrinus polinominio skirstinio hipotezes nustatyta, kad galima išskirti 2 autorių.

SUMMARY

This master's thesis aims to determine the identity of different prose creators. By analyzing five books by different authors. Several tasks were set to achieve the final goal. The first task was to determine which linguistic features of the works in question are suitable for statistical analysis. To accomplish this task, using software R, I prepared the texts from the books. This was completed by separating the words and punctuation marks, dividing the sentences into words and letters, counting the words and letters, determining the frequency of repeated words and letters. I also performed statistical analysis with the obtained data for repetitive words and sentences of different lengths for each author. The second task consisted of investigating which methods of mathematical statistics are suitable for achieving the overall goal. I investigate that (lingvo) statistical data analysis, correlation analysis, one-way analysis of variance, and computer statistics methods can be used to determine the identity of different prose writers. The third task was to find out how the authors differ from each other by applying the chosen methods of mathematical statistics. We managed to find out that the letters t, h, ė, and u can be used to identify the fifth author. The letter č helps identify the fourth author. It was also noted that the letters c, z, d, k, m, p, and ž are used as a percentage in the work of all authors equally and are completely unsuitable for the identification of any specific author. After testing the hypotheses of the polynomial distribution, it was found that 2 authors can be distinguished by their writing style.

1 priedas

Kiekvieno autoriaus pasikartojančių žodžių skaičius

	1 autorius	2 autorius	3 autorius	4 autorius	5 autorius	Bendrai
1 kartą	725	753	570	783	1082	3913
2 kartus	92	108	92	90	110	492
3 kartus	32	36	39	33	34	174
4 kartus	19	11	10	14	15	69
5 kartus	3	11	7	7	5	33
6 kartus	9	7	4	10	7	37
7 kartus	4	4	7	3	6	24
8 kartus	3	4	2	3	3	15
9 kartus	4	5	0	2	2	13
10 kartus	4	2	5	3	2	16
11 kartų	2	3	1	1	3	10
12 kartų	1	1	1	1	1	5
13 kartų	3	3	1	1	2	10
14 kartų	1	1	0	0	2	4
15 kartų	2	0	2	0	0	4
16 kartų	0	1	1	0	1	3
17 kartų	0	0	1	3	0	4
18 kartų	0	1	1	1	0	3
19 kartų	1	0	0	0	0	1
20 kartų	0	1	0	0	0	1
21 kartą	0	0	0	1	0	1
22 kartus	0	1	0	1	1	3
29 kartus	0	0	0	1	0	1
30 kartų	0	0	0	1	0	1
34 kartus	1	0	0	0	0	1
36 kartus	1	0	0	0	0	1
46 kartus	0	0	1	0	0	1
48 kartus	0	0	0	0	1	1
51 kartą	0	0	0	1	0	1
56 kartus	1	0	0	0	0	1
66 kartus	0	1	0	0	0	1

2 priedas

Dažniausiai vartojami žodžiai

- jau, jį, lyg, vis – 20 kartų
- kai – 21 kartą
- nors – 22 kartus
- tiek – 23 kartus
- kas, nieko, prie – 24 kartus
- dar, jo – 25 kartus
- čia – 26 kartus
- ar, aš, man, gal – 28 kartus
- nuo, Beatričė – 29 kartus
- jos – 32 kartus
- po – 33 kartus
- Timis – 34 kartus
- ką – 36 kartus
- jis – 37 kartus
- taip – 43 kartus
- ant – 44 kartus
- ne – 50 kartų
- su – 53 kartus
- kaip – 54 kartus
- tai – 55 kartus
- o – 59 kartus
- ji – 60 kartų
- iš – 62 kartus
- buvo – 63 kartus
- tik – 66 kartus
- bet – 67 kartus
- kad – 78 kartus
- į – 109 kartus
- ir – 267 kartus

3 priedas

Koreliacijos koeficientų lygybė tarp žodžių ir raidžių skaičiaus sakinyje

$$\begin{cases} H_0\text{: 1 ir 3 autoriaus koreliacijos koeficientai lygūs,} \\ H_1\text{: 1 ir 3 autoriaus koreliacijos koeficientai nelygūs.} \end{cases}$$

$r_1 = 0{,}960, z_1 = 1{,}946, n = 200;$

$r_2 = 0{,}940, z_2 = 1{,}738, m = 200.$

$$Z = \frac{z_1 - z_2}{\sqrt{\frac{1}{n-3} + \frac{1}{m-3}}} = \frac{1{,}946 - 1{,}738}{\sqrt{\frac{1}{200-3} + \frac{1}{200-3}}} = \frac{0{,}208}{\sqrt{\frac{2}{197}}} = \frac{0{,}208}{0{,}101} = 2{,}059.$$

Kadangi $|z| = 2{,}059 > 1{,}959 = z_{0{,}005}$, tai H_0 atmetama.

$$\begin{cases} H_0\text{: 1 ir 5 autoriaus koreliacijos koeficientai lygūs,} \\ H_1\text{: 1 ir 5 autoriaus koreliacijos koeficientai nelygūs.} \end{cases}$$

$r_1 = 0{,}960, z_1 = 1{,}946, n = 200;$

$r_2 = 0{,}960, z_2 = 1{,}946, m = 200.$

$$Z = \frac{z_1 - z_2}{\sqrt{\frac{1}{n-3} + \frac{1}{m-3}}} = \frac{1{,}946 - 1{,}946}{\sqrt{\frac{1}{200-3} + \frac{1}{200-3}}} = \frac{0}{\sqrt{\frac{2}{197}}} = 0.$$

Kadangi $|z| = 0 \leq 1{,}959 = z_{0{,}005}$, tai H_0 neatmetama.

$$\begin{cases} H_0\text{: 2 ir 3 autoriaus koreliacijos koeficientai lygūs,} \\ H_1\text{: 2 ir 3 autoriaus koreliacijos koeficientai nelygūs.} \end{cases}$$

$r_1 = 0{,}975, z_1 = 2{,}185, n = 200;$

$r_2 = 0{,}940, z_2 = 1{,}738, m = 200.$

$$Z = \frac{z_1 - z_2}{\sqrt{\frac{1}{n-3} + \frac{1}{m-3}}} = \frac{2{,}185 - 1{,}738}{\sqrt{\frac{1}{200-3} + \frac{1}{200-3}}} = \frac{0{,}447}{\sqrt{\frac{2}{197}}} = \frac{0{,}447}{0{,}101} = 4{,}426.$$

Kadangi $|z| = 4{,}426 > 1{,}959 = z_{0{,}005}$, tai H_0 atmetama.

$$\begin{cases} H_0\text{: 2 ir 4 autoriaus koreliacijos koeficientai lygūs,} \\ H_1\text{: 2 ir 4 autoriaus koreliacijos koeficientai nelygūs.} \end{cases}$$

$r_1 = 0{,}975, z_1 = 2{,}185, n = 200;$

$r_2 = 0{,}958, z_2 = 1{,}946, m = 200.$

$$Z = \frac{z_1 - z_2}{\sqrt{\frac{1}{n-3} + \frac{1}{m-3}}} = \frac{2{,}185 - 1{,}946}{\sqrt{\frac{1}{200-3} + \frac{1}{200-3}}} = \frac{0{,}239}{\sqrt{\frac{2}{197}}} = \frac{0{,}239}{0{,}101} = 2{,}366.$$

Kadangi $|z| = 2{,}366 > 1{,}959 = z_{0{,}005}$, tai H_0 atmetama.

$$\begin{cases} H_0\text{: 2 ir 5 autoriaus koreliacijos koeficientai lygūs,} \\ H_1\text{: 2 ir 5 autoriaus koreliacijos koeficientai nelygūs.} \end{cases}$$

$r_1 = 0{,}975, z_1 = 2{,}185, n = 200;$

$r_2 = 0{,}960, z_2 = 1{,}946, m = 200.$

$$Z = \frac{z_1 - z_2}{\sqrt{\frac{1}{n-3} + \frac{1}{m-3}}} = \frac{2{,}185 - 1{,}946}{\sqrt{\frac{1}{200-3} + \frac{1}{200-3}}} = \frac{0{,}239}{\sqrt{\frac{2}{197}}} = \frac{0{,}239}{0{,}101} = 2{,}366.$$

Kadangi $|z| = 2{,}366 > 1{,}959 = z_{0{,}005}$, tai H_0 atmetama.

$$\begin{cases} H_0\text{: 3 ir 4 autoriaus koreliacijos koeficientai lygūs,} \\ H_1\text{: 3 ir 4 autoriaus koreliacijos koeficientai nelygūs.} \end{cases}$$

$r_1 = 0{,}940, z_1 = 1{,}738, n = 200;$

$r_2 = 0{,}958, z_2 = 1{,}946, m = 200.$

$$Z = \frac{z_1 - z_2}{\sqrt{\frac{1}{n-3} + \frac{1}{m-3}}} = \frac{1{,}738 - 1{,}946}{\sqrt{\frac{1}{200-3} + \frac{1}{200-3}}} = \frac{-0{,}208}{\sqrt{\frac{2}{197}}} = \frac{-0{,}208}{0{,}101} = -2{,}060.$$

Kadangi $|z| = 2{,}060 > 1{,}959 = z_{0{,}005}$, tai H_0 atmetama.

$$\begin{cases} H_0\text{: 3 ir 5 autoriaus koreliacijos koeficientai lygūs,} \\ H_1\text{: 3 ir 5 autoriaus koreliacijos koeficientai nelygūs.} \end{cases}$$

$r_1 = 0{,}940, z_1 = 1{,}738, n = 200;$

$r_2 = 0{,}960, z_2 = 1{,}946, m = 200.$

$$Z = \frac{z_1 - z_2}{\sqrt{\frac{1}{n-3} + \frac{1}{m-3}}} = \frac{1{,}738 - 1{,}946}{\sqrt{\frac{1}{200-3} + \frac{1}{200-3}}} = \frac{-0{,}208}{\sqrt{\frac{2}{197}}} = \frac{-0{,}208}{0{,}101} = -2{,}060.$$

Kadangi $|z| = 2{,}060 > 1{,}959 = z_{0{,}005}$, tai H_0 atmetama.

$$\begin{cases} H_0\text{: 4 ir 5 autoriaus koreliacijos koeficientai lygūs,} \\ H_1\text{: 4 ir 5 autoriaus koreliacijos koeficientai nelygūs.} \end{cases}$$

$r_1 = 0{,}958, z_1 = 1{,}946, n = 200;$

$r_2 = 0{,}960, z_2 = 1{,}946, m = 200.$

$$Z = \frac{z_1 - z_2}{\sqrt{\frac{1}{n-3} + \frac{1}{m-3}}} = \frac{1{,}946 - 1{,}946}{\sqrt{\frac{1}{200-3} + \frac{1}{200-3}}} = \frac{0}{\sqrt{\frac{2}{197}}} = 0.$$

Kadangi $|z| = 0 \leq 1{,}959 = z_{0{,}005}$, tai H_0 neatmetama.

1 ir 3 autorių palyginimas:

$$\widehat{p_1} = \frac{1127}{8705} = 0{,}12947;$$

$$\widehat{p_2} = \frac{905}{7040} = 0{,}12855;$$

$$\bar{p} = \frac{2032}{15745} = 0{,}12906;$$

$$Z = \frac{\widehat{p_1} - \widehat{p_2}}{\sqrt{\bar{p}(1-\bar{p})\left(\frac{1}{n}+\frac{1}{m}\right)}} = \frac{0{,}12947 - 0{,}12855}{\sqrt{0{,}129(1-0{,}129)\left(\frac{1}{8705}+\frac{1}{7040}\right)}} =$$

$$= \frac{0{,}00092}{\sqrt{0{,}1124(0{,}00011+0{,}00014)}} = \frac{0{,}00092}{\sqrt{0{,}0000281}} = \frac{0{,}00092}{0{,}0053} = 0{,}174;$$

$$z_{\frac{\alpha}{2}} = z_{\frac{0{,}05}{2}} = z_{0{,}025} = 1{,}959.$$

Kadangi $|Z| = 0{,}174 \leq 1{,}959 = z_{0{,}025}$, tai H_0 neatmetama.

1 ir 4 autorių palyginimas:

$$\widehat{p_1} = \frac{1127}{8705} = 0{,}12947;$$

$$\widehat{p_2} = \frac{1191}{9147} = 0{,}13021;$$

$$\bar{p} = \frac{2318}{17852} = 0{,}12985;$$

$$Z = \frac{\widehat{p_1} - \widehat{p_2}}{\sqrt{\bar{p}(1-\bar{p})\left(\frac{1}{n}+\frac{1}{m}\right)}} = \frac{0{,}12947 - 0{,}13021}{\sqrt{0{,}130(1-0{,}130)\left(\frac{1}{8705}+\frac{1}{9147}\right)}} =$$

$$= \frac{-0{,}00074}{\sqrt{0{,}1131(0{,}00011+0{,}00011)}} = \frac{-0{,}00074}{\sqrt{0{,}0000248}} = \frac{-0{,}00074}{0{,}0050} = -0{,}148;$$

$$z_{\frac{\alpha}{2}} = z_{\frac{0{,}05}{2}} = z_{0{,}025} = 1{,}959.$$

Kadangi $|Z| = 0{,}148 \leq 1{,}959 = z_{0{,}025}$, tai H_0 neatmetama.

1 ir 5 autorių palyginimas:

$$\widehat{p_1} = \frac{1127}{8705} = 0{,}12947;$$

$$\widehat{p_2} = \frac{1413}{11198} = 0{,}12618;$$

$$\bar{p} = \frac{2540}{19903} = 0{,}12762;$$

$$Z = \frac{\widehat{p_1} - \widehat{p_2}}{\sqrt{\bar{p}(1-\bar{p})\left(\frac{1}{n}+\frac{1}{m}\right)}} = \frac{0{,}12947 - 0{,}12618}{\sqrt{0{,}128(1-0{,}128)\left(\frac{1}{8705}+\frac{1}{11198}\right)}} =$$

$$= \frac{0{,}00329}{\sqrt{0{,}112(0{,}00011+0{,}00009)}} = \frac{0{,}00329}{\sqrt{0{,}0000224}} = \frac{0{,}00329}{0{,}0047} = 0{,}7;$$

$z_{\frac{\alpha}{2}} = z_{\frac{0{,}05}{2}} = z_{0{,}025} = 1{,}959.$

Kadangi $|Z| = 0{,}7 \leq 1{,}959 = z_{0{,}025}$, tai H_0 neatmetama.

2 ir 4 autorių palyginimas:

$$\widehat{p_1} = \frac{1245}{8938} = 0{,}13929;$$

$$\widehat{p_2} = \frac{1191}{9147} = 0{,}13021;$$

$$\bar{p} = \frac{2436}{18085} = 0{,}13470;$$

$$Z = \frac{\widehat{p_1} - \widehat{p_2}}{\sqrt{\bar{p}(1-\bar{p})\left(\frac{1}{n}+\frac{1}{m}\right)}} = \frac{0{,}13929 - 0{,}13021}{\sqrt{0{,}135(1-0{,}135)\left(\frac{1}{8938}+\frac{1}{9147}\right)}} = \frac{0{,}00908}{\sqrt{0{,}117(0{,}00011+0{,}00011)}}$$

$$= \frac{0{,}00908}{\sqrt{0{,}000026}} = \frac{0{,}00908}{0{,}0051} = 1{,}780;$$

$z_{\frac{\alpha}{2}} = z_{\frac{0{,}05}{2}} = z_{0{,}025} = 1{,}959.$

Kadangi $|Z| = 1{,}780 \leq 1{,}959 = z_{0{,}025}$, tai H_0 neatmetama.

2 ir 5 autorių palyginimas:

$$\widehat{p_1} = \frac{1245}{8938} = 0{,}13929;$$

$$\widehat{p_2} = \frac{1413}{11198} = 0{,}12618;$$

$$\bar{p} = \frac{2658}{20136} = 0{,}13200;$$

$$Z = \frac{\widehat{p_1} - \widehat{p_2}}{\sqrt{\bar{p}(1-\bar{p})\left(\frac{1}{n}+\frac{1}{m}\right)}} = \frac{0{,}13929 - 0{,}12618}{\sqrt{0{,}132(1-0{,}132)\left(\frac{1}{8938}+\frac{1}{11198}\right)}} =$$

$$= \frac{0{,}01311}{\sqrt{0{,}115(0{,}00011+0{,}00009)}} = \frac{0{,}01311}{\sqrt{0{,}000023}} = \frac{0{,}01311}{0{,}0048} = 2{,}731;$$

$z_{\frac{\alpha}{2}} = z_{\frac{0{,}05}{2}} = z_{0{,}025} = 1{,}959.$

Kadangi $|Z| = 2{,}731 > 1{,}959 = z_{0{,}025}$, tai H_0 atmetama.

3 ir 4 autorių palyginimas:

$$\widehat{p_1} = \frac{905}{7040} = 0{,}12855;$$

$$\widehat{p_2} = \frac{1191}{9147} = 0{,}13021;$$

$$\bar{p} = \frac{2096}{16187} = 0{,}12949;$$

$$Z = \frac{\widehat{p_1} - \widehat{p_2}}{\sqrt{\bar{p}(1-\bar{p})\left(\frac{1}{n}+\frac{1}{m}\right)}} = \frac{0{,}12855 - 0{,}13021}{\sqrt{0{,}129(1-0{,}129)\left(\frac{1}{7040}+\frac{1}{9147}\right)}} = \frac{-0{,}00166}{\sqrt{0{,}112(0{,}00014+0{,}00011)}}$$

$$= \frac{-0{,}00166}{\sqrt{0{,}000028}} = \frac{-0{,}00166}{0{,}0053} = -0{,}313;$$

$$z_{\frac{\alpha}{2}} = z_{\frac{0{,}05}{2}} = z_{0{,}025} = 1{,}959.$$

Kadangi $|Z| = 0{,}313 \leq 1{,}959 = z_{0{,}025}$, tai H_0 neatmetama.

3 ir 5 autorių palyginimas:

$$\widehat{p_1} = \frac{905}{7040} = 0{,}12855;$$

$$\widehat{p_2} = \frac{1413}{11198} = 0{,}12618;$$

$$\bar{p} = \frac{2318}{18238} = 0{,}12710;$$

$$Z = \frac{\widehat{p_1} - \widehat{p_2}}{\sqrt{\bar{p}(1-\bar{p})\left(\frac{1}{n}+\frac{1}{m}\right)}} = \frac{0{,}12855 - 0{,}12618}{\sqrt{0{,}127(1-0{,}127)\left(\frac{1}{7040}+\frac{1}{11198}\right)}} =$$

$$= \frac{0{,}00237}{\sqrt{0{,}111(0{,}00014+0{,}00009)}} = \frac{0{,}00237}{\sqrt{0{,}000026}} = \frac{0{,}00237}{0{,}0051} = 0{,}465;$$

$$z_{\frac{\alpha}{2}} = z_{\frac{0{,}05}{2}} = z_{0{,}025} = 1{,}959.$$

Kadangi $|Z| = 0{,}465 \leq 1{,}959 = z_{0{,}025}$, tai H_0 neatmetama.

4 ir 5 autorių palyginimas:

$$\widehat{p_1} = \frac{1191}{9147} = 0{,}13021;$$

$$\widehat{p_2} = \frac{1413}{11198} = 0{,}12618;$$

$$\bar{p} = \frac{2604}{20345} = 0{,}12799;$$

$$Z = \frac{\widehat{p_1} - \widehat{p_2}}{\sqrt{\bar{p}(1-\bar{p})\left(\frac{1}{n}+\frac{1}{m}\right)}} = \frac{0{,}13021 - 0{,}12618}{\sqrt{0{,}128(1-0{,}128)\left(\frac{1}{9147}+\frac{1}{11198}\right)}} =$$

$$= \frac{0{,}00403}{\sqrt{0{,}112(0{,}00011+0{,}00009)}} = \frac{0{,}00403}{\sqrt{0{,}000022}} = \frac{0{,}00403}{0{,}0047} = 0{,}857;$$

$z_{\frac{\alpha}{2}} = z_{\frac{0{,}05}{2}} = z_{0{,}025} = 1{,}959.$

Kadangi $|Z| = 0{,}857 \leq 1{,}959 = z_{0{,}025}$, tai H_0 neatmetama.