



EUROPOS SĄJUNGA
Europos socialinis fondas



ŠVIETIMO IR MOKSLO MINISTERIJA



KURKIME ATEITĮ DRAUGE!

Projektas

**“Empirinių duomenų ir informacijos HSM tyrimams kaupimas ir valdymas : Lietuvos HSM
duomenų archyvas (LiDA)”**

SFMIS Nr. BPD2004-ESF-2.5.0-03-05/0042

**PAVYZDINIS METODOLOGINIS MOKOMASIS STUDIJŲ
PAKETAS**

STATISTINĖ KIEKYBINIŲ DUOMENŲ ANALIZĖ SU SPSS IR STATA

GALUTINĖ REDAKCIJA

(Autorinio kūrinio užsakymo sutartis Nr. A-6357, 2008-02-15)

KAUNAS, 2008

KŪRINIO UŽSAKOVAS	
Perkančioji organizacija:	Kauno technologijos universitetas
Perkančiosios organizacijos atstovas :	Algis Krupavičius Kęstučio g. 8-206, Kaunas 8 (37) 300115 pvai@ktu.lt
KŪRINIO AUTORIUS	
Kūrinio autorius	Bendra autorių grupė (pagal 2007-08-23 Jungtinės veiklos sutartį)
Bendra autorių atstovas:	Vytautas Janilionis Topolių 2-72, Kaunas 8 (686) 61605 vytautas.janilionis@ktu.lt

Autoriai:

Doc. Vytautas Janilionis

(parašas)

Dr. Vaidas Morkevičius

(parašas)

Dr. Rimantas Rauleckas

(parašas)

Galutinę autorinio kūrinio redakciją parengė:

Vaidas Morkevičius
(vardas, pavardė)

2008 m. birželio 27 d.
(data)

(parašas)

Pavyzdinį metodologinį mokomąjį studijų paketą „Statistinė kiekybinių duomenų analizė su SPSS ir STATA“ sudaro **TRYŠ DALYS**:

I DALIS. Pavyzdinė skaitmeninė duomenų bazė, į ją įtrauktų tyrimų aprašymai ir kintamųjų sąvadaai bei paaiškinimai (p. 8)

II DALIS. Įvadinio kurso į statistinę analizę mokomoji medžiaga (p. 59)

III DALIS. Statistinės analizės pavyzdžių naudojant pavyzdinę skaitmeninę duomenų bazę medžiaga (p. 205)

PATEIKIMO FORMA:	
1. Pavyzdinės skaitmeninės duomenų bazės laikmenos	Elektronine forma – CD-R (2 vnt.); formatai: SPSS, STATA ir ASCII („tab-delimited“), atitinkamai *.sav, *.dta ir *.csv laikmenos.
2. Tyrimų aprašymų ir kintamųjų sąvadų bei paaiškinimų, įvadinio kurso į statistinę analizę mokomosios medžiagos, statistinės analizės pavyzdžių naudojant pavyzdinę skaitmeninę duomenų bazę medžiagos	2 vnt. atspausdinta, 5 vnt. elektronine forma – CD-R (*.html , *.doc, *.rtf ir *.pdf formatais).

BENDRAS TURINYS

I DALIS. TYRIMŲ ĮTRAUKTŲ Į PAVYZDINĘ SKAITMENINĘ DUOMENŲ BAZĘ APRAŠYMAI IR KINTAMŲJŲ SĄVADAI BEI PAAIŠKINIMAI	8
1. ĮVADAS.....	9
2. 1990, 1996, 1999 m. EUROPOS IR PASAULIO VERTYBIŲ TYRIMŲ LIETUVOJE APRAŠYMAS IR KINTAMŲJŲ SĄVADAS BEI PAAIŠKINIMAI.....	11
2.1. EVS ir WVS tikslai ir objektas.....	11
2.2. EVS ir WVS vykdymo istorija	12
2.3. Pagrindiniai EVS ir WVS rezultatai.....	12
2.4. Teorinės EVS ir WVS analizės perspektyvos	18
2.5. 1990, 1996, 1999 m. Europos ir pasaulio vertybių tyrimų Lietuvoje kintamųjų sąvadas bei paaiškinimai	19
3. 2002, 2004, 2006 m. EUROPOS SOCIALINIO TYRIMO APRAŠYMAS IR KINTAMŲJŲ SĄVADAS BEI PAAIŠKINIMAI	22
3.1. ESS tikslai ir objektas	22
3.2. ESS vykdymo istorija	22
3.3. Pagrindiniai ESS rezultatai ir publikacijos	22
3.4. Teorinės ESS analizės perspektyvos	23
3.5. 2002, 2004, 2006 m. ESS kintamųjų sąvadas bei paaiškinimai	24
4. 2006 m. GEOKULTŪRINIŲ NUOSTATŲ TYRIMO APRAŠYMAS APRAŠYMAS IR KINTAMŲJŲ SĄVADAS BEI PAAIŠKINIMAI.....	32
4.1. GNT tikslai ir objektas.....	32
4.2. GNT vykdymo istorija.....	32
4.3. Pagrindiniai GNT rezultatai ir publikacijos.....	32
4.4. Teorinės GNT analizės perspektyvos.....	33
4.5. 2006 m. GNT kintamųjų sąvadas bei paaiškinimai.....	34
5. 2007 m. „FREEDOM IN THE WORLD“ TYRIMO APRAŠYMAS IR KINTAMŲJŲ SĄVADAS BEI PAAIŠKINIMAI	45
5.1. LPT tikslai ir objektas.....	45
5.2. LPT vykdymo istorija	45
5.3. Pagrindiniai LPT rezultatai ir publikacijos.....	45
5.4. Teorinės LPT analizės perspektyvos	46
5.5. 2007 m. LPT kintamųjų sąvadas bei paaiškinimai.....	47
6. 2004 m. „NAUJOJO BALTIJOS BAROMETRO“ TYRIMO APRAŠYMAS IR KINTAMŲJŲ SĄVADAS BEI PAAIŠKINIMAI	50
6.1. NBB tyrimo tikslai ir objektas.....	50
6.2. NBB tyrimo vykdymo istorija	50
6.3. Pagrindiniai NBB tyrimo rezultatai ir publikacijos.....	50
6.4. Teorinės NBB tyrimo analizės perspektyvos.....	50
6.5. 2004 m. NBB kintamųjų sąvadas bei paaiškinimai	52
7. LITERATŪRA.....	58
II DALIS. ĮVADINIO KURSO Į STATISTINĘ ANALIZĘ MOKOMOJI MEDŽIAGA	59
1. ĮVADAS	60
1. APRAŠOMOJI STATISTIKA	61
1.1. Dažniai	61
1.2. Empirinė skirstinio funkcija.....	63
1.3. Skaitinės duomenų charakteristikos.....	64
1.3.1. Padėties charakteristikos.....	65

1.3.2. Sklaidos charakteristikos	67
1.3.3. Asimetrijos ir eksceso koeficientai	68
1.4. Stačiakampė diagrama.....	69
1.5. Diagrama STIEBAS ir LAPAS	70
2. PASIKLIAUTINIEJI INTERVALAI.....	72
2.1. Pagrindinės sąvokos	72
2.2. Parametrų taškiniai įverčiai	74
2.3. Pasikliautinių intervalų skaičiavimas.....	75
2.4. Normaliojo skirstinio vidurkio pasikliautinojo intervalo radimo pavyzdys.....	78
2.5. Normaliojo skirstinio vidurkio pasikliautinojo intervalo radimo pavyzdys.....	80
2.6. Įvykio tikimybės pasikliautinojo intervalo radimo pavyzdys	82
3. HIPOTEZIŲ TIKRINIMAS.....	83
3.1. Pagrindinės sąvokos	83
3.2. Hipotezių tikrinimo klaidos	84
3.3. Bendroji hipotezės tikrinimo schema	85
3.4. p-reikšmė ir kriterijaus galia.....	86
3.5. Išvadų formulavimas.....	88
3.6. Hipotezės apie populiacijos vidurkį tikrinimas	89
3.7. Hipotezės apie proporciją tikrinimas	91
3.8. Suderinamumo hipotezių tikrinimas	92
3.8.1. Kolmogorovo Smirnovio suderinamumo kriterijus.....	94
3.8.2. Chi kvadratu suderinamumo kriterijus	95
4. VIDURKIŲ PALYGINIMAS. DISPERSINĖ ANALIZĖ	99
4.1. Hipotezės dėl dviejų populiacijų vidurkių skirtumo tikrinimas	99
4.1.2. Dviejų nepriklausomų imčių T-kriterijus	99
4.1.2. Mano-Vitnio-Vilkoksono rangų sumų kriterijus dviems nepriklausomoms imtims.....	102
4.2. Vienfaktorinė dispersinė analizė	104
4.2.1. Vienfaktorinės dispersinės analizės modeliai.....	104
4.2.2. Pilnas randomizuotas vienfaktorinės dispersinės analizės modelis.....	105
4.2.3. Daugkartiniai lyginimai.....	107
4.2.4. Vienfaktorinės dispersinės analizės pavyzdys.....	108
4.3. Kruskalo-Voliso kriterijus k-nepriklausomoms imtims	111
5. POŽYMIŲ PRIKLAUSOMUMO TYRIMAS. KORELIACINĖ ANALIZĖ.....	113
5.1. Pirsono koreliacijos koeficientas	113
5.2. Spirmeno ranginės koreliacijos koeficientas.....	115
5.3. Požymių priklausomumo lentelės	115
5.4. Požymių nepriklausomumo hipotezės tikrinimas.....	117
5.5. Ryšio stiprumo matai tvarkos skalėje.....	118
5.6. Ryšio stiprumo matai vardų skalėje	120
6. REGRESINĖ ANALIZĖ	122
6.1. Regresinės analizės uždaviniai	122
6.2. Regresijos funkcijos koeficientai.....	123
6.2.1. Koeficientų taškiniai ir intervaliniai įverčiai	123
6.2.2. Hipotezių apie koeficientų lygybę nuliui tikrinimas	125
6.3. Standartinė regresijos paklaida ir apibrėžtumo koeficientas.....	127
6.4. Prognozavimo pasikliautinieji intervalai	130
6.5. Regresijos prielaidų tikrinimas	130
6.6. Regresinės analizės taikymo klaidos.....	133
6.7. Regresinės analizės pavyzdžiai.....	135
6.7.1. Išlaidų reklamai analizė	135
6.7.2. Kauno miesto butų kainų analizė.....	142
7. KLAUSTERINĖ ANALIZĖ	148

7.1. Klasifikavimo uždaviniai	148
7.2. Klasterinės analizės modelis	150
7.3. Artumo matai	152
7.3.1. Artumo matai intervalų ir santykių skalėse	152
7.3.2. Dvireikšmių vektorių artumo matai	153
7.4. Klasterių sudarymo metodai	155
7.5. Klasifikavimo kokybės kriterijai	158
7.6. Hierarchiniai klasifikavimo algoritmai	159
7.7. Klasterinės analizės pavyzdys	159
8. FAKTORINĖ ANALIZĖ	167
8.1. Įvadas	167
8.1. Faktorinės analizės modelis	168
8.2. Faktorinės analizės metodai	170
8.3. Faktorinės analizės etapai	170
8.3.1. Aprašomoji statistika ir duomenų tikamumas faktorinei analizei	171
8.3.2. Faktorių išskyrimas	171
8.3.3. Faktorių skaičiaus nustatymas	172
8.3.4. Faktorių sukimas ir interpretavimas	175
8.4. Faktorinės analizės pavyzdys	176
9. DAUGIAMATČIŲ SKALIŲ METODAS (MDS)	186
9.1. Įvadas	186
9.2. MDS modeliai	187
9.3. MDS etapai	189
9.3.1. Duomenų rinkimas	189
9.3.2. Modelio parinkimas	190
9.3.3. Erdvės dimensijų skaičiaus nustatymas	192
9.3.4. Rezultatų interpretavimas	193
9.3.5. Modelio tinkamumo tyrimas	193
9.4. MDS pavyzdys	193
LITERATŪRA	203
III DALIS. STATISTINĖS ANALIZĖS PAVYZDŽIŲ NAUDOJANT PAVYZDINĘ SKAITMENINĘ DUOMENŲ BAZĘ MEDŽIAGA	205
1. ĮVADAS	206
2. APRAŠOMOJI STATISTIKA	207
2.1. Dažniai, intervaliniai dažniai ir santykiniai dažniai	208
2.2. Duomenų sklaidos, padėties ir formos charakteristikos	234
3. PASIKLIAUTINIEJI INTERVALAI	242
3.1. Populiacijos vidurkio pasikliautinis intervalas	242
3.2. Populiacijos proporcijos (tikimybės) pasikliautinis intervalas	250
4. HIPOTEZIŲ TIKRINIMAS	258
4.1. Hipotezių apie populiacijos vidurkį tikrinimas	260
4.2. Suderinamumo hipotezių tikrinimas	263
4.3. Hipotezių apie populiacijos proporciją tikrinimas	266
5. VIDURKIŲ PALYGINIMAS. DISPERSINĖ ANALIZĖ (ANOVA)	276
5.1. Vidurkių palyginimo analizė	276
5.1.1. Vidurkių palyginimas su SPSS	276
5.1.2. Vidurkių palyginimas su Stata	278
5.2. Dispersinė analizė (ANOVA)	280
5.2.1. Dispersinė analizė su SPSS	280
5.2.2. Dispersinė analizė su Stata	286
5.3. Neparametriniai kriterijai	289

5.3.1. Mano-Vitnio-Vilkoksono rangų sumų kriterijus dviems nepriklausomoms imtims su SPSS	290
5.3.2. Mano-Vitnio-Vilkoksono rangų sumų kriterijus dviems nepriklausomoms imtims su Stata	292
5.3.3. Kraskelo-Voliso kriterijus k-nepriklausomoms imtims su SPSS	296
5.3.4. Kraskelo-Voliso kriterijus k-nepriklausomoms imtims su Stata	297
6. POŽYMIŲ PRIKLAUSOMUMO TYRIMAS. KORELIACINĖ ANALIZĖ	300
6.1. Intervalų skalės koreliacinė analizė	300
6.1.1. Intervalų skalės koreliacijos koeficientai su SPSS	300
6.1.2. Intervalų skalės koreliacijos koeficientai su Stata	303
6.2. Požymių priklausomumo lentelės. Požymių nepriklausomumo hipotezės tikrinimas (Chi kvadratu kriterijus)	307
6.2.1. Požymių priklausomumo 2*2 lentelės su SPSS	307
6.2.2. Požymių priklausomumo 2*2 lentelės su Stata	309
6.3. Ryšio stiprumo matai tvarkos skalėje.....	311
6.3.1. Ryšio stiprumo matai tvarkos skalėje su SPSS	311
6.3.2. Ryšio stiprumo matai tvarkos skalėje su Stata	313
6.4. Ryšio stiprumo matai vardų skalėje	315
6.4.1. Ryšio stiprumo matai vardų skalėje su SPSS	315
6.4.2. Ryšio stiprumo matai vardų skalėje su Stata	320
7. REGRESINĖ ANALIZĖ	322
7.1. Daugialypė tiesinė regresinė analizė	322
7.1.1. Daugialypė tiesinė regresinė analizė su SPSS	322
7.1.2. Daugialypė tiesinė regresinė analizė su Stata	330
8. KLASTERINĖ ANALIZĖ	338
8.1. Hierarchinė klasterinė analizė.....	338
8.1.1. Hierarchinė klasterinė analizė su SPSS	338
8.1.2. Hierarchinė klasterinė analizė su Stata	343
9. FAKTORINĖ ANALIZĖ	348
9.1. Tiriamoji faktorinė analizė	348
9.1.1. Tiriamoji faktorinė analizė su SPSS	348
9.1.2. Tiriamoji faktorinė analizė su Stata	356
10. DAUGIAMAČIŲ SKALIŲ ANALIZĖ (MDS)	364
10.1. Klasikinio MDS modelio pavyzdys su Stata	366
10.2. Moderniojo nemetrinio MDS modelio pavyzdys su SPSS ir Stata	373
11. LITERATŪRA	393

**I DALIS. TYRIMŲ ĮTRAUKTŲ Į PAVYZDINĘ
SKAITMENINĘ DUOMENŲ BAZĘ
APRAŠYMAI IR KINTAMŲJŲ SĄVADAI BEI
PAAIŠKINIMAI**

Autorius:

Dr. Rimantas Rauleckas

(parašas)

Kaunas, 2008

1. ĮVADAS

Šioje dalyje pateikiami į pavyzdinę skaitmeninę duomenų bazę įtrauktų tyrimų aprašymai, iš tų tyrimų atrinktų kintamųjų sąvadaai ir savarankiškai duomenų analizei reikalingi paaiškinimai.

Į pavyzdinę skaitmeninę duomenų bazę atrinkti kintamieji iš šių tyrimų:

1990, 1996, 1999 m. Europos ir pasaulio vertybių tyrimų Lietuvoje,

2002, 2004, 2006 m. Europos socialinio tyrimo,

2006 m. Lietuvos gyventojų geokultūrinių nuostatų tyrimo,

2007 m. Laisvės lygio pasaulio šalyse tyrimo bei

2004 m. Naujojo Baltijos barometro tyrimo.

Tyrimų duomenys pateikti trimis populiariais formatais – SPSS, Stata ir tekstiniu ASCII („tab-delimited“) – atskirose laikmenose. Laikmenų patalpintų į pridamus CD-R (2 vnt. elektronine forma) sąrašas:

Laikmenos	Tyrimai
EVS-WVS_1990-1999_LT.dat EVS-WVS_1990-1999_LT.dta EVS-WVS_1990-1999_LT.sav	1990, 1996, 1999 m. Europos ir pasaulio vertybių tyrimų Lietuvoje duomenys
ESS1_ed06_LT.dat ESS1_ed06_LT.dta ESS1_ed06_LT.sav	2002 m. Europos socialinio tyrimo duomenys
ESS2_e03_LT.dat ESS2_e03_LT.dta ESS2_e03_LT.sav	2004 m. Europos socialinio tyrimo duomenys
ESS3_e03.1_LT.dat ESS3_e03.1_LT.dta ESS3_e03.1_LT.sav	2006 m. Europos socialinio tyrimo duomenys
geokult_2006.dat geokult_2006.dta geokult_2006.sav	2006 m. Lietuvos gyventojų geokultūrinių nuostatų tyrimo duomenys
FH_2007.dat FH_2007.dta FH_2007.sav	2007 m. Laisvės lygio pasaulio šalyse tyrimo duomenys
NBB_2004_LT.dat NBB_2004_LT.dta	2004 m. Naujojo Baltijos barometro tyrimo duomenys

Į tyrimų aprašymus įtraukta informacija apie kiekvieno tyrimo tikslus, objektą, vykdymo istoriją, pagrindinius rezultatus bei teorinės analizės perspektyvas.

Taip pat pateikiami iš tyrimų duomenų atrinktų kintamųjų sąvadai bei savarankiškai vartotojo statistinei analizei reikalingi paaiškinimai:

kintamųjų kiekis,

kintamųjų vardai: kintamųjų identifikaciniai pavadinimai,

kintamųjų žymės (pavadinimai): trumpi kintamųjų pavadinimai nusakantys kintamųjų prasmę,

kintamųjų reikšmių žymės:

kintamųjų įgaunamų reikšmių aibė skaitine ar tekstine forma, kuri atitinka duomenų bazėje įvestus kodus,

reikšmės kodams suteikti aprašymai, nusakantys reikšmių ir joms priskirtų kodų prasmę,

kintamųjų prasmė ir turinys:

klausimyne pateikti klausimai,

trumpi kintamųjų prasmės paaiškinimai, sutampantys arba praplečiantys kintamųjų pavadinimuose pateiktą informaciją; indeksų ar skalių atveju paaiškinama vidinė kintamųjų struktūra, pvz., 6 dalykų (šeima, darbas, laisvalaikis, draugai, religija, politika) ranginis vertinimas (nuo 1 – svarbiausias iki 6 – mažiausiai svarbus),

bendresni konstruktai ar klasės, kurioms priklauso kintamieji,

duomenų analizei reikalingi paaiškinimai:

kintamųjų skalės tipai,

trūkstamų reikšmių žymėjimas,

kiti su duomenų analize susiję komentarai.

Tyrimų aprašymus ir kintamųjų sąvadus bei paaiškinimus sudaro 7 skyriai, kurių bendra apimtis 50 puslapių.

2. 1990, 1996, 1999 M. EUROPOS IR PASAULIO VERTYBIŲ TYRIMŲ LIETUVOJE APRAŠYMAS IR KINTAMŲJŲ SĄVADAS BEI PAAIŠKINIMAI

Šiame skyriuje pateikiamas Europos ir pasaulio vertybių tyrimų Lietuvoje (toliau – EVS ir WVS) 1990, 1996, 1999 m. aprašymas: tyrimų tikslai, objektas, vykdymo istorija, pagrindiniai rezultatai ir teorinės perspektyvos. Be to, pateikiamas šių tyrimų kintamųjų sąvadas bei paaiškinimai: kintamųjų kiekis, kintamųjų vardai, kintamųjų žymės, kintamųjų reikšmių žymės, kintamųjų prasmė ir turinys, duomenų analizei reikalingi paaiškinimai.

2.1. EVS ir WVS tikslai ir objektas

Europos ir pasaulio vertybių tyrimus inicijavę mokslininkai siekia empirinių duomenų pagalba atskleisti Europos ir pasaulio gyventojų nuomones, nuostatas, vertybes bei palyginti jų kaitą tarp šalių ir laike. Europos vertybių tyrimu siekta įvertinti pagrindinių socialinių vertybių Europoje skirtumus ir kaitą, empirikai įvertinti, ar atsiranda vieningas kultūrinis Europos identitetas. Pirmasis EVS kėlė bendrus klausimus, ar europiečiai turi homogenišką ir ilgainiui išliekanti vertybių rinkinį, ar vertybės (ypač krikščioniškos vertybės ir praktikos) Europoje kinta ir kokiomis kryptimis, kokios šių procesų pasekmės Europos vienybei.

EVS objektas – šiuolaikinės Europos gyventojų nuomonės, įsitikinimai, nuostatos ir vertybės. Šios orientacijos matuojamos standartizuotais klausimynais svarbiausiose gyvenimo srityse: religija, moralė, politika, demokratija, ekonomika, technologija, tradicinės vertybės, gyvenimo stilius, ekologija, šeima, vedybos, vyro ir moters vaidmenys, darbas, asmeniniai finansai.

Respondentų buvo klausama, ar tokie veiksmai kaip savižudybė, mokesčių vengimas, melavimas, eutanazija, skyrybos ir abortai gali būti pateisinami. Respondentai atsakinėjo apie grupes ir asociacijas, kurioms jie priklauso, kuriose laisvanoriškai dirbo, žmonių grupes, kurių nenorėtų turėti kaimynystėje, savo sveikatos būklę, įsitikinimus apie tai, ar jie turi laisvą pasirinkimą ir kontroliuoja savo gyvenimą. Nemažai klausimų klausama apie gyvenimo prasmę ir tikslus, tokių kaip respondentų požiūris į mokslo pasiekimų vertę, gėrio ir blogio perskyrą, religinius įsitikinimus ir elgesį. Respondentų klausama apie jų nuostatas religijos, moralės, politikos, seksualinės laisvės, vedybinio gyvenimo, vienišo tėvo/motinos, vaikų auginimo atžvilgiu ir darbo, šeimos, politikos ir religijos svarbą jų gyvenime. Klausimai, susiję su darbu apėmė svarbiausius finansinius ir socialinius motyvatorius darbe, pasididžiavimą darbine veikla, pasitenkinimą užimama padėtimi darbe, požiūrius į savininkų/valstybės/darbuotojų vykdomą verslo valdymą. Taip pat, klausama apie pasaulinės ekonomikos stabilumą ir ar respondentai buvo patenkinti savo finansine situacija. Tyrime teiraujamasi respondentų nuomones apie įvairias politinio veiksmo formas, svarbiausius šalies tikslus, pasitikėjimą įvairiomis pilietinėmis ir valstybinėmis institucijomis, ar eitų į karą dėl savo šalies. Demografinė informacija apima šeimos pajamas, asmenų namuose skaičių, vietovės dydį, gyvenamąją vietą, respondento amžių, lytį, užsiėmimą, išsilavinimą, religiją ir religingumą, priklausymą politinėms partijoms ir profesinėms sąjungoms, saviidentifikaciją politinės kairės/dešinės skalėje.

Tyrimo populiacija – suaugę asmenys, turintys 18 ir daugiau metų apie 80-tyje visuomenių, atstovaujančių apie 60 pasaulio šalių. Kiekvienoje šalyje apklausta bent 1000 respondentų. Daugumoje šalių buvo naudojamas stratifikuotas daugelio stadijų atsitiktinis tikimybinis imties atrankos būdas (Ronald Inglehart and Inter-university Consortium for Political and Social Research 2004). Lietuvoje, nepatikrintais duomenimis, naudotas daugelio stadijų lizdinis (cluster) imties sudarymo būdas pirmame (vietovių Lietuvoje) atrankos etape, atsitiktinio maršruto antrame (namų ūkių) atrankos etape, taip pat kvotinis ir Kisho lentelių atrankos būdai paskutiniuose (namų ūkių narių) atrankos etapuose. Su respondентаis buvo atliekamas asmeninis interviu. Pirmąjį tyrimą Lietuvoje 1990 m. birželio-rugpjūčio mėnesiais atliko Vilniaus universiteto Sociologijos

laboratorija (atsakingas asmuo Rasa Ališauskienė), apklausta N=1000 respondentų. Antrąjį tyrimą Lietuvoje 1996 m. spalio-lapkričio mėnesiais atliko UAB „Baltijos Tyrimai“ (atsakingas asmuo Rasa Ališauskienė), apklausta N=1,009 respondentų. Trečiasis tyrimas atliktas UAB „Baltijos Tyrimai“ 1999 m. lapkričio 5 d. – gruodžio 14 d., apklausta N=1,018 respondentų, atsakingi asmenys Stanislovas Juknevičius (Lietuvos kultūros ir meno institutas) ir Rasa Ališauskienė (Socialinių tyrimų institutas) (Halman 2001).

2.2. EVS ir WVS vykdymo istorija

Pirmoji EVS banga vykdyta 1981 m. dešimtyje vakarų Europos šalių, dėl didelio susidomėjimo kiek vėliau išaugusi iki 26 šalių grupės. EVS grupei vadovavo Jan Kerkhofs ir Ruud de Moor, patariamasis komitetas įtraukė Gordon Heald, Juan Linz, Elisabeth Noelle-Neumann, Jacques Rabier ir Helene Riffault. Šių tyrimų rezultatai parodė, kad tikėtina, jog vyksta dideli ir nuspėjami kultūriniai pokyčiai šiose šalyse.

Antroji EVS banda prasidėjo 1990 m., ja siekta stebėti aptiktus pokyčius geografiškai platesnėje perspektyvoje. EVS vyko visose Europos šalyse, įtraukiant Rytų ir Centrinės Europos šalis (pirmą kartą įtraukta ir Lietuva), o taip pat JAV ir Kanadą. Tyrimo vykdomasis komitetas susidarė iš Ruud de Moor (pirmininkas), Jan Kerkhofs (pirmininko pavaduotojas), Karel Dobbelaere, Loek Halman, Stephen Harding, Felix Heunks, Ronald Inglehart, Renate Koecher, Jacques Rabier ir Noel Timms. Ronald Inglehart (Mičigano universitetas) organizavo ir koordinavo tyrimus ne Europos šalyse ir keliuose Rytų Europos šalyse: ši banga yra kartu ir pirmoji Pasaulio vertybių tyrimo (WVS, angl. WVS-World Values Survey) banga. Iš viso šioje pirmoje WVS bangoje dalyvavo apie 40 šalių iš visų gyvenamų kontinentų. 1993 m. spalį tyrėjai susirinko Ispanijoje įvertinti pirmų dviejų tyrimų bangų rezultatus. Nustačius gana aiškius kaitos požymius tarp 1981 ir 1990, nuspręsta tęsti stebėseną ir atlikti papildomas tyrimų bangas ir pradėjo ruošti 1995 m. tyrimų bangai.

1995 m. vykdyta trečioji tyrimų ir antroji WVS banga. Šios tyrimų bangos metu ypač siekta plačiau padengti ne vakarietiškas visuomenes ir analizuoti demokratinės politinės kultūros vystymąsi naujosiose demokratijose. EVS grupė nedalyvavo šioje 1995 m. bangoje.

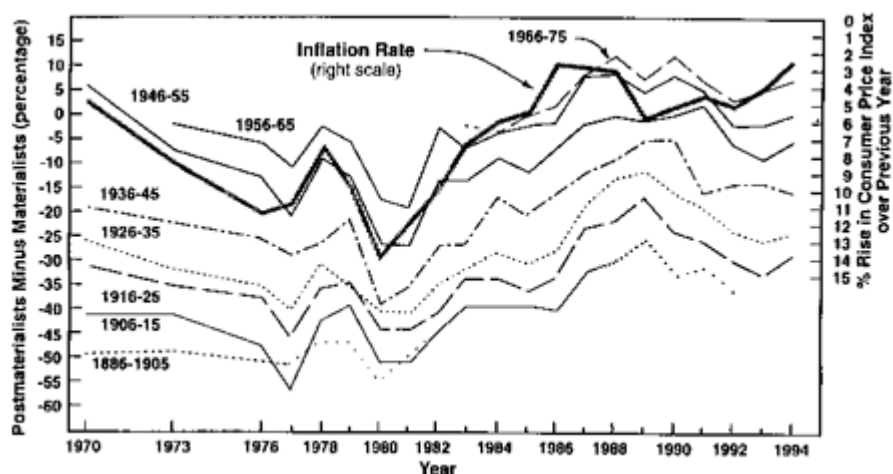
1999-2001 m. ketvirtąjį tyrimų (trečiąją EVS bei trečiąją WVS) bangą vykdė WVS ir EVS grupės kartu. Trečioji EVS banga apėmė 32 Vakarų, Rytų ir Centrinės Europos šalis (taip pat ir Lietuvą). WVS trečioji 2000 m. banga, apėmė apie 60 šalių, prioritetą skiriant islamiškų ir Afrikos visuomenių tyrimams. Šis tyrimų bangą buvo istoriškai didžiausias kada nors atliktas visuomenių įsitikinimų sistemos tyrimas, surenkant reprezentatyvius duomenis iš 81 visuomenės, kurios atstovauja 85 proc. pasaulio gyventojų. Visuomenės ir šalys pilnai apėmė visą skirtumų intervalą pasaulyje pagal pajamų gyventojui (nuo 300\$ iki 35 000\$ per metus), politinio ir ekonominio režimo (nuo autoritarinio režimo iki senųjų demokratijų; nuo rinkos ekonomikų iki kų tik išsivadavusių iš planinės ekonomikos šalių), kultūrinės (nuo pabrėžiančių socialinį konformizmą ir grupės viršenybę iki propaguojančių saviraišką ir žmogaus laisvę) ir religinės (nuo krikščionybės iki islamo, konfucianizmo, induizmo) tradicijos kriterijus.

2005 m. vykdyta penktoji tyrimų (ketvirtoji WVS banga).

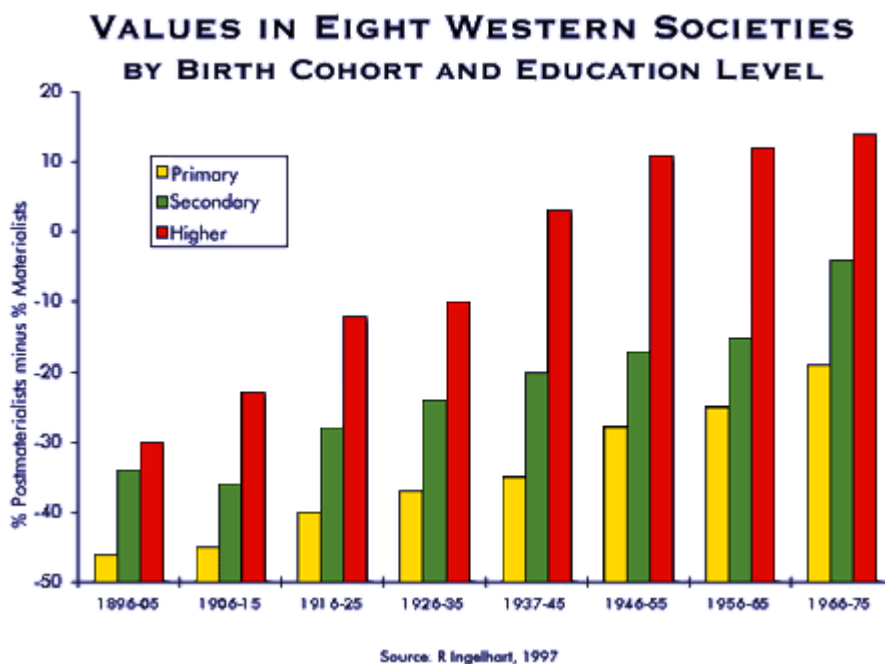
2.3. Pagrindiniai EVS ir WVS rezultatai

EVS ir WVS bangų rezultatai sustiprino nuomonę, kad Europoje ir pasaulyje palaipsniui vyksta gili modernios kultūros transformacija, tačiau skirtingu greičiu skirtingose šalyse, priklausimai nuo jos socioekonominio išsivystymo lygio ir istorinio palikimo, ir vyraujanti šios transformacijos kryptis yra gana aiškiai nuspėjama (Ronald Inglehart and Inter-university Consortium for Political and Social Research 2004).

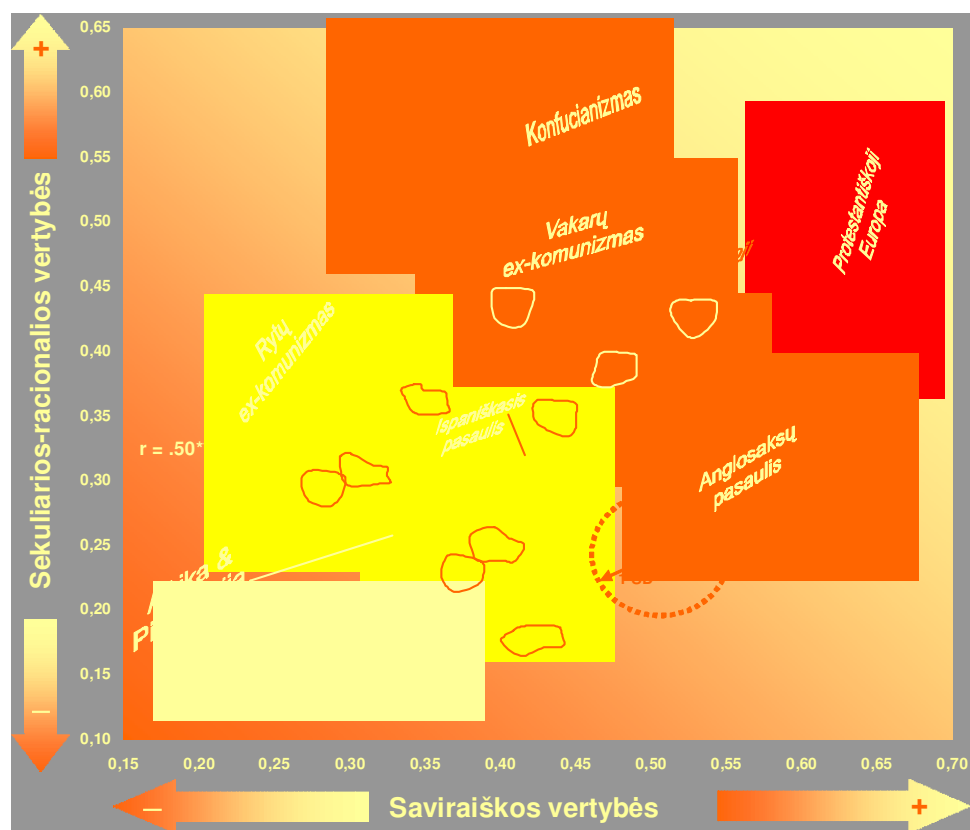
Remiantis EVS ir WVS duomenimis, buvo parodyta, kad vakarų industrinėse šalyse ir visame pasaulyje vyksta dideli vertybių pokyčiai, nes jaunesnės kartos žmonės, pasižymintys didesnės saviraiškos ir sekuliariomis vertybėmis, natūraliai keičia senesnės kartos žmones, šiems natūraliai išmirštant.



2.1 pav. Vertybių kaita kohortose 1970-1994 m šešiose vakarų Europos šalyse (postmaterialistų procentas, atėmus materialistų dalį). Šaltinis: R. Inglehart. Modernization and Postmodernization. Princeton University Press, 1997.



2.2 pav. Postmaterialios vertybės gimimo kohortos ir išsilavinimo pjūviu. Šaltinis: R. Inglehart. Modernization and Postmodernization. Princeton University Press, 1997.

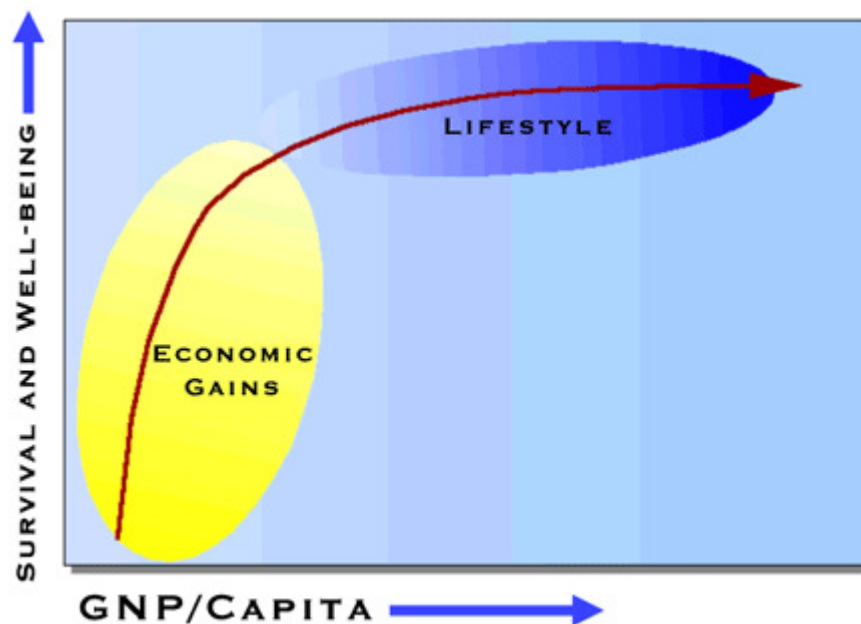


2.3 pav. Šalių grupės tarp dviejų vertybių dimensijų. Šaltinis: R. Inglehart. *Modernization and Postmodernization*. Princeton University Press, 1997.

Žemėlapis atspindi tai, kad pagrindinės vertybės yra glaudžiai susijusios (jų tarpusavio statistinė koreliacija yra stipri) ir didesniame abstrakcijos lygyje sudaro dvi atskiras vertybių dimensijas, kurios gerai apibūdina skirtingas kultūras (šalių grupes).

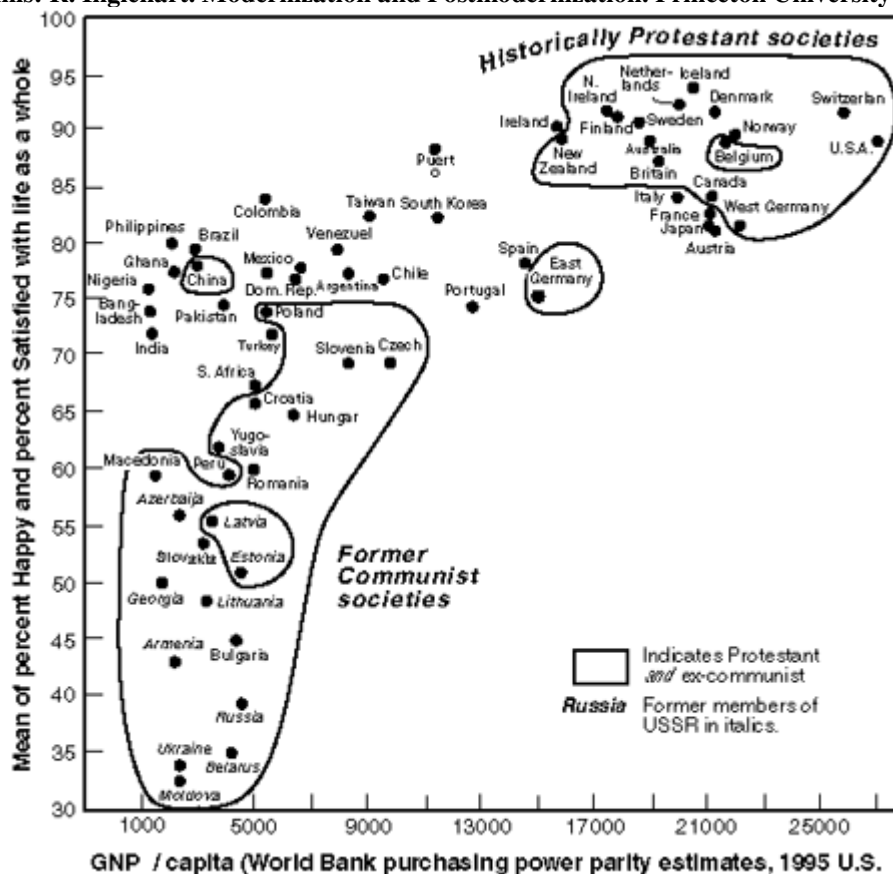
(1) tradicinės/ sekuliaris - racionalios ir (2) išgyvenimo/saviraiškos vertybės. Šios dvi dimensijos paaiškina virš 70 % dešimties indikatorių, naudotų faktorinėje visų WVS šalių vertybių analizėje, dispersijos, be to, šios dimensijos yra labai stipriai susijusios ir su kitomis svarbiomis orientacijomis.

Tradicinių/sekliarių-racionalių vertybių dimensija atskiria visuomenes, kuriose religija yra labai svarbi ir kuriose ne. Tradicinės visuomenės taip pat pabrėžia tėvų-vaikų ryšį, paklusimą autoritetui, pasižymi tvirtais absoliučiais standartais (vertybėmis), tradicinėmis šeimos vertybėmis, atmesdami skyrybas, abortus, eutanaziją, savižudybę; tradicinės visuomenės atskiria aukštas pasididžiavimo savo tauta lygis, nacionalistinės nuostatos. Sekuliaris-racionalias visuomenes apibūdina priešingo poliškumo vertybės.



Source: R. Inglehart, 1997

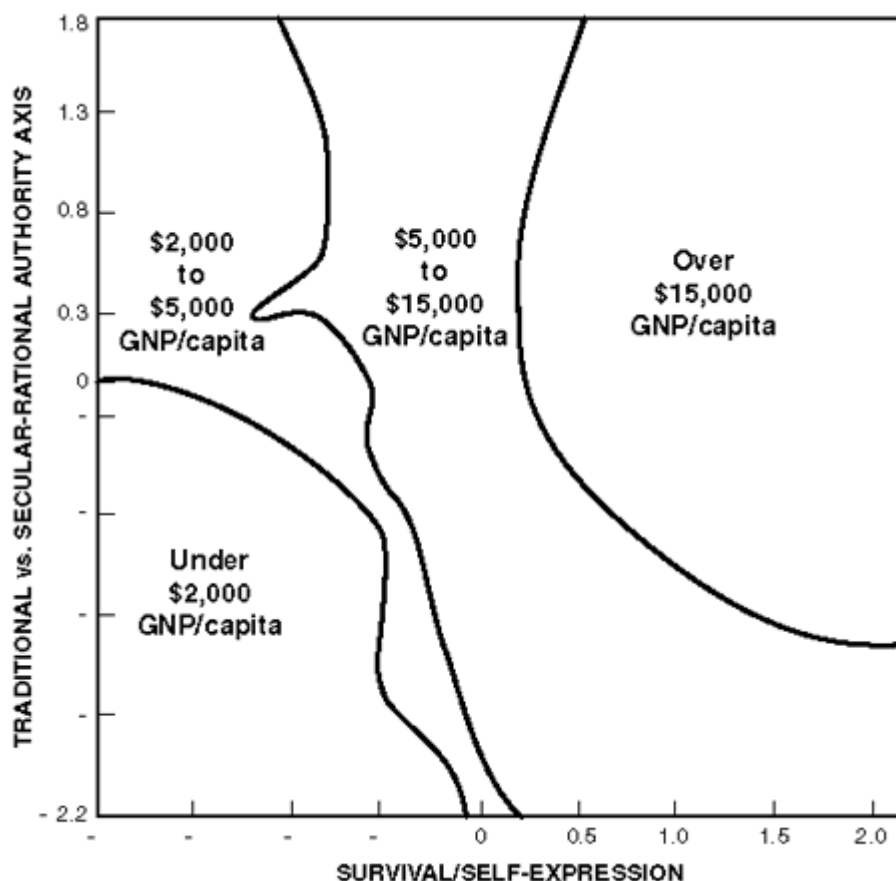
2.4 pav. Išgyvenimo ir gerovės vertybių priklausomybė nuo šalių bendro vidaus produkto, tenkančio vienam gyventojui. Šaltinis: R. Inglehart. Modernization and Postmodernization. Princeton University Press, 1997.



2.5 pav. Subjektyvios gerovės priklausomybė nuo ekonominio išsivystymo (šalių bendro vidaus produkto, tenkančio vienam gyventojui) ir istoriškai panašios patirties šalių grupės. $R=0.70$ $N=65$ $p<0.0000$. Šaltinis: R. Inglehart, H-D. Klingemann. Genes, Culture and Happiness. MIT Press, 2000.

Antroji tarpkultūrinių vertybių skirtumų dimensija yra susijusi su perėjimu iš industrinės visuomenės, kuriai būdingos išgyvenimo (saugumo ir pan.) vertybės, į post-industrinę, kuriai būdingos saviraiškos vertybės. Tai aiškinama tuo, kad labai stipriai padidėjus gerovei tam tikrose šalyse, vis didesnei gyventojui daliai išgyvenimo problema niekada nekyla, todėl įvyko perėjimas

nuo fizinio ir ekonominio saugumo poreikių patenkinimo ir su jais susijusių vertybių puoselėjimo į subjektyvios gerovės, saviraiškos, gyvenimo kokybės vertybių išplitimą.

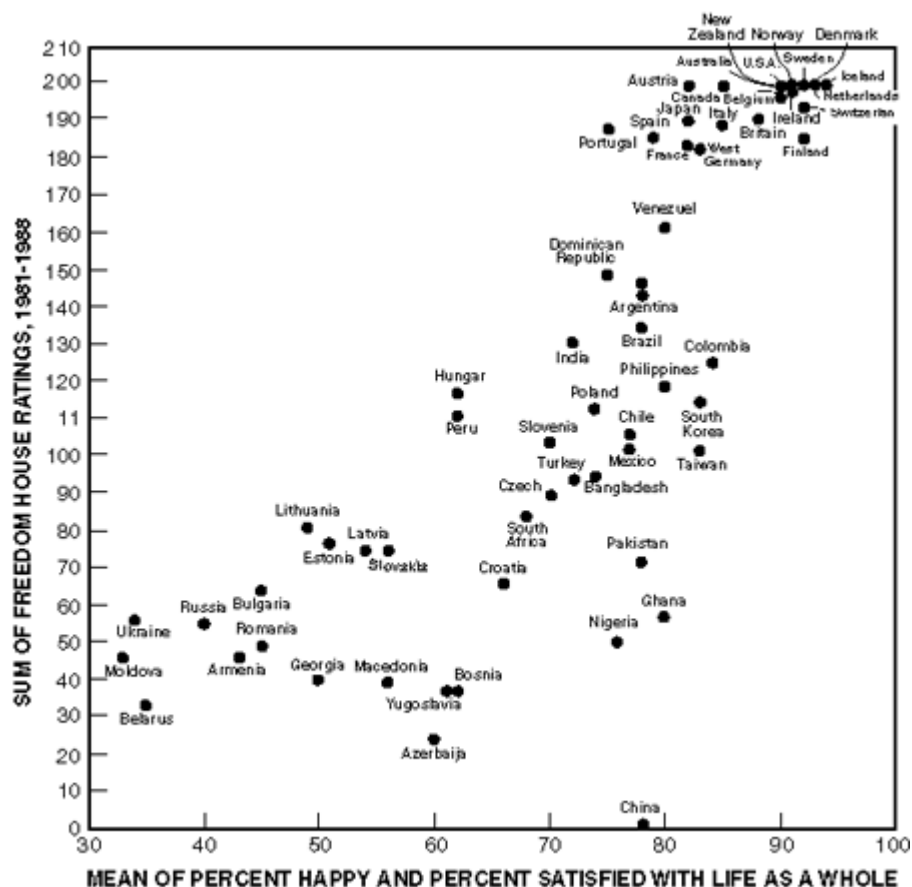


2.6 pav. 65 visuomenių ekonominio išsivystymo rodikliais (1995 m. BVP/vienam gyventojui, Pasaulio banko "World Development Report", 1997, pp. 214-215) pagrįsta ekonominio išsivystymo lygio ryšys su dviejų vertybinių dimensijų lygiais. Šaltinis: Inglehart, Baker. American Sociological Review, 2000.

Inglehart ir Baker (2000) parodo, kad industrinėse šalyse vertybės pakito nuo tradicinių link sekuliarių. Tačiau bendresnis - modernizacijos – procesas nėra tiesinis ir vyksta toliau: po industrializacijos visuomenės tampa žinių visuomenėmis: šiame etape išgyvenimo vertybes keičia saviraiškos vertybės.

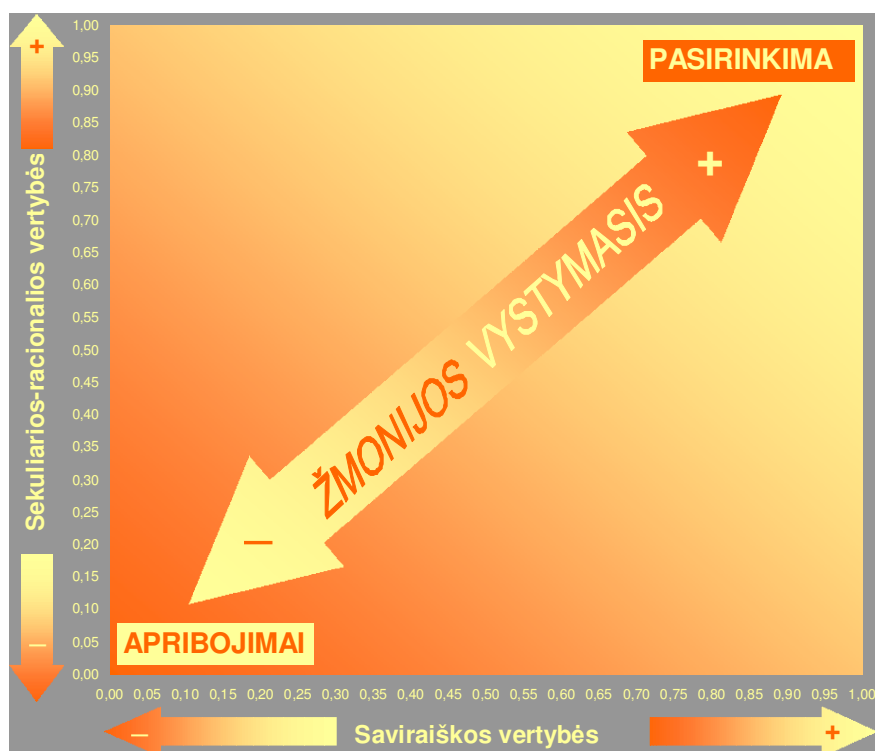
Visuomenės arba individų lygiu tradicinių/sekuliarių vertybių dimensija atskiria Materialistus ir Postmaterialistus, kuriuos skiria priklausomybė skirtingoms kartoms, vienos kurių turėjo kovoti už išgyvenimo, o kitoms išgyvenimas (saugumas) buvo savaime suprantama duotybė. Saviraiškos vertybes apibūdina didelis dėmesys aplinkos apsaugai, aukšta tolerancija kitokiems požiūriams, siekis aktyviai dalyvauti ekonominių ir politinių klausimų sprendime, tolerancija kitoms grupėms (užsieniečiams, kitos seksualinės orientacijos žmonėms), lyčių lygybei; vaikų auklėjimo praktikoje mažiau akcentuojamas sunkus darbas, o daugiau vaizduotės ir tolerancijos vystymas. Galų gale, kad saviraiškos vertybėmis pasižyminčias visuomenes išskiria aukštas tarpasmeninio pasitikėjimo lygis.

Manoma, kad šie pokyčiai, sukuria tolerancijos ir pasitikėjimo atmosferą, kurioje propaguojamos individo laisvės ir saviraiškos vertybės, ugdomas aktyvavus politiškai asmuo ir kuri yra būtina demokratijos prielaida.



2.7 pav. Subjektyvios gerovės (vidutinė šalies faktoriaus vertė, gauta faktorizavus laimės, pasitenkinimo gyvenimu ir subjektyvios gerovės kintamuosius) priklausomybė nuo demokratiškumo stiprumo (Freedom House pilietinių laisvių ir politinių teisių indeksų suma 1981-1998 m., atėmus iš 236) ($r=.78$ $N=62$ $p=.0000$). Šaltinis: R.Inglehart, H-D. Klingemann. Genes, Culture and Happiness. MIT Press, 2000.

Bene abstrakčiausia yra pasiūlyta Žmonijos vystimosi teorija, numatanti žmonijos vystimąsi platesnio geografiškai ir gilesnio raiškos prasme sekuliarių-racionalių ir saviraiškos vertybių išplitimo link.



2.8 pav. Žmonijos vystimosi apibendrinta schema. Šaltinis: Welzel, Christian, Ronald Inglehart, Hans-Dieter Klingemann. „The Theory of Human Development: A Cross-Cultural Analysis“ European Journal Of Political Research 42,3 (2003), p. 341-379.

2.4. Teorinės EVS ir WVS analizės perspektyvos

EVS nesiekta patvirtinti ar atmesti specifinių hipotezių, teorinių idėjų. Tačiau metodologiniu požiūriu, EVS projekto tyrėjai laikosi hierarchinio nuostatų ir vertybių struktūros modelio, t.y., kad nuomonės ir elgesio intencijos atskleidžia bendresnę nuostatą, kuri paaiškina arba motyvuoja elgesį tam tikrų objektų ir situacijų atžvilgiu, taip pat, kad egzistuoja latentinė (t.y., tiesiogiai nematuojama klausimyno pagalba) motyvacinė struktūra, kuriai suteikė pavadinimą vertybė ir kuri yra kur kas bendresnis elgesį motyvuojantis principas negu nuostata, galintis paaiškinti konkrečias nuostatas. Šis teorinis-metodologinis pagrindas yra paimtas iš socialinės psichologijos tradicijos (pvz., Ajzen ir Fishbein 1980). EVS tyrėjai, naudodami faktorinę, latentinių klasių/bruožų (angl. latent class analysis, latent trait analysis) analizę, ieško, validuoja skirtingais laiko periodais ir skirtingose šalyse atliktų tyrimų metu fiksuojamas vertybes.

WVS yra glaudžiai susijęs su R. Ingleharto darbais ir jų taša, todėl ir tyrimo eiga bei klausimyno papildymai remiasi postmaterialistų/materialistų, saviraiškos/saugumo vertybių ir tradicinės/sekuliarios visuomenės skirtimis, elgesio ir nuostatų skirtumų tarp kartų ir įvairių kultūrų įvairiose gyvenimo sferose paieška.

2.5. 1990, 1996, 1999 m. Europos ir pasaulio vertybių tyrimų Lietuvoje kintamųjų sąvadas bei paaiškinimai

Atrinktų kintamųjų kiekis: 8

Nr.	Kintamųjų vardai	Kintamųjų žymės	Kintamųjų reikšmių žymės (bendri kodai: „-1“ - nežino, „-2“ - neatsakė; „-3“ - netaikytina, „-4“ - neklausta, „-5“ - neatsakyta)	Kintamųjų prasmė ir turinys	Duomenų analizei reikalingi paaiškinimai
1.	e033	Saviidentifikavimas politinės kairės-dešinės skalėje	1- kairė, 2, 3, 4, 5, 6, 7, 8, 9, 10 – dešinė	Klasikinė skalė, leidžianti kompaktiškai (vienos skalės pagalba) įvertinti asmens politinę poziciją.	Tik vienoje bangoje! Tvarkos skalės kintamasis
2.	f028	Ar dažnai paskutiniu metu einate į bažnyčią ar religinius susirinkimus?	1 – dažniau negu kartą per savaitę 2 – kartą per savaitę 3 – kartą per mėnesį 4 – tik per Kalėdas ir Velykas 5 – per religines šventes 6 – kartą per metus 7 – rečiau nei kartą per metus 8 – niekada, beveik niekada	Vienas iš religingumo indikatorių	Tvarkos skalės kintamasis
3.	s018	Svoris (imties dydis 1000)	Skaičiai nuo 0,148 iki 9,518	Kiekvienų metų imties dydis suvienodinamas į 1000 respondentų skaičių, tam, kad būtų lengviau lyginti tarp skirtingų metų, o taip pat pasveriant pagal demografines respondentų charakteristikas. Kiekvieno respondento atsakymams priskirtas koeficientas, iš kurio dauginama atsakymo reikšmė. Koeficientas parinktas tokiu būdu, kad pakoreguotų imties charakteristikų nuokrypį nuo populiacijos charakteristikų (pvz., jei apklausta didesnis procentas asmenų su aukštesniu išsilavinimu negu yra tarp visų šalies gyventojų virš 18 m., tada svorio koeficientai šiai grupei bus mažesnis negu respondentams be	Intervalinis kintamasis. Naudojamas tik duomenų svėrimui ir daugiau jokioms kitoms analizėms atlikti

Nr.	Kintamųjų vardai	Kintamųjų žymės	Kintamųjų reikšmių žymės (bendri kodai: „-1“ - nežino, „-2“ - neatsakė; „-3“ – netaikytina, „-4“ – neklausta, „-5“ – neatsakyta)	Kintamųjų prasmė ir turinys	Duomenų analizei reikalingi paaiškinimai
				aukštojo išsilavinimo, tokiu būdu sumažinant paklaidą, skaičiuojant statistines charakteristikas).	
4.	s020	Tyrimo metai	1990, 1997, 1999	Tyrimo atlikimo metai	Intervalinis kintamasis
5.	x007	Šeimyninė padėtis	1 - Vedęs/-usi, 2 - Gyvena kartu, nevedę, 3 - Išsiskyręs/-usi, 4 - Gyvena atskirai, 5 - Našlys/-ė, 6 - Nevedęs/-usi, 7 - Išsiskyręs/-usi, gyvena atskirai arba našlys/-e, 8 - Gyvena ne kartu, bet turi pastovius santykius (vedę, kohabitacija)	Šeimyninė padėtis	Vardų skalės kintamasis
6.	x028	Darbinis statusas	1 - Pilna darbo diena, 2 - Ne pilna darbo diena, 3 - Savarankiškai dirbantis/-i, 4 - Pensijoje, 5 - Namų šeimininkė, 6 - Studentas/-ė, 7 - Bedarbis/-ė, 8 - Kita	Darbinis statusas	Vardų skalės kintamasis
7.	tradrat5	Sekuliarios vs. tradicinės vertybės	Kintamųjų reikšmės intervale (1,5; 3,3)	Indeksas, matuojantis tradicinių vertybių dominavimą lyginant su sekuliariomis, racionaliomis vertybėmis. Tradicinės vertybės apima šiuos kintamuosius: • Dievas respondento gyvenime yra labai svarbu; • Vaikui yra svarbiau išmokti paklusnumo ir religinio tikėjimo negu nepriklausomumo ir apsisprendimo („Autonomijos indeksas“); • Abortai niekada nepateisinami; • Respondentas turi stiprų nacionalinio	Intervalinis kintamasis

Nr.	Kintamųjų vardai	Kintamųjų žymės	Kintamųjų reikšmių žymės (bendri kodai: „-1“ - nežino, „-2“ - neatsakė; „-3“ – netaikytina, „-4“ – neklausta, „-5“ – neatsakyta)	Kintamųjų prasmė ir turinys	Duomenų analizei reikalingi paaiškinimai
				pasididžiavimo jausmą; Respondentas daugiau linkęs gerbti valdžią. Sekuliarios vertybės apibūdinamos priešingai tradicinių vertybių turiniui. Žemensnės indekso vertės atitinka daugiau tradicinius principus, o aukštesni – sekuliarus/racionalius principus. Indeksas sudarytas suvidurkinus paminėtus kintamuosius. Kintamasis yra apskaičiuotas faktorinės analizės su Varimax pasukimu būdu (Inglehart ir Baker 2000).	
8.	survself	Išlikimo vs. saviraiškos vertybės	Kintamųjų reikšmės intervale (-3,3; 1,7)	Išlikimo/saviraiškos orientacijos indeksas matuoja išlikimo vs. saviraiškos vertybes (Inglehart ir Baker 2000). Išlikimo vertybės apima šiuos kintamuosius: - Respondentas teikia prioritetą ekonominiam ir fiziniam saugumui, o ne saviraiškai ir gyvenimo kokybei (keturių indikatorių materialistinių/postmaterialistinių vertybių indeksas). - Respondentas save apibūdina kaip nelabai laimingą. - Respondentas nėra pasirašęs ir nepasirašinėtų peticijos. - Homoseksualumas niekada nepateisinamas. - Reikia būti atsargiam pasitikint žmonėmis. Saviraiškos vertybių lygis yra reikšmės, atitinkančios išvardintų kintamųjų priešingą polių. Kintamasis yra apskaičiuotas faktorinės analizės su Varimax pasukimu būdu (Inglehart ir Baker 2000).	Intervalinis kintamasis

3. 2002, 2004, 2006 M. EUROPOS SOCIALINIO TYRIMO APRAŠYMAS IR KINTAMŲJŲ SĄVADAS BEI PAAIŠKINIMAI

Šiame skyriuje pateikiamas Europos socialinio tyrimo (toliau – ESS) 2002, 2004, 2006 m. aprašymas: tyrimo tikslai, objektas, vykdymo istorija, pagrindiniai rezultatai ir teorinės perspektyvos. Be to, pateikiamas šių tyrimų kintamųjų sąvadas bei paaiškinimai: kintamųjų kiekis, kintamųjų vardai, kintamųjų žymės, kintamųjų reikšmių žymės, kintamųjų prasmė ir turinys, duomenų analizei reikalingi paaiškinimai.

3.1. ESS tikslai ir objektas

Europos socialinio tyrimo tikslas – aukščiausiu moksliniu standartu atlikti testinį ir pavyzdinį žmoniškųjų vertybių kaitos Europoje tyrimą.

Šiuo tyrimu visų pirma siekiama stebėti ir interpretuoti besikeičiančias visuomenės nuostatas ir vertybes Europoje bei jų sąveiką su besikeičiančiomis Europos institucijomis, antrą – vystyti ir panaudoti patobulintus metodus tarptautinių apklausų vykdymui Europoje ir už jos ribų, trečia – sukurti ir sukaupti Europos socialinių indikatorių, įtraukiant ir nuostatų indikatorius, bazę.

ESS objektas yra Europos šalių gyventojų socialiniai (apimantys politinę, ekonominę, religinę, kultūrinę sferas) įsitikinimai, nuostatos, vertybės ir atitinkamos praktikos (elgesys) socialinėje realybėje. Klausimynas sudarytas iš pagrindinio klausimų modulio, kurio turinys nesikeičia ir iš kelių papildomų klausimų modulių, kurie į tyrimą įtraukiami ne kiekvienoje tyrimų bangoje. Pagrindinio klausimų modulio klausimai leidžia stebėti tokius socialiniu kintamuosius kaip žiniasklaidos vartojimo ypatybės, socialinis ir politinis pasitikėjimas, politinis aktyvumas, dalyvavimas, sociopolitinės orientacijos, valstybės valdymas ir efektyvumas, moralinės, politinės ir socialinės vertybės, socialinė atskirtis, nacionalinis ir religinis tapatumas, gerovė, sveikata ir saugumas, demografiniai ir socio-ekonominiai kintamieji.

Apklausos pabaigoje respondentui pateikiamas papildomas klausimynas, kuriuo detaliau įvertinamos žmoniškiosios vertybės ir patikrinamas pagrindinio klausimyno klausimų patikimumas ir tikslumas.

3.2. ESS vykdymo istorija

Europos socialinis tyrimas (angl. European Social Survey – ESS) vykdomas virš 30 – ties šalių kas antrais metais. Pirmą tyrimų bangą buvo vykdyta 2002/2003 m. 22-ose šalyse, antroji - 2004/2005 m. 26-iose šalyse ir trečioji - 2006/2007 m. 25-iose šalyse. Lietuva planuoja prisijungti prie tyrimo ketvirtosios tyrimų bangos metu 2008/2009 m.

Šis tęstinis projektas yra finansuojamas Europos komisijos, Europos mokslo fondo ir kiekvienos šalies nacionalinių finansuotojų. Atliekamos apklausos pasižymi ypatingai aukštais kokybės standartais.

Tyrime naudojamas griežtai atsitiktinis tikimybinis respondentų imties parinkimo metodas, reikalaujamas minimalus respondentų atsako dažnumas - 70% ir griežta klausimynų vertimo bei atgalinio vertimo metodika. Asmeninis interviu trunka apie valandą.

Daugiau informacijos pateikiama adresu <http://www.europeansocialsurvey.org>.

3.3. Pagrindiniai ESS rezultatai ir publikacijos

Gausus ESS tyrimais grįstų publikacijų sąrašas (virš 200 įrašų) pateikiamas tyrimo puslapyje <http://www.europeansocialsurvey.org>.

3.4. Teorinės ESS analizės perspektyvos

Tyrime remiamasi Schwartz žmogiškųjų vertybių matavimo testu, kitais tipiniais žmoniškąsias vertybes, nuostatas, įsitikinimus, nuomones, veiksmo intencijas matuojančiais instrumentais.

3.5. 2002, 2004, 2006 m. ESS kintamųjų sąvadas bei paaiškinimai

ESS tyrimo duomenys yra suskaidyti pagal atlikimo bangą, t. y. kiekvienai tyrimo bangai pateikta atskira laikmena. Kiekvienoje iš jų yra:

Bendra informacija apie tyrimą, kurią atspindi pirmi 4 kintamieji: „name“, „essround“, „edition“ ir „proddate“.

Respondento identifikacinį numerį atspindintis kintamasis „idno“.

Duomenų svėrimui skirti 2 paskutiniai kintamieji „pweight“ ir „dweight“.

Nr.	Kintamųjų vardai	Kintamųjų žymės	Kintamųjų reikšmių žymės	Kintamųjų prasmė ir turinys	Duomenų analizei reikalingi paaiškinimai
1.	name	ESS duomenų failo pavadinimas	ESS2002/2003, ESS2e03, ESS3e03_1	ESS duomenų failo pavadinimas	Vardų skalės kintamasis
2.	essround	ESS vykdymo banga	1 – pirmoji banga, 2 – antroji banga, 3 – trečioji banga	Tyrimų bangos numeris (1, 2, 3)	Tvarkos skalės kintamasis
3.	edition	ESS duomenų failo versijos numeris	6.0, 3.0, 3.1 (versijų numeriai)	ESS duomenų failo versijos numeris	Vardų skalės kintamasis
4.	proddate	ESS duomenų failo sudarymo data	Datos	ESS duomenų failo sudarymo data	Vardų skalės kintamasis
5.	idno	Respondento identifikacinis numeris	Sveiki skaičiai	Respondento identifikacinis numeris	Vardų skalės kintamasis
6.	pweight	Populiacijos dydžio svoris (turi būti naudojamas kartu su dweight)	Skaičiai	Šalies gyventojų skaičiaus svoris, naudojamas kartu su „dweight“	Intervalų skalės kintamasis
7.	dweight	Tyrimo dizaino svoris	Skaičiai	Tyrimo dizaino svoris	Intervalų skalės kintamasis

Toliau kiekvienoje laikmenoje yra originalūs atrinkti kintamieji ir tie patys kintamieji transformuoti taip, kad tiktų pavyzdinei statistinei analizei atlikti (pekoduoti, sulietuvinti ir pan.). Iš viso per visas 3 ESS tyrimo vykdymo bangas sukurta (transformuojant originalius kintamuosius) 15 kintamųjų.

Nr.	Kintamųjų vardai	Kintamųjų žymės	Kintamųjų reikšmių žymės	Kintamųjų prasmė ir turinys	Duomenų analizei reikalingi paaiškinimai
2002 m. banga					
1.	salis	Šalis	Tinkamos reikšmės: AT Austrija BE Belgija CH Šveicarija CZ Čekija DK Danija ES Ispanija FI Suomija FR Prancūzija GB Jungtinė karalystė GR Graikija HU Vengrija IE Airija IL Izraelis IT Italija Lenkija LU Liuksemburgas NL Olandija NO Norvegija PT Portugalija SI Slovėnija Švedija Vokietija	Apklaustos vykdymo šalis	Vardų skalės kintamasis; perkoduotas iš originalaus kintamojo „cntry“.
2.	salis_sk	Šalis	Tinkamos reikšmės: 1 – Lenkija, 2 – Vokietija, 3 – Švedija; Trūkstamos reikšmės: 99 – kitos	Apklaustos vykdymo šalis	Vardų skalės kintamasis; perkoduotas iš originalaus kintamojo „cntry“; skirtas atlikti analizei, kai negalima naudoti tekstinio formato kintamųjų (pavyzdžiui, braižant kryžmines stačiakampes diagramas).
3.	domejimasis_politika	Kiek domisi politika	Tinkamos reikšmės:	Domėjimosi politika	Tvarkos skalės kintamasis; perkoduotas iš

Nr.	Kintamųjų vardai	Kintamųjų žymės	Kintamųjų reikšmių žymės	Kintamųjų prasmė ir turinys	Duomenų analizei reikalingi paaiškinimai
			1 – labai domisi, 2 – domisi pakankamai, 3 – mažai domisi, 4 – visiškai nesidomi; Trūkstamos reikšmės: 7 – atsisakė nurodyti, 8 – nežino, 9 – neatsakė.	lygis	originalaus kintamojo „polintr“.
4.	religingumas	Savęs priskyrimas kokiai nors religijai ar denominacijai	Tinkamos reikšmės: 1 – priskiria, 2 – nepriskiria; Trūkstamos reikšmės: 7 – atsisakė nurodyti, 8 – nežino, 9 – neatsakė	Savęs priskyrimas kokiai nors religijai ar denominacijai	Dvireikšmis kintamasis; perkoduotas iš originalaus kintamojo „rlgblg“.
5.	religija	Religija ar denominacija, kuriai priskiria save šiuo metu	Tinkamos reikšmės: 1 – Katalikų, 2 – Protestantų, 3 – Rytų ortodoksų, 4 – Kitos krikščionių denominacijos, 5 – Žydų, 6 – Islamo, 7 – Rytų religijų, 8 – Kitų nekrikščionių religijų, Trūkstamos reikšmės: 66 – Netaikytina, 77 – Atsisakė nurodyti, 99 – Neatsakė.	Religija ar denominacija, kuriai priskiria save šiuo metu	Vardų skalės kintamasis; perkoduotas iš originalaus kintamojo „rlgdnm“.

Nr.	Kintamųjų vardai	Kintamųjų žymės	Kintamųjų reikšmių žymės	Kintamųjų prasmė ir turinys	Duomenų analizei reikalingi paaiškinimai
6.	religingumo_laipsnis	Kiek esate religingas?	Tinkamos reikšmės: 0 – visiškai nereliginas,- a, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10 – labai religingas,-a; Trūkstamos reikšmės: 77 – atsisakė, 88 – nežino, 99 – neatsakė	Religingumo laipsnis	Tvarkos skalės kintamasis; perkoduotas iš originalaus kintamojo „rlgdgr“; dėl didelio reikšmių skaičiaus tinka intervalų skalei pritaikytoms statistinėms technikoms
7.	metai	Gimimo metai	Tinkamos reikšmės: Keturženkliai skaičiai (pvz., 1975); Trūkstamos reikšmės: 7777 – Atsisakė 8888 – Nežino 9999 – Nėra atsakymo	Respondento gimimo metai	Intervalų skalės kintamasis; perkoduotas iš originalaus kintamojo „yrbrn“.
8.	amzius	Amžius	Tinkamos reikšmės: Dviženkliai skaičiai (pvz., 29); Trūkstamos reikšmės: 7777 – Atsisakė 8888 – Nežino 9999 – Nėra atsakymo	Respondento amžius	Intervalų skalės kintamasis; perskaičiuotas iš „metai“ kintamojo (2002 – metai = amzius)
2004 m. banga					
9.	salis	Šalis	Tinkamos reikšmės: 1 – Austrija 2 – Belgija 3 – Šveicarija 4 – Čekija 5 – Vokietija 6 – Danija	Apklauso vykdomo šalis	Vardų skalės kintamasis; perkoduotas iš originalaus kintamojo „cntry“.

Nr.	Kintamųjų vardai	Kintamųjų žymės	Kintamųjų reikšmių žymės	Kintamųjų prasmė ir turinys	Duomenų analizei reikalingi paaiškinimai
			7 – Estija 8 – Ispanija 9 – Suomija 10 – Prancūzija 11 – JK 12 – Graikija 13 – Vengrija 14 – Airija 15 – Islandija 16 – Liuksemburgas 17 – Olandija 18 – Norvegija 19 – Lenkija 20 – Portugalija 21 – Švedija 22 – Slovenija 23 – Slovakija 24 – Turkija 25 – Ukraina		
10.	partijos_nar	Partijos narys,-ė	Tinkamos reikšmės: 0 – ne, 1 – taip; Trūkstamos reikšmės: 7 – atsisakė nurodyti, 8 – nežinojo, 9 – neatsakė.	Respondentas priklauso kuriai nors partijai, ar nepriklauso	Dvireikšmis kintamasis; perkoduotas iš originalaus kintamojo „mmbprty“.

Nr.	Kintamųjų vardai	Kintamųjų žymės	Kintamųjų reikšmių žymės	Kintamųjų prasmė ir turinys	Duomenų analizei reikalingi paaiškinimai
11.	pasit_ek	Pasitenkinimas tuo, kaip veikia šalies ekonomika	Tinkamos reikšmės: 0 – visiškai nepatenkintas,-a, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10 – visiškai patenkintas,-a; Trūkstamos reikšmės: 77 – atsisakė nurodyti, 88 – nežinojo, 99 – neatsakė.	Pasitenkinimas tuo, kaip veikia šalies ekonomika	Tvarkos skalės kintamasis; perkoduotas iš originalaus kintamojo „stfeco“; dėl didelio reikšmių skaičiaus tinka intervalų skalei pritaikytoms statistinėms technikoms
2006 m. banga					
12.	salis	Šalis	Tinkamos reikšmės: 1 – Austrija 2 – Belgija 3 – Bulgarija 4 – Šveicarija 5 – Kipras 6 – Vokietija 7 – Danija 8 – Estija 9 – Ispanija 10 – Suomija 11 – Prancūzija 12 – Jungtinė karalystė 13 – Vengrija 14 – Airija 15 – Olandija 16 – Norvegija 17 – Lenkija 18 – Portugalija 19 – Rusija	Apklauso vykdomo šalis	Vardų skalės kintamasis; perkoduotas iš originalaus kintamojo „cntry“.

Nr.	Kintamųjų vardai	Kintamųjų žymės	Kintamųjų reikšmių žymės	Kintamųjų prasmė ir turinys	Duomenų analizei reikalingi paaiškinimai
			20 – Švedija 21 – Slovėnija 22 – Slovakija 23 – Ukraina		
13.	prod_boikotas	Boikotavo tam tikrų produktų pirkimą per pastaruosius 12 mėn.	Tinkamos reikšmės: 1 – taip, 0 – ne; Trūkstamos reikšmės: 7 – atsisakė nurodyti, 8 – nežinojo, 9 – neatsakė.	Ar respondentas boikotavo tam tikrų produktų pirkimą per pastaruosius 12 mėn.	Dvireikšmis kintamasis; perkoduotas iš originalaus kintamojo „bctprd“ (reikšmė 2 = 0).
14.	pasit_gyv_lg	Pasitenkinimas savo gyvenimo lygiu	Tinkamos reikšmės: 0 – visiškai nepatenkintas,-a, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10 – visiškai patenkintas,-a; Trūkstamos reikšmės: 77 – atsisakė nurodyti, 88 – nežinojo, 99 – neatsakė.	Pasitenkinimas savo gyvenimo lygiu	Tvarkos skalės kintamasis; perkoduotas iš originalaus kintamojo „stfsdlv“.

Nr.	Kintamųjų vardai	Kintamųjų žymės	Kintamųjų reikšmių žymės	Kintamųjų prasmė ir turinys	Duomenų analizei reikalingi paaiškinimai
15.	pajamu_lygin	Su kuo labiausiai lygina savo pajamas	Tinkamos reikšmės: 1 – Kolegomis darbe, 2 – Šeimos nariais, draugais ir kitais žmonėmis, 3 – Nelygina su niekuo Trūkstamos reikšmės: 6 – Netaikytina, 7 – Atsisakė nurodyti, 8 – Nežinojo, 9 – Neatsakyta.	Su kuo labiausiai lygina savo pajamas	Vardų skalės kintamasis; perkoduotas iš originalaus kintamojo „inccmpw“.

4. 2006 M. GEOKULTŪRINIŲ NUOSTATŲ TYRIMO APRAŠYMAS APRAŠYMAS IR KINTAMŲJŲ SĄVADAS BEI PAAIŠKINIMAI

Šiame skyriuje pateikiamas Geokultūrinių nuostatų tyrimo (toliau – GNT) 2006 m. aprašymas: tyrimo tikslai, objektas, vykdymo istorija, pagrindiniai rezultatai ir teorinės perspektyvos. Be to, pateikiamas šių tyrimų kintamųjų sąvadas bei paaiškinimai: kintamųjų kiekis, kintamųjų vardai, kintamųjų žymės, kintamųjų reikšmių žymės, kintamųjų prasmė ir turinys, duomenų analizei reikalingi paaiškinimai.

4.1. GNT tikslai ir objektas

Šio tyrimo pagrindinis tikslas buvo išsiaiškinti Lietuvos visuomenės geokultūrinės nuostatos bei atskleisti jas įtakančius faktorius.

GNT objektas yra Lietuvos visuomenės geokultūrinės nuostatos. Tyrimo metu buvo tiriama, kokias kalbas respondentai moka; kokiomis kalbomis žiūri TV; kokiomis kalbomis skaito spaudą; kokiomis kalbomis klauso radiją; kokiomis kalbomis naršo internetą, klauso muziką; informacijos kanalai; pasitikėjimas įvairių šalių žiniasklaida; kur norėtų gyventi; įvairūs gyvenimo aspektai 6 šalyse; Lietuvos sąjungininkai ir priešai; kada buvo geriau gyventi; ar turi draugų kitataučių.

4.2. GNT vykdymo istorija

Tyrimo trukmė: 2006 m. spalio 1 d. – 2007 m. kovo 31 d. Pagrindiniai tyrėjai: Pilietinės visuomenės institutas, duomenų rinkėjas (lauko darbams kontraktuota organizacija): Vilmos.

Vykdyta reprezentatyvi Lietuvos gyventojų apklausa, atrankos metodas – reprezentatyvi tikimybinė atranka, duomenų rinkimo būdas – respondento namuose.

Lietuvos geopolitinės minties tradicijoje bei užsienio politikos praktikoje santykinai mažai dėmesio skiriama pačios visuomenės geokultūrinių nuostatų tyrimams. Tokia situacija gali lemti atotrūkį tarp Lietuvos visuomenės geokultūrinių preferencijų ir valstybės kuriamų bei įgyvendinamų užsienio ar vidaus politikos strategijų. Todėl, siekiant užpildyti šią spragą, Pilietinės visuomenės institutas atliko vieną iš pirmųjų Lietuvos visuomenės geokultūrinių nuostatų tyrimą, o jo išvadas pristatė visuomenei, žiniasklaidai, akademinėi bendruomenei bei politikams.

4.3. Pagrindiniai GNT rezultatai ir publikacijos

GNT tyrimo rezultatai prieinami tyrimo puslapyje <http://www.civitas.lt/lt/?pid=74&id=52> dr. Ainės Ramonaitės, Nerijaus Maliukevičiaus ir dr. Mindaugo Degučio studijoje „Tarp Rytų ir Vakarų: Lietuvos visuomenės geokultūrinės nuostatos“. Studijoje nagrinėjamos Lietuvos visuomenės geokultūrinės nuostatos ir atsparumas kitų šalių informacinėms kultūrinėms įtakoms. Aiškinamasi, ar provakarietišką Lietuvos elito orientaciją palaiko gyventojai. Ar visuomenėje paplitusi sovietinė nostalgija kartu rodo ir prorusišką nusiteikimą? Ar pasiekia savo tikslų Rusijos informacinė politika, siekianti išlaikyti Baltijos valstybes savo kultūrinės įtakos erdvėje? Studijos autoriai taip pat stengėsi išsiaiškinti, ar ir kiek Lietuva yra pažeidžiama. Kokia Lietuvos visuomenės dalis gyvena Rusijos informacinėje erdvėje? Ar Lietuvos visuomenė toleruoja kitų valstybių įtaką Lietuvos politikams? Ar ji būtų atspari informacinio karo atakoms?

Studijos „Tarp Rytų ir Vakarų: Lietuvos visuomenės geokultūrinės nuostatos“ santrauka prieinama adresu http://www.civitas.lt/admin/get_file.php?id=37.

4.4. Teorinės GNT analizės perspektyvos

Tyrime remiamasi įvairiomis skalėmis, kurios įvairiapusškai vertina įvairias Lietuvos gyventojų geokultūrinės nuostatos, Lietuvos informacinės erdvės įtakas.

4.5. 2006 m. GNT kintamųjų sąvadas bei paaiškinimai

2006 m. Lietuvos gyventojų Geokultūrinių nuostatų tyrimo duomenis sudaro du blokai: originalūs kintamieji ir sudarytų skalių kintamieji.

Pirmiausiai laikmenoje pateikti originalūs kintamieji, kurių iš viso yra 84.

Nr.	Kintamųjų vardai	Kintamųjų vardai originaliame faile	Kintamųjų žymės	Kintamųjų reikšmių žymės	Kintamųjų prasmė ir turinys	Duomenų analizei reikalingi paaiškinimai
1.	L12T_01	L12T_01	12. Politinės santvarkos vertinimas. Buvusi TSRS	1 - Labai blogai, 2, 3, 4, 5 - Labai gerai	Politinės santvarkos vertinimas. Buvusi TSRS	Tvarkos skalės kintamasis
2.	L12T_02	L12T_02	12. Švietimo sistemos vertinimas. Buvusi TSRS	1 - Labai blogai, 2, 3, 4, 5 - Labai gerai	Švietimo sistemos vertinimas. Buvusi TSRS	Tvarkos skalės kintamasis
3.	L12T_03	L12T_03	12. Sveikatos apsaugos sistemos vertinimas. Buvusi TSRS	1 - Labai blogai, 2, 3, 4, 5 - Labai gerai	Sveikatos apsaugos sistemos vertinimas. Buvusi TSRS	Tvarkos skalės kintamasis
4.	L12T_04	L12T_04	Socialinės apsaugos vertinimas. Buvusi TSRS	1 - Labai blogai, 2, 3, 4, 5 - Labai gerai	Socialinės apsaugos vertinimas. Buvusi TSRS	Tvarkos skalės kintamasis
5.	L12T_05	L12T_05	Saugumo, teisėsaugos darbo vertinimas. Buvusi TSRS	1 - Labai blogai, 2, 3, 4, 5 - Labai gerai	Saugumo, teisėsaugos darbo vertinimas. Buvusi TSRS	Tvarkos skalės kintamasis
6.	L12T_06	L12T_06	Įsidarbinimo galimybių vertinimas. Buvusi TSRS	1 - Labai blogai, 2, 3, 4, 5 - Labai gerai	Įsidarbinimo galimybių vertinimas. Buvusi TSRS	Tvarkos skalės kintamasis
7.	L12T_07	L12T_07	Ekonominės gerovės vertinimas. Buvusi TSRS	1 - Labai blogai, 2, 3, 4, 5 - Labai gerai	Ekonominės gerovės vertinimas. Buvusi TSRS	Tvarkos skalės kintamasis
8.	L12T_08	L12T_08	Kultūros, meno vertinimas. Buvusi TSRS	1 - Labai blogai, 2, 3, 4, 5 - Labai gerai	Kultūros, meno vertinimas. Buvusi TSRS	Tvarkos skalės kintamasis
9.	L12T_09	L12T_09	Žmogaus teisių vertinimas. Buvusi TSRS	1 - Labai blogai, 2, 3, 4, 5 - Labai gerai	Žmogaus teisių vertinimas. Buvusi TSRS	Tvarkos skalės kintamasis
10.	L12T_10	L12T_10	Demokratijos lygio vertinimas. Buvusi TSRS	1 - Labai blogai, 2, 3, 4, 5 - Labai gerai	Demokratijos lygio vertinimas. Buvusi TSRS	Tvarkos skalės kintamasis

Nr.	Kintamųjų vardai	Kintamųjų vardai originaliame faile	Kintamųjų žymės	Kintamųjų reikšmių žymės	Kintamųjų prasmė ir turinys	Duomenų analizei reikalingi paaiškinimai
11.	L12T_11	L12T_11	Laisvės (galimybė reikšti savo mintis, užsiimti veikla, keliauti ir t.t.) vertinimas. Buvusi TSRS	1 - Labai blogai, 2, 3, 4, 5 - Labai gerai	Laisvės (galimybė reikšti savo mintis, užsiimti veikla, keliauti ir t.t.) vertinimas. Buvusi TSRS	Tvarkos skalės kintamasis
12.	L12T_12	L12T_12	Teisingumo užtikrinimo vertinimas. Buvusi TSRS	1 - Labai blogai, 2, 3, 4, 5 - Labai gerai	Teisingumo užtikrinimo vertinimas. Buvusi TSRS	Tvarkos skalės kintamasis
13.	L12T_13	L12T_13	Lygių galimybių visiems visuomenės nariams vertinimas. Buvusi TSRS	1 - Labai blogai, 2, 3, 4, 5 - Labai gerai	Lygių galimybių visiems visuomenės nariams vertinimas. Buvusi TSRS	Tvarkos skalės kintamasis
14.	L12T_14	L12T_14	Moralės normų laikymosi vertinimas. Buvusi TSRS	1 - Labai blogai, 2, 3, 4, 5 - Labai gerai	Moralės normų laikymosi vertinimas. Buvusi TSRS	Tvarkos skalės kintamasis
15.	L12E_01	L12E_01	Politinės santvarkos vertinimas. Vakarų Europa	1 - Labai blogai, 2, 3, 4, 5 - Labai gerai	Politinės santvarkos vertinimas. Vakarų Europa	Tvarkos skalės kintamasis
16.	L12E_02	L12E_02	Švietimo sistemos vertinimas. Vakarų Europa	1 - Labai blogai, 2, 3, 4, 5 - Labai gerai	Švietimo sistemos vertinimas. Vakarų Europa	Tvarkos skalės kintamasis
17.	L12E_03	L12E_03	Sveikatos apsaugos sistemos vertinimas. Vakarų Europa	1 - Labai blogai, 2, 3, 4, 5 - Labai gerai	Sveikatos apsaugos sistemos vertinimas. Vakarų Europa	Tvarkos skalės kintamasis
18.	L12E_04	L12E_04	Socialinės apsaugos vertinimas. Vakarų Europa	1 - Labai blogai, 2, 3, 4, 5 - Labai gerai	Socialinės apsaugos vertinimas. Vakarų Europa	Tvarkos skalės kintamasis
19.	L12E_05	L12E_05	Saugumo, teisėsaugos darbo vertinimas. Vakarų Europa	1 - Labai blogai, 2, 3, 4, 5 - Labai gerai	Saugumo, teisėsaugos darbo vertinimas. Vakarų Europa	Tvarkos skalės kintamasis
20.	L12E_06	L12E_06	Įsidarbinimo galimybių vertinimas. Vakarų Europa	1 - Labai blogai, 2, 3, 4, 5 - Labai gerai	Įsidarbinimo galimybių vertinimas. Vakarų Europa	Tvarkos skalės kintamasis
21.	L12E_07	L12E_07	Ekonominės gerovės vertinimas. Vakarų Europa	1 - Labai blogai, 2, 3, 4, 5 - Labai gerai	Ekonominės gerovės vertinimas. Vakarų Europa	Tvarkos skalės kintamasis
22.	L12E_08	L12E_08	Kultūros, meno vertinimas. Vakarų Europa	1 - Labai blogai, 2, 3, 4, 5 - Labai gerai	Kultūros, meno vertinimas. Vakarų Europa	Tvarkos skalės kintamasis

Nr.	Kintamųjų vardai	Kintamųjų vardai originaliame faile	Kintamųjų žymės	Kintamųjų reikšmių žymės	Kintamųjų prasmė ir turinys	Duomenų analizei reikalingi paaiškinimai
23.	L12E_09	L12E_09	Žmogaus teisių vertinimas. Vakarų Europa	1 - Labai blogai, 2, 3, 4, 5 - Labai gerai	Žmogaus teisių vertinimas. Vakarų Europa	Tvarkos skalės kintamasis
24.	L12E_10	L12E_10	Demokratijos lygio vertinimas. Vakarų Europa	1 - Labai blogai, 2, 3, 4, 5 - Labai gerai	Demokratijos lygio vertinimas. Vakarų Europa	Tvarkos skalės kintamasis
25.	L12E_11	L12E_11	Laisvės (galimybė reikšti savo mintis, užsiimti veikla, keliauti ir t.t.) vertinimas. Vakarų Europa	1 - Labai blogai, 2, 3, 4, 5 - Labai gerai	Laisvės (galimybė reikšti savo mintis, užsiimti veikla, keliauti ir t.t.) vertinimas. Vakarų Europa	Tvarkos skalės kintamasis
26.	L12E_12	L12E_12	Teisingumo užtikrinimo vertinimas. Vakarų Europa	1 - Labai blogai, 2, 3, 4, 5 - Labai gerai	Teisingumo užtikrinimo vertinimas. Vakarų Europa	Tvarkos skalės kintamasis
27.	L12E_13	L12E_13	Lygių galimybių visiems visuomenės nariams vertinimas. Vakarų Europa	1 - Labai blogai, 2, 3, 4, 5 - Labai gerai	Lygių galimybių visiems visuomenės nariams vertinimas. Vakarų Europa	Tvarkos skalės kintamasis
28.	L12E_14	L12E_14	Moralės normų laikymosi vertinimas. Vakarų Europa	1 - Labai blogai, 2, 3, 4, 5 - Labai gerai	Moralės normų laikymosi vertinimas. Vakarų Europa	Tvarkos skalės kintamasis
29.	L12J_01	L12J_01	Politinės santvarkos vertinimas. JAV	1 - Labai blogai, 2, 3, 4, 5 - Labai gerai	Politinės santvarkos vertinimas. JAV	Tvarkos skalės kintamasis
30.	L12J_02	L12J_02	Švietimo sistemos vertinimas. JAV	1 - Labai blogai, 2, 3, 4, 5 - Labai gerai	Švietimo sistemos vertinimas. JAV	Tvarkos skalės kintamasis
31.	L12J_03	L12J_03	Sveikatos apsaugos sistemos vertinimas. JAV	1 - Labai blogai, 2, 3, 4, 5 - Labai gerai	Sveikatos apsaugos sistemos vertinimas. JAV	Tvarkos skalės kintamasis
32.	L12J_04	L12J_04	Socialinės apsaugos vertinimas. JAV	1 - Labai blogai, 2, 3, 4, 5 - Labai gerai	Socialinės apsaugos vertinimas. JAV	Tvarkos skalės kintamasis
33.	L12J_05	L12J_05	Saugumo, teisėsaugos darbo vertinimas. JAV	1 - Labai blogai, 2, 3, 4, 5 - Labai gerai	Saugumo, teisėsaugos darbo vertinimas. JAV	Tvarkos skalės kintamasis
34.	L12J_06	L12J_06	Įsidarbinimo galimybių vertinimas. JAV	1 - Labai blogai, 2, 3, 4, 5 - Labai gerai	Įsidarbinimo galimybių vertinimas. JAV	Tvarkos skalės kintamasis
35.	L12J_07	L12J_07	Ekonominės gerovės vertinimas. JAV	1 - Labai blogai, 2, 3, 4, 5 - Labai gerai	Ekonominės gerovės vertinimas. JAV	Tvarkos skalės kintamasis

Nr.	Kintamųjų vardai	Kintamųjų vardai originaliame faile	Kintamųjų žymės	Kintamųjų reikšmių žymės	Kintamųjų prasmė ir turinys	Duomenų analizei reikalingi paaiškinimai
36.	L12J_08	L12J_08	Kultūros, meno vertinimas. JAV	1 - Labai blogai, 2, 3, 4, 5 - Labai gerai	Kultūros, meno vertinimas. JAV	Tvarkos skalės kintamasis
37.	L12J_09	L12J_09	Žmogaus teisių vertinimas. JAV	1 - Labai blogai, 2, 3, 4, 5 - Labai gerai	Žmogaus teisių vertinimas. JAV	Tvarkos skalės kintamasis
38.	L12J_10	L12J_10	Demokratijos lygio vertinimas. JAV	1 - Labai blogai, 2, 3, 4, 5 - Labai gerai	Demokratijos lygio vertinimas. JAV	Tvarkos skalės kintamasis
39.	L12J_11	L12J_11	Laisvės (galimybė reikšti savo mintis, užsiimti veikla, keliauti ir t.t.) vertinimas. JAV	1 - Labai blogai, 2, 3, 4, 5 - Labai gerai	Laisvės (galimybė reikšti savo mintis, užsiimti veikla, keliauti ir t.t.) vertinimas. JAV	Tvarkos skalės kintamasis
40.	L12J_12	L12J_12	Teisingumo užtikrinimo vertinimas. JAV	1 - Labai blogai, 2, 3, 4, 5 - Labai gerai	Teisingumo užtikrinimo vertinimas. JAV	Tvarkos skalės kintamasis
41.	L12J_13	L12J_13	Lygių galimybių visiems visuomenės nariams vertinimas. JAV	1 - Labai blogai, 2, 3, 4, 5 - Labai gerai	Lygių galimybių visiems visuomenės nariams vertinimas. JAV	Tvarkos skalės kintamasis
42.	L12J_14	L12J_14	Moralės normų laikymosi vertinimas. JAV	1 - Labai blogai, 2, 3, 4, 5 - Labai gerai	Moralės normų laikymosi vertinimas. JAV	Tvarkos skalės kintamasis
43.	L12L_01	L12L_01	Politinės santvarkos vertinimas. Dabartinė Lietuva	1 - Labai blogai, 2, 3, 4, 5 - Labai gerai	Politinės santvarkos vertinimas. Dabartinė Lietuva	Tvarkos skalės kintamasis
44.	L12L_02	L12L_02	Švietimo sistemos vertinimas. Dabartinė Lietuva	1 - Labai blogai, 2, 3, 4, 5 - Labai gerai	Švietimo sistemos vertinimas. Dabartinė Lietuva	Tvarkos skalės kintamasis
45.	L12L_03	L12L_03	Sveikatos apsaugos sistemos vertinimas. Dabartinė Lietuva	1 - Labai blogai, 2, 3, 4, 5 - Labai gerai	Sveikatos apsaugos sistemos vertinimas. Dabartinė Lietuva	Tvarkos skalės kintamasis
46.	L12L_04	L12L_04	Socialinės apsaugos vertinimas. Dabartinė Lietuva	1 - Labai blogai, 2, 3, 4, 5 - Labai gerai	Socialinės apsaugos vertinimas. Dabartinė Lietuva	Tvarkos skalės kintamasis

Nr.	Kintamųjų vardai	Kintamųjų vardai originaliame faile	Kintamųjų žymės	Kintamųjų reikšmių žymės	Kintamųjų prasmė ir turinys	Duomenų analizei reikalingi paaiškinimai
47.	L12L_05	L12L_05	Saugumo, teisėsaugos darbo vertinimas. Dabartinė Lietuva	1 - Labai blogai, 2, 3, 4, 5 - Labai gerai	Saugumo, teisėsaugos darbo vertinimas. Dabartinė Lietuva	Tvarkos skalės kintamasis
48.	L12L_06	L12L_06	Įsidarbinimo galimybių vertinimas. Dabartinė Lietuva	1 - Labai blogai, 2, 3, 4, 5 - Labai gerai	Įsidarbinimo galimybių vertinimas. Dabartinė Lietuva	Tvarkos skalės kintamasis
49.	L12L_07	L12L_07	Ekonominės gerovės vertinimas. Dabartinė Lietuva	1 - Labai blogai, 2, 3, 4, 5 - Labai gerai	Ekonominės gerovės vertinimas. Dabartinė Lietuva	Tvarkos skalės kintamasis
50.	L12L_08	L12L_08	Kultūros, meno vertinimas. Dabartinė Lietuva	1 - Labai blogai, 2, 3, 4, 5 - Labai gerai	Kultūros, meno vertinimas. Dabartinė Lietuva	Tvarkos skalės kintamasis
51.	L12L_09	L12L_09	Žmogaus teisių vertinimas. Dabartinė Lietuva	1 - Labai blogai, 2, 3, 4, 5 - Labai gerai	Žmogaus teisių vertinimas. Dabartinė Lietuva	Tvarkos skalės kintamasis
52.	L12L_10	L12L_10	Demokratijos lygio vertinimas. Dabartinė Lietuva	1 - Labai blogai, 2, 3, 4, 5 - Labai gerai	Demokratijos lygio vertinimas. Dabartinė Lietuva	Tvarkos skalės kintamasis
53.	L12L_11	L12L_11	Laisvės (galimybė reikšti savo mintis, užsiimti veikla, keliauti ir t.t.) vertinimas. Dabartinė Lietuva	1 - Labai blogai, 2, 3, 4, 5 - Labai gerai	Laisvės (galimybė reikšti savo mintis, užsiimti veikla, keliauti ir t.t.) vertinimas. Dabartinė Lietuva	Tvarkos skalės kintamasis
54.	L12L_12	L12L_12	Teisingumo užtikrinimo vertinimas. Dabartinė Lietuva	1 - Labai blogai, 2, 3, 4, 5 - Labai gerai	Teisingumo užtikrinimo vertinimas. Dabartinė Lietuva	Tvarkos skalės kintamasis
55.	L12L_13	L12L_13	Lygių galimybių visiems visuomenės nariams vertinimas. Dabartinė Lietuva	1 - Labai blogai, 2, 3, 4, 5 - Labai gerai	Lygių galimybių visiems visuomenės nariams vertinimas. Dabartinė Lietuva	Tvarkos skalės kintamasis

Nr.	Kintamųjų vardai	Kintamųjų vardai originaliame faile	Kintamųjų žymės	Kintamųjų reikšmių žymės	Kintamųjų prasmė ir turinys	Duomenų analizei reikalingi paaiškinimai
56.	L12L_14	L12L_14	Moralės normų laikymosi vertinimas. Dabartinė Lietuva	1 - Labai blogai, 2, 3, 4, 5 - Labai gerai	Moralės normų laikymosi vertinimas. Dabartinė Lietuva	Tvarkos skalės kintamasis
57.	L12R_01	L12R_01	Politinės santvarkos vertinimas. Dabartinė Rusija	1 - Labai blogai, 2, 3, 4, 5 - Labai gerai	Politinės santvarkos vertinimas. Dabartinė Rusija	Tvarkos skalės kintamasis
58.	L12R_02	L12R_02	Švietimo sistemos vertinimas. Dabartinė Rusija	1 - Labai blogai, 2, 3, 4, 5 - Labai gerai	Švietimo sistemos vertinimas. Dabartinė Rusija	Tvarkos skalės kintamasis
59.	L12R_03	L12R_03	Sveikatos apsaugos sistemos vertinimas. Dabartinė Rusija	1 - Labai blogai, 2, 3, 4, 5 - Labai gerai	Sveikatos apsaugos sistemos vertinimas. Dabartinė Rusija	Tvarkos skalės kintamasis
60.	L12R_04	L12R_04	Socialinės apsaugos vertinimas. Dabartinė Rusija	1 - Labai blogai, 2, 3, 4, 5 - Labai gerai	Socialinės apsaugos vertinimas. Dabartinė Rusija	Tvarkos skalės kintamasis
61.	L12R_05	L12R_05	Saugumo, teisėsaugos darbo vertinimas. Dabartinė Rusija	1 - Labai blogai, 2, 3, 4, 5 - Labai gerai	Saugumo, teisėsaugos darbo vertinimas. Dabartinė Rusija	Tvarkos skalės kintamasis
62.	L12R_06	L12R_06	Įsidarbinimo galimybių vertinimas. Dabartinė Rusija	1 - Labai blogai, 2, 3, 4, 5 - Labai gerai	Įsidarbinimo galimybių vertinimas. Dabartinė Rusija	Tvarkos skalės kintamasis
63.	L12R_07	L12R_07	Ekonominės gerovės vertinimas. Dabartinė Rusija	1 - Labai blogai, 2, 3, 4, 5 - Labai gerai	Ekonominės gerovės vertinimas. Dabartinė Rusija	Tvarkos skalės kintamasis
64.	L12R_08	L12R_08	Kultūros, meno vertinimas. Dabartinė Rusija	1 - Labai blogai, 2, 3, 4, 5 - Labai gerai	Kultūros, meno vertinimas. Dabartinė Rusija	Tvarkos skalės kintamasis
65.	L12R_09	L12R_09	Žmogaus teisių vertinimas. Dabartinė Rusija	1 - Labai blogai, 2, 3, 4, 5 - Labai gerai	Žmogaus teisių vertinimas. Dabartinė Rusija	Tvarkos skalės kintamasis

Nr.	Kintamųjų vardai	Kintamųjų vardai originaliame faile	Kintamųjų žymės	Kintamųjų reikšmių žymės	Kintamųjų prasmė ir turinys	Duomenų analizei reikalingi paaiškinimai
66.	L12R_10	L12R_10	Demokratijos lygio vertinimas. Dabartinė Rusija	1 - Labai blogai, 2, 3, 4, 5 - Labai gerai	Demokratijos lygio vertinimas. Dabartinė Rusija	Tvarkos skalės kintamasis
67.	L12R_11	L12R_11	Laisvės (galimybė reikšti savo mintis, užsiimti veikla, keliauti ir t.t.) vertinimas. Dabartinė Rusija	1 - Labai blogai, 2, 3, 4, 5 - Labai gerai	Laisvės (galimybė reikšti savo mintis, užsiimti veikla, keliauti ir t.t.) vertinimas. Dabartinė Rusija	Tvarkos skalės kintamasis
68.	L12R_12	L12R_12	Teisingumo užtikrinimo vertinimas. Dabartinė Rusija	1 - Labai blogai, 2, 3, 4, 5 - Labai gerai	Teisingumo užtikrinimo vertinimas. Dabartinė Rusija	Tvarkos skalės kintamasis
69.	L12R_13	L12R_13	Lygių galimybių visiems visuomenės nariams vertinimas. Dabartinė Rusija	1 - Labai blogai, 2, 3, 4, 5 - Labai gerai	Lygių galimybių visiems visuomenės nariams vertinimas. Dabartinė Rusija	Tvarkos skalės kintamasis
70.	L12R_14	L12R_14	Moralės normų laikymosi vertinimas. Dabartinė Rusija	1 - Labai blogai, 2, 3, 4, 5 - Labai gerai	Moralės normų laikymosi vertinimas. Dabartinė Rusija	Tvarkos skalės kintamasis
71.	L12B_01	L12B_01	Politinės santvarkos vertinimas. Dabartinė Baltarusija	1 - Labai blogai, 2, 3, 4, 5 - Labai gerai	Politinės santvarkos vertinimas. Dabartinė Baltarusija	Tvarkos skalės kintamasis
72.	L12B_02	L12B_02	Švietimo sistemos vertinimas. Dabartinė Baltarusija	1 - Labai blogai, 2, 3, 4, 5 - Labai gerai	Švietimo sistemos vertinimas. Dabartinė Baltarusija	Tvarkos skalės kintamasis
73.	L12B_03	L12B_03	Sveikatos apsaugos sistemos vertinimas. Dabartinė Baltarusija	1 - Labai blogai, 2, 3, 4, 5 - Labai gerai	Sveikatos apsaugos sistemos vertinimas. Dabartinė Baltarusija	Tvarkos skalės kintamasis
74.	L12B_04	L12B_04	Socialinės apsaugos vertinimas. Dabartinė Baltarusija	1 - Labai blogai, 2, 3, 4, 5 - Labai gerai	Socialinės apsaugos vertinimas. Dabartinė Baltarusija	Tvarkos skalės kintamasis

Nr.	Kintamųjų vardai	Kintamųjų vardai originaliame faile	Kintamųjų žymės	Kintamųjų reikšmių žymės	Kintamųjų prasmė ir turinys	Duomenų analizei reikalingi paaiškinimai
75.	L12B_05	L12B_05	Saugumo, teisėsaugos darbo vertinimas. Dabartinė Baltarusija	1 - Labai blogai, 2, 3, 4, 5 - Labai gerai	Saugumo, teisėsaugos darbo vertinimas. Dabartinė Baltarusija	Tvarkos skalės kintamasis
76.	L12B_06	L12B_06	Įsidarbinimo galimybių vertinimas. Dabartinė Baltarusija	1 - Labai blogai, 2, 3, 4, 5 - Labai gerai	Įsidarbinimo galimybių vertinimas. Dabartinė Baltarusija	Tvarkos skalės kintamasis
77.	L12B_07	L12B_07	Ekonominės gerovės vertinimas. Dabartinė Baltarusija	1 - Labai blogai, 2, 3, 4, 5 - Labai gerai	Ekonominės gerovės vertinimas. Dabartinė Baltarusija	Tvarkos skalės kintamasis
78.	L12B_08	L12B_08	Kultūros, meno vertinimas. Dabartinė Baltarusija	1 - Labai blogai, 2, 3, 4, 5 - Labai gerai	Kultūros, meno vertinimas. Dabartinė Baltarusija	Tvarkos skalės kintamasis
79.	L12B_09	L12B_09	Žmogaus teisių vertinimas. Dabartinė Baltarusija	1 - Labai blogai, 2, 3, 4, 5 - Labai gerai	Žmogaus teisių vertinimas. Dabartinė Baltarusija	Tvarkos skalės kintamasis
80.	L12B_10	L12B_10	Demokratijos lygio vertinimas. Dabartinė Baltarusija	1 - Labai blogai, 2, 3, 4, 5 - Labai gerai	Demokratijos lygio vertinimas. Dabartinė Baltarusija	Tvarkos skalės kintamasis
81.	L12B_11	L12B_11	Laisvės (galimybė reikšti savo mintis, užsiimti veikla, keliauti ir t.t.) vertinimas. Dabartinė Baltarusija	1 - Labai blogai, 2, 3, 4, 5 - Labai gerai	Laisvės (galimybė reikšti savo mintis, užsiimti veikla, keliauti ir t.t.) vertinimas. Dabartinė Baltarusija	Tvarkos skalės kintamasis
82.	L12B_12	L12B_12	Teisingumo užtikrinimo vertinimas. Dabartinė Baltarusija	1 - Labai blogai, 2, 3, 4, 5 - Labai gerai	Teisingumo užtikrinimo vertinimas. Dabartinė Baltarusija	Tvarkos skalės kintamasis
83.	L12B_13	L12B_13	Lygių galimybių visiems visuomenės nariams vertinimas. Dabartinė Baltarusija	1 - Labai blogai, 2, 3, 4, 5 - Labai gerai	Lygių galimybių visiems visuomenės nariams vertinimas. Dabartinė Baltarusija	Tvarkos skalės kintamasis

Nr.	Kintamųjų vardai	Kintamųjų vardai originaliame faile	Kintamųjų žymės	Kintamųjų reikšmių žymės	Kintamųjų prasmė ir turinys	Duomenų analizei reikalingi paaiškinimai
84.	L12B_14	L12B_14	Moralės normų laikymosi vertinimas. Dabartinė Baltarusija	1 - Labai blogai, 2, 3, 4, 5 - Labai gerai	Moralės normų laikymosi vertinimas. Dabartinė Baltarusija	Tvarkos skalės kintamasis

Toliau laikmenoje pateikti sudarytų skalių kintamieji, kurių iš viso yra 12.

Nr.	Kintamųjų vardai	Kintamųjų žymės	Kintamųjų reikšmių žymės	Kintamųjų prasmė ir turinys	Duomenų analizei reikalingi paaiškinimai
1.	tsrs_soc	Socialinės ir kultūrinės gerovės vertinimas: TSRS	Nėra, kiekybinis kintamasis	Skalė: Socialinės ir kultūrinės gerovės vertinimas: TSRS	Skalė sudaryta iš originalių kintamųjų: L12T_02 L12T_03 L12T_04 L12T_06 L12T_08. Skalės reikšmės yra originalių kintamųjų reikšmių vidurkis.
2.	tsrs_pol	Politinės gerovės vertinimas: TSRS	Nėra, kiekybinis kintamasis	Skalė: Politinės gerovės vertinimas: TSRS	Skalė sudaryta iš originalių kintamųjų: L12T_01 L12T_09 L12T_10 L12T_11 L12T_12 L12T_13. Skalės reikšmės yra originalių kintamųjų reikšmių vidurkis.
3.	veur_soc	Socialinės ir kultūrinės gerovės vertinimas: Vakarų Europa	Nėra, kiekybinis kintamasis	Skalė: Socialinės ir kultūrinės gerovės vertinimas: Vakarų Europa	Skalė sudaryta iš originalių kintamųjų: L12E_02 L12E_03 L12E_04 L12E_06 L12E_08. Skalės reikšmės yra originalių kintamųjų reikšmių vidurkis.

Nr.	Kintamųjų vardai	Kintamųjų žymės	Kintamųjų reikšmių žymės	Kintamųjų prasmė ir turinys	Duomenų analizei reikalingi paaiškinimai
4.	veur_pol	Politinės gerovės vertinimas: Vakarų Europa	Nėra, kiekybinis kintamasis	Skalė: Politinės gerovės vertinimas: Vakarų Europa	Skalė sudaryta iš originalių kintamųjų: L12E_01 L12E_09 L12E_10 L12E_11 L12E_12 L12E_13. Skalės reikšmės yra originalių kintamųjų reikšmių vidurkis.
5.	jav_soc	Socialinės ir kultūrinės gerovės vertinimas: JAV	Nėra, kiekybinis kintamasis	Skalė: Socialinės ir kultūrinės gerovės vertinimas: JAV	Skalė sudaryta iš originalių kintamųjų: L12J_02 L12J_03 L12J_04 L12J_06 L12J_08. Skalės reikšmės yra originalių kintamųjų reikšmių vidurkis.
6.	jav_pol	Politinės gerovės vertinimas: JAV	Nėra, kiekybinis kintamasis	Skalė: Politinės gerovės vertinimas: JAV	Skalė sudaryta iš originalių kintamųjų: L12J_01 L12J_09 L12J_10 L12J_11 L12J_12 L12J_13. Skalės reikšmės yra originalių kintamųjų reikšmių vidurkis.
7.	liet_soc	Socialinės ir kultūrinės gerovės vertinimas: Lietuva	Nėra, kiekybinis kintamasis	Skalė: Socialinės ir kultūrinės gerovės vertinimas: Lietuva	Skalė sudaryta iš originalių kintamųjų: L12L_02 L12L_03 L12L_04 L12L_06 L12L_08. Skalės reikšmės yra originalių kintamųjų reikšmių vidurkis.
8.	liet_pol	Politinės gerovės vertinimas: Lietuva	Nėra, kiekybinis kintamasis	Skalė: Politinės gerovės vertinimas: Lietuva	Skalė sudaryta iš originalių kintamųjų: L12L_01 L12L_09 L12L_10 L12L_11 L12L_12 L12L_13. Skalės reikšmės yra originalių kintamųjų reikšmių vidurkis.

Nr.	Kintamųjų vardai	Kintamųjų žymės	Kintamųjų reikšmių žymės	Kintamųjų prasmė ir turinys	Duomenų analizei reikalingi paaiškinimai
9.	rus_soc	Socialinės ir kultūrinės gerovės vertinimas: Rusija	Nėra, kiekybinis kintamasis	Skalė: Socialinės ir kultūrinės gerovės vertinimas: Rusija	Skalė sudaryta iš originalių kintamųjų: L12R_02 L12R_03 L12R_04 L12R_06 L12R_08. Skalės reikšmės yra originalių kintamųjų reikšmių vidurkis.
10.	rus_pol	Politinės gerovės vertinimas: Rusija	Nėra, kiekybinis kintamasis	Skalė: Politinės gerovės vertinimas: Rusija	Skalė sudaryta iš originalių kintamųjų: L12R_01 L12R_09 L12R_10 L12R_11 L12R_12 L12R_13. Skalės reikšmės yra originalių kintamųjų reikšmių vidurkis.
11.	bltr_soc	Socialinės ir kultūrinės gerovės vertinimas: Baltarusija	Nėra, kiekybinis kintamasis	Skalė: Socialinės ir kultūrinės gerovės vertinimas: Baltarusija	Skalė sudaryta iš originalių kintamųjų: L12B_02 L12B_03 L12B_04 L12B_06 L12B_08. Skalės reikšmės yra originalių kintamųjų reikšmių vidurkis.
12.	bltr_pol	Politinės gerovės vertinimas: Baltarusija	Nėra, kiekybinis kintamasis	Skalė: Politinės gerovės vertinimas: Baltarusija	Skalė sudaryta iš originalių kintamųjų: L12B_01 L12B_09 L12B_10 L12B_11 L12B_12 L12B_13. Skalės reikšmės yra originalių kintamųjų reikšmių vidurkis.

5. 2007 M. „FREEDOM IN THE WORLD“ TYRIMO APRAŠYMAS IR KINTAMŲJŲ SĄVADAS BEI PAAIŠKINIMAI

Šiame skyriuje pateikiamas Laisvės pasaulyje 2007 m. tyrimo (toliau – LPT) aprašymas: tyrimo tikslai, objektas, vykdymo istorija, pagrindiniai rezultatai ir teorinės perspektyvos. Be to, pateikiamas šio tyrimo kintamųjų sąvadas bei paaiškinimai: kintamųjų kiekis, kintamųjų vardai, kintamųjų žymės, kintamųjų reikšmių žymės, kintamųjų prasmė ir turinys, duomenų analizei reikalingi paaiškinimai.

5.1. LPT tikslai ir objektas

Laisvės pasaulyje 2007 m. tyrimo (Freedom in the World 2007), atliekamo organizacijos Freedom House (JAV), pagrindu sudaryta duomenų bazė apima 193 šalių laisvių ir teisių lygio įvertinimus: politinių teisių, pilietinių teisių įverčius ir iš jų išplaukiantį šalies statuso (ar šalis yra laisva, iš dalies laisva ar nelaisva) įvertinimą.

Tyrimu įvertinamos (matuojamos) šalies laisvės ir teisės remiasi ekspertų vertinimais ir atspindi tai, kaip šalies gyventojai realiai patiria laisves ir teises, priešingai nei valdžios ir vyriausybės efektyvumo tiesioginiai matavimai.

Originaliame tyrime kiekviena šalis, o taip pat ir diskutuoti regionai yra ne tik yra įvertinami skaitmeniškai, bet ir yra aprašyta jų istorinis vystimasis bei esama teisių ir laisvių padėtis šalyje.

Tyrimo laisvė apibrėžiama kaip galimybė veikti spontaniškai srityse, kurios yra už valdžios ir kitų dominavimo centrų kontrolės. Laisvė matuojama įvertinant politines teises (dalyvavimo rinkimų procese galimybės: teisė nevaržomai balsuoti už aiškias alternatyvas teisėtuose rinkimuose, būti kandidatu ir būti išrinktam, prisijungti prie politinių partijų ir organizacijų, rinkti atstovus, kurie turi lemiamą įtaką vykdomoms politikoms ir kurie yra atskaitingi rinkėjams) ir pilietines laisves (saviraiškos ir įsitikinimų laisvė, asociacijų ir organizacijų sudarymo teisės, įstatymų valdžios buvimas, asmeninės autonomijos ir individo laisvės).

Laisvės pasaulyje (Freedom in the World) tyrimas pateikia kasmetinį laisvės visame pasaulyje įvertinimą. Tyrimu siekta įvertinti liberalios demokratijos padėtį, įvertinant patiriamų politinių laisvių ir pilietinių teisių kokybę ir kiekybę, kiekvienoje pasaulio šalyje.

Pagrindinis tyrimo objektas – liberalios demokratijos politinio režimo laisvių ir teisių kiekvienoje pasaulio šalyje lygis.

5.2. LPT vykdymo istorija

Organizacijos “Freedom House” pirmieji interviu („Balance Sheet of Freedom”) apie laisves įvairiose šalyse prasidėjo VI-jame XX a. dešimtmetyje. 1972 m. pradėtas išsamesnis kasmetinis tyrimas „The Comparative Study of Freedom“. Raymond Gastil, tyrėjas iš Vašingtono universiteto (Sietlas, JAV) sukūrė metodologiją, kaip vertinti politines teises ir pilietines laisves bei kaip atskirti laisvas nuo iš dalies laisvų ir nelaisvų šalių. Nuo 1978 m. tyrimo rezultatai pateikiami knygos forma pavadinimu “Freedom in the World”.

5.3. Pagrindiniai LPT rezultatai ir publikacijos

Tyrimo rezultatai (skaitiniai duomenys ir aprašymai) kasmet pateikiami knygoje “Freedom in the World” bei interneto svetainėje <http://www.freedomhouse.org>.

5.4. Teorinės LPT analizės perspektyvos

Tyrimo teorinė perspektyva remiasi liberalios demokratijos sąvoka ir jos apibrėžimu Visuotinėje žmogaus teisių deklaracijoje. Tyrime taikoma teorinė perspektyva ir metodologija detaliau aptariama skyrelyje apie tyrimo kintamuosius (žr. sekantį skyrelį) bei platesnėje perspektyvoje Gerardo L. Munck ir Jay Verkuilen (2002) ir Kenneth A. Bollen bei Pamela Paxton (2000).

Reikėtų pažymėti, kad Raymond Gastil ir Freedom House pateikiami šalių laisvės įverčiai nėra vieninteliai, skirtingais metodais gaunami įverčiai pateikiami Kenneth A. Bollen, Tatu Vanhanen, Leonard R. Sussman, Arthur Banks, Mark J. Gasiorowski, Adam Przeworski, Ted Gurr ir kitų mokslininkų darbuose.

LPT tyrimas įdomus politologams ir politikos mokslų studentams, siekiantiems įvertinti liberalios demokratijos sampratoje užfiksuotų laisvių ir teisių esamą būseną arba istorinį kitimą konkrečioje šalyje, regione ar visame pasaulyje.

5.5. 2007 m. LPT kintamųjų sąvadas bei paaiškinimai

Atrinktų kintamųjų kiekis: 12

Nr.	Kintamųjų vardai	Kintamųjų žymės	Kintamųjų reikšmių žymės	Kintamųjų prasmė ir turinys	Duomenų analizei reikalingi paaiškinimai
1.	salis	Šalis	193 Šalių pavadinimai anglų k.	Šalies pavadinimas	Vardų skalės kintamasis
2.	europa	Europos šalis	1 – Europos šalis, 0 – ne Europos šalis	Europai geografiškai arba kultūriškai priskiriamos šalys	Dvireikšmis kintamasis
3.	pt	Politinės teisės	1 rodo aukščiausią laisvės kiekį, 7 – mažiausią laisvės kiekį	Politinių teisių (angl. Political Rights) indeksas sudarytas iš 3 subskalių (A, B, C), kurių suma gali pasiekti max. 40 balų, pervedamų į 7 balų skalę*, kur 1 rodo aukščiausią laisvės kiekį, 7 – mažiausią laisvės kiekį tokiu būdu: 36-40 balų – 1; 30-35 balų - 2; 24-29 balų – 3; 18-23 balų – 4; 12-17 balų – 5; 6-11 balų – 6; 0-5 balų – 7.	Intervalų skalės kintamasis
4.	pl	Pilietinės teisės	1 rodo aukščiausią laisvės kiekį, 7 – mažiausią laisvės kiekį	Pilietinių teisių (angl. Civil Liberties) indeksas sudarytas iš 4 subskalių (D, E, F, G), kurių suma gali pasiekti max. 60 balų, pervedamų į 7 balų skalę**, kur 1 rodo aukščiausią laisvės kiekį, 7 – mažiausią laisvės kiekį tokiu būdu: 53-60 balų – 1; 44-52 balų - 2; 35-43 balų – 3; 26-34 balų – 4; 17-25 balų – 5; 8-16 balų – 6; 0-7 balų – 7.	Intervalų skalės kintamasis
5.	statusas	F (laisva), PF (iš dalies laisva), NF (nelaisva)	F – laisva (angl. Free), PF - iš dalies laisva (angl. Partly Free), NF - nelaisva (angl.	Sudaryta iš PR ir CL kintamųjų aritmetinio vidurkio: 1.0 - 2.5 balų - laisva (Free); 3.0 - 5.0 balų - iš dalies laisva (Partly Free);	Intervalų skalės kintamasis

Nr.	Kintamųjų vardai	Kintamųjų žymės	Kintamųjų reikšmių žymės	Kintamųjų prasmė ir turinys	Duomenų analizei reikalingi paaiškinimai
			Not Free)	5.5 - 7.0 balų - nelaisva (Not Free).	
6.	a	Rinkimų procesas	0 – nėra teisių; ... ; 12 – maksimalios teisės	Politinių teisių sub-indeksas, sudarytas 3 klausimų pagrindu, max. 12 balų***	Intervalų skalės kintamasis
7.	b	Politinis pliuralizmas ir dalyvavimas (politinių teisių sub-kategorija)	0 – nėra teisių; ... ; 16 – maksimalios teisės	Politinių teisių sub-indeksas, sudarytas 4 klausimų pagrindu, max. 16 balų***	Intervalų skalės kintamasis
8.	c	Valdžios funkcionavimas (politinių teisių sub-kategorija)	0 – nėra teisių; ... ; 12 – maksimalios teisės	Politinių teisių sub-indeksas, sudarytas 3 klausimų pagrindu, max. 12 balų***	Intervalų skalės kintamasis
9.	d	Saviraiškos ir įsitikinimų laisvė (pilietinių teisių sub-kategorija)	0 – nėra teisių; ... ; 16 – maksimalios teisės	Pilietinių teisių sub-indeksas, sudarytas 4 klausimų pagrindu, max. 16 balų***	Intervalų skalės kintamasis
10.	e	Asociacijų ir organizacijų sudarymo teisės (pilietinių teisių sub-kategorija)	0 – nėra teisių; ... ; 12 – maksimalios teisės	Pilietinių teisių sub-indeksas, sudarytas 3 klausimų pagrindu, max. 12 balų***	Intervalų skalės kintamasis
11.	f	Įstatymų valdžia (pilietinių teisių sub-kategorija)	0 – nėra teisių; ... ; 16 – maksimalios teisės	Pilietinių teisių sub-indeksas, sudarytas 4 klausimų pagrindu, max. 16 balų***	Intervalų skalės kintamasis
12.	g	Asmeninės autonomijos ir individo laisvės (pilietinių teisių sub-kategorija)	0 – nėra teisių; ... ; 16 – maksimalios teisės	Pilietinių teisių sub-indeksas, sudarytas 4 klausimų pagrindu, max. 16 balų	Intervalų skalės kintamasis

* Politinės teisės:

1 balas – Šalys ir teritorijos geriausiai užtikrina politines teises, pradedant laisvais ir teisingais rinkimais. Tie, kas yra išrenkami realiai valdo, yra konkuruojančios partijos ar kitos grupės, opozicija vaidina svarbų vaidmenį ir turi realią galią. Mažumos turi pakankamą savivaldą arba gali dalyvauti valdyme neformalaus konsensuso pagrindu.

2 balai– Šalys ir teritorijos yra mažiau laisvos negu gavusios 1 balą. Laisvės kokybę gali mažinti politinė korupcija, smurtas, politinė mažumų diskriminacija, užsienio ar karinių grupuočių įtaka politikai.

3, 4, 5 balai– Tie patys veiksniai kaip ir 2 balų grupėje gali mažinti laisvę šalyse su šiais politinių teisių balais. Kiti elementai gali būti pilietinis karas, didelis karinių struktūrų dalyvavimo politikoje laipsnis, išlikusi karališka galia, nesąžiningi rinkimai, vienos partijos dominavimas. Tačiau šalys ir

teritorijos turi kai kuriuos politinių teisių elementus: laisvę organizuoti panašias į politines gruputes, pakankamai laisvus referendumus ar kitas visuomenės įtakos valdžiai priemones.

6 balai– Šalys ir teritorijos yra valdomos karinių chuntų, vienpartinio diktato, religinių hierarchijų arba autokratų. Režimai gali leisti tik minimalius politinių teisių pasireiškimus, kaip nedidelio laipsnio mažumų atstovavimą ar autonomiją. Kai kurios valstybės yra tradicinės monarchijos, kurios politinių teisių trūkumą švelnina konsultacijų su gyventojais, politinės diskusijos toleravimo, pilietinių peticijų priėmimo priemonėmis.

7 balai– Politinės teisės neegzistuoja dėl ypatingai represinio režimo arba kai vyksta pilietinis karas represinio režimo sąlygomis. Šalys ir teritorijos šioje grupėje gali pasižymėti ypatingu smurto lygiu arba karo vadų, kurie dominuoja kaip politinė jėga, nesant autoritetingos, funkcionalios valdžios, valdymu.

**** Pilietinės laisvės:**

1 balas – Šalys ir teritorijos geriausiai užtikrina pilietines laisves: saviraiškos, susirinkimų, asociacijos, išsilavinimo ir religijos laisvę. Išsiskiria įstatymų valdžios sistemos stiprumu. Pasižymi laisva ekonomine veikla ir linkusios siekti galimybių lygybės.

2 balai– Šalys ir teritorijos turi netobulumų keliais pilietinių teisių aspektais, bet yra sąlyginai laisvos.

3, 4, 5 balai– Priespaudos lygis egzistuoja: cenzūra, politinis teroras, trukdymas laisvoms asociacijoms. Grupės, esančios opozicijoje, imasi politinio teroro.

6 balai– Šalyse ir teritorijose yra stipriai varžomos saviraiškos ir asociacijos laisvės, beveik visada yra politinių kalinių ir kitų politinio teroro pasireiškimų. Tik kelios dalinės teisės egzistuoja, pvz., kai kurios religinės ir socialinės laisvės, kai kurios labai ribotos verslo veiklos ir sąlyginai laisva privati diskusija.

7 balai– Šalys ir teritorijos neturi praktiškai jokios laisvės. Šias visuomenes apibūdina visaapimanti ir pagrįsta represijų baimė.

*** Detaliau žr. http://www.freedomhouse.org/template.cfm?page=351&ana_page=333&year=2007

6. 2004 M. „NAUJOJO BALTIJOS BAROMETRO“ TYRIMO APRAŠYMAS IR KINTAMŲJŲ SĄVADAS BEI PAAIŠKINIMAI

Šiame skyriuje pateikiamas tyrimo „Naujasis Baltijos barometras“ (toliau – NBB) 2004 m. aprašymas: tyrimo tikslai, objektas, vykdymo istorija, pagrindiniai rezultatai ir teorinės perspektyvos. Be to, pateikiamas šio tyrimo kintamųjų sąvadas bei paaiškinimai: kintamųjų kiekis, kintamųjų vardai, kintamųjų žymės, kintamųjų reikšmių žymės, kintamųjų prasmė ir turinys, duomenų analizei reikalingi paaiškinimai.

6.1. NBB tyrimo tikslai ir objektas

„Naujasis Baltijos barometras“ – tai vienu metu trijose Baltijos šalyse atliekamas sociologinis tyrimas apie politines nuostatas bei požiūrį į ekonomines reformas. Iš viso – apie 200 indikatorių. Tyrimai buvo atlikti: 1993, 1995, 1996, 2000, 2001, 2004 m. Analogiški tyrimai („Naujasis Demokratijos barometras“) atliekami visose pokomunistinėse šalyse. Projekto vadovas prof. Richard Rose, Strathclyde universitetas (Škotija). Tyrimas iš esmės savo indikatoriais panašus į Europos vertybių, Pasaulio vertybių tyrimus, tačiau jo pridėtinė vertė yra trijose Baltijos šalyse vykusių pokyčių užfiksavime nuo 1993 m. iki 2004 m.

Tyrimo duomenys skelbiami puslapyje www.balticvoices.org). Tyrimą Lietuvoje vykdė Visuomenės nuomonės ir rinkos tyrimų centras "Vilmorus" (duomenų rinkėjas, t.y., lauko darbams kontraktuota organizacija). Čia pateikiama informacija remiasi <http://www.vilmorus.lt> tinklapyje pateiktais duomenimis.

Pagrindinis tyrimo objektas – dabartinio ir sovietinio politinio bei ekonominio režimo vertinimai, ekonominės ir politinės padėties vertinimai, politiniai kairės – dešinės įsitikinimai, tapatumo klausimai, kalbos, etniniai santykiai, socialinės ir demografinės charakteristikos.

6.2. NBB tyrimo vykdymo istorija

NBB pradėtas 1993 m. ir atskleidė nuomones trijose daugiaetninėse visuomenėse – Lietuvoje, Latvijoje, Estijoje. Paskutinis tyrimas atliktas 2004 m.

6.3. Pagrindiniai NBB tyrimo rezultatai ir publikacijos

NBB tyrimo rezultatai skelbiami www.balticvoices.org tinklapyje.

6.4. Teorinės NBB tyrimo analizės perspektyvos

Tyrimas remiasi esmine politinės kultūros tyrimų idėja ir prielaida, kad visuomenės vertybės ir nuostatos yra demokratinių režimų išsilaikymo ir konsolidacijos pagrindas. Šiame tyrime pagrindinis klausimas yra demokratijos (arba režimo) palaikymo (pritarimo), išmatuoto įvairiomis formomis, lygis pokomunistinėse visuomenėse. Režimo palaikymo arba politinio palaikymo tyrimų teorinės ištakos sietinos su Easton (1965, 1975), Parsons (1967) darbais, iš kurių buvo išvystyta politinio palaikymo ir jo formų teorija arba greičiau tyrimų sritis. Iš esmės, šia teoriją sudaro šie pagrindiniai elementai: politinio palaikymo objektai (principai, institucijos, grupės ir pan.), kurie apsprendžia palaikymo objekto nekonkretumą (difuzinis palaikymas) arba konkretumą (specifinis palaikymas) ir ryšio su palaikomu objektu pobūdis (jausminis, vertinimo, kognityvinis (instrumentinis) ryšys).

6.1 lentelė. Politinio palaikymo teorijos elementai: objektai ir palaikymo pobūdis

Politinio palaikymo objektai	Palaikymo tipas	Palaikymo prasmė	Empiriniai indikatoriai	Papildomos palaikymo dimensijos (Fuchs 1993 p. 238; cit. Klingemann 1999)
Politinė bendruomenė	Difuzinis palaikymas	Jausminio prisirišimo prie nacionalinės bendruomenės lygis; pasiryžimas bendradarbiauti politiškai	Pasididžiavimo jausmas būti savo šalies piliečiu	Jausminiai, moraliniai, instrumentiniai politinio palaikymo aspektai
Režimas, režimo principai		Režimo kaip idealo ar principo palaikymas	Demokratinio režimo pasirinkimas, lyginant su kitais režimais	
Režimo vertybės, režimo veiksmingumas		Režimo praktinis veikimas konkrečioje šalyje ir konkrečiu laiku	Pasitenkinimas tuo, kaip demokratija veikia šalyje; pasitenkinimas esama [namų ūkio] ekonomine situacija	
Režimo struktūra, režimo institucijos		Politinių institucijų (parlamento, politinių partijų, vyriausybės, prezidento, teismų, policijos, biurokratijos, kt.) palaikymas	Asmeninis pasitikėjimas institucijomis	
Valdžia, politiniai veikėjai	Specifinis palaikymas	Konkrečios partijos ar asmens palaikymas	Pasitenkinimas tuo, kaip dabar esantys valdžioje asmenys tvarko šalies reikalus	

NBB tyrimo duomenys dažniausiai naudojami siekiant įvertinti tuo metu buvusį demokratijos palaikymo lygį, palaikymo lygio kaitą, palaikymo lygio galimas priežastis (įvertinant, su kokiais kitais kintamaisiais demokratijos palaikymas koreliuoja). Palaikymo priežastys aiškinamos kultūriniu palikimu, kai politinis įvairių politinių objektų palaikymas išplaukia iš esamų istorinių-kultūrinių normų ir jas perduodančios individo socializacijos („kultūrinė perspektyva“), o institucinis (arba „racionalaus pasirinkimo perspektyva“) aiškinimas remiasi tuo, kad politinis palaikymas atsiranda sėkmingai veikiant institucijoms.

6.5. 2004 m. NBB kintamųjų sąvadas bei paaiškinimai

Atrinktų kintamųjų kiekis: 25

Nr.	Kintamųjų vardai	Kintamųjų vardai originaliame faile	Kintamųjų žymės	Kintamųjų reikšmių žymės	Kintamųjų prasmė ir turinys	Duomenų analizei reikalingi paaiškinimai
1.	politik	b01	Kiek jūs domitės politika?	1 Labai domiuosi 2 Kažkiek domiuosi 3 Mažai domiuosi 4 Visiškai nesidomiu 9 - Nežinau	Domėjimo politika lygis	Tvarkos skalės kintamasis
2.	nvo	d01	Ar jūs priklausote kokiai nors sporto, meno, bendruomeninei ar labdarinei organizacijai?	1 – taip, 2 – ne	Priklausymas kokiai nors NVO (išskyrus partijas)	Dvireikšmis kintamasis
3.	partpal	d02	Ar palaikote kurią nors partiją?	1 – taip, 2 – ne	Partijos palaikymas (formaliai - per narystę - arba neformaliai – saviidentifikacija)	Dvireikšmis kintamasis
4.	teismai	d04_01	Pasitikėjimas institucijomis: teismai	1 – jokio pasitikėjimo; ... ; 7 - didelis pasitikėjimas; 9 – nežino	Pasitikėjimas institucijomis: teismai	Tvarkos skalės kintamasis
5.	partijos	d04_02	Pasitikėjimas institucijomis: politinės partijos	1 – jokio pasitikėjimo; ... ; 7 - didelis pasitikėjimas; 9 – nežino	Pasitikėjimas institucijomis: politinės partijos	Tvarkos skalės kintamasis
6.	armija	d04_03	Pasitikėjimas institucijomis: kariuomenė	1 – jokio pasitikėjimo; ... ; 7 - didelis pasitikėjimas; 9 – nežino	Pasitikėjimas institucijomis: kariuomenė	Tvarkos skalės kintamasis
7.	seimas	d04_04	Pasitikėjimas institucijomis: Seimas	1 – jokio pasitikėjimo; ... ; 7 - didelis pasitikėjimas; 9 –	Pasitikėjimas institucijomis: Seimas	Tvarkos skalės kintamasis

Nr.	Kintamųjų vardai	Kintamųjų vardai originaliame faile	Kintamųjų žymės	Kintamųjų reikšmių žymės	Kintamųjų prasmė ir turinys	Duomenų analizei reikalingi paaiškinimai
				nežino		
8.	policija	d04_05	Pasitikėjimas institucijomis: policija	1 – jokio pasitikėjimo; ... ; 7 - didelis pasitikėjimas; 9 – nežino	Pasitikėjimas institucijomis: policija	Tvarkos skalės kintamasis
9.	prezid	d04_06	Pasitikėjimas institucijomis: Prezidentas	1 – jokio pasitikėjimo; ... ; 7 - didelis pasitikėjimas; 9 – nežino	Pasitikėjimas institucijomis: Prezidentas	Tvarkos skalės kintamasis
10.	profsaj	d04_07	Pasitikėjimas institucijomis: profesinės sąjungos	1 – jokio pasitikėjimo; ... ; 7 - didelis pasitikėjimas; 9 – nežino	Pasitikėjimas institucijomis: profesinės sąjungos	Tvarkos skalės kintamasis
11.	baznyčia	d04_08	Pasitikėjimas institucijomis: Bažnyčia	1 – jokio pasitikėjimo; ... ; 7 - didelis pasitikėjimas; 9 – nežino	Pasitikėjimas institucijomis: Bažnyčia	Tvarkos skalės kintamasis
12.	ekonom	a01b	Kaip jūs vertinate dabartinę Lietuvos ekonominę sistemą?	1 +100 2 +90 3 +80 4 +70 5 +60 6 +50 7 +40 8 +30 9 +20 10 +10 11 0 12 -10 13 -20 14 -30 15 -40 16 -50	Neigiamas esamos Lietuvos ekonominės sistemos vertinimas. Vienas iš režimo ekonominio veiksmingumo vertinimo matų. Naudojamas prognozuojant politinio palaikymo lygį, balsavimo pasirinkimus ir kt.	Intervalų skalės kintamasis

Nr.	Kintamųjų vardai	Kintamųjų vardai originaliame faile	Kintamųjų žymės	Kintamųjų reikšmių žymės	Kintamųjų prasmė ir turinys	Duomenų analizei reikalingi paaiškinimai
				17 -60 18 -70 19 -80 20 -90 21 -100 99 - Nežinau		
13.	ekonteig	a01b	Kaip jūs vertinate dabartinę Lietuvos ekonominę sistemą?	1 -100 2 -90 3 -80 4 -70 5 -60 6 -50 7 -40 8 -30 9 -20 10 -10 11 0 12 +10 13 +20 14 +30 15 +40 16 +50 17 +60 18 +70 19 +80 20 +90 21 +100	Teigiamas esamos Lietuvos ekonominės sistemos vertinimas. Reikšmės perkoduotas atvirkštine tvarka (geriausias vertinimas atitinka „21“, blogiausias - „1“). Vienas iš režimo ekonominio veiksmingumo vertinimo matų. Naudojamas prognozuojant politinio palaikymo lygį, balsavimo pasirinkimus ir kt.	Intervalų skalės kintamasis Kintamojo reikšmės perkoduotos atvirkščiai, kad didžiausia reikšmė žymėtų palankiausią ekonominės sistemos vertinimą
14.	amzius	g01a	Respondento amžius, metais		Respondento amžius, metais	Intervalų skalės kintamasis

Nr.	Kintamųjų vardai	Kintamųjų vardai originaliame faile	Kintamųjų žymės	Kintamųjų reikšmių žymės	Kintamųjų prasmė ir turinys	Duomenų analizei reikalingi paaiškinimai
15.	lytis	g02	Respondento lytis	1 – vyras, 0 – moteris	Respondento lytis	Dvireikšmis kintamasis
16.	pajamos	g07	Bendros šeimos pajamos paskutinį mėnesį		Bendros šeimos pajamos paskutinį mėnesį	Tvarkos skalės kintamasis
17.	religing	g08b	Kaip dažnai jūs einate į bažnyčią?	1 - Kartą per savaitę 2 - Viena ar du kartus per mėnesį 3 - Keletą kartų per metus 4 - Apytiksliai kartą per metus 5 - Rečiau negu kartą per metus 6 - Niekada neinu į bažnyčią	Respondento religingumo lygmuo	Tvarkos skalės kintamasis
18.	vietove	g11	Vietovė	1 - Vilnius 2 - Kaunas, Klaipėda, Šiauliai, Panevėžys 3 - Rajono centras 4 - Kiti miestai 5 - Kaimo vietovė	Vietovė, kurioje gyvena respondentas	Vardų skalės kintamasis
19.	korupcij		Korupcija lygio vertinimas	1 – dauguma pareigūnų yra korumpuoti; 0 – mažiau negu pusė pareigūnų yra korumpuoti	Dvireikšmis kintamasis, perskaičiuotas iš b11, „1“ priskiriant prasmę „4 Beveik visi pareigūnai yra korumpuoti“ ir „3 Dauguma pareigūnų yra korumpuoti“, o „0“ priskiriant reikšmės „1 Labai nedaug pareigūnų yra korumpuoti“, „2	Dvireikšmis kintamasis

Nr.	Kintamųjų vardai	Kintamųjų vardai originaliame faile	Kintamųjų žymės	Kintamųjų reikšmių žymės	Kintamųjų prasmė ir turinys	Duomenų analizei reikalingi paaiškinimai
					Mažiau negu pusė yra korumpuoti“	
20.	pasidem		Pasitikėjimas demokratinės valstybės institucijomis	Reikšmės intervale –[3;3]	Pasitikėjimo institucijomis kintamųjų pagrindu (d04_01, d04_02, d04_03, d04_04, d04_05, d04_06, d04_08), išskyrus pasitikėjimą profsajungomis (d04_07) gauto faktorinės (pagrindinių komponentų) analizės pirmojo faktoriaus, kuriame didžiausią svorį turi pasitikėjimas Seimu, politinėmis partijomis, teismais, policija, reikšmė	Intervalų skalės kintamasis
21.	pasitrad		Pasitikėjimas tradicinio autoriteto institucijomis	Reikšmės intervale –[3;3]	Pasitikėjimo institucijomis kintamųjų pagrindu (d04_01, d04_02, d04_03, d04_04, d04_05, d04_06, d04_08), išskyrus pasitikėjimą profsajungomis (d04_07) gauto faktorinės (pagrindinių	Intervalų skalės kintamasis

Nr.	Kintamųjų vardai	Kintamųjų vardai originaliame faile	Kintamųjų žymės	Kintamųjų reikšmių žymės	Kintamųjų prasmė ir turinys	Duomenų analizei reikalingi paaiškinimai
					komponenčių) analizės antrojo faktoriaus, kuriame didžiausią svorį turi pasitikėjimas Prezidentu, bažnyčia, ginkluotosiomis pajėgomis, reikšmė	
22.	politik	b01	Kiek jūs domitės politika?	1 Labai domiuosi 2 Kažkiek domiuosi 3 Mažai domiuosi 4 Visiškai nesidomiu 9 - Nežinau	Domėjimo politika lygis	Tvarkos skalės kintamasis
23.	nvo	d01	Ar jūs priklausote kokiai nors sporto, meno, bendruomeninei ar labdarinei organizacijai?	1 – taip, 2 – ne	Priklausymas kokiai nors NVO (išskyrus partijas)	Dvireikšmis kintamasis
24.	partpal	d02	Ar palaikote kurią nors partiją?	1 – taip, 2 – ne	Partijos palaikymas (formaliai - per narystę - arba neformaliai – saviidentifikacija)	Dvireikšmis kintamasis
25.	teismai	d04_01	Pasitikėjimas institucijomis: teismai	1 – jokio pasitikėjimo; ... ; 7 - didelis pasitikėjimas; 9 – nežino	Pasitikėjimas institucijomis: teismai	Tvarkos skalės kintamasis

7. LITERATŪRA

- Ajzen, I. ir M. Fishbein (1980). Understanding attitudes and predicting social behavior. Englewood Cliffs, NJ, Prentice-Hall.
- Bollen, K. A. ir P. Paxton (2000). "Subjective Measures of Liberal Democracy." *Comparative Political Studies* 33(1): 58-86.
- Easton, D. (1965). *A Systems Analysis of Political Life*. New York, Wiley.
- Easton, D. (1975). "A re-assessment of the concept of political support." *British Journal of Political Science* 5(4): 435-457.
- Fuchs, D. (1993). Trends of Political Support in the Federal Republic of Germany. In *Political Culture in Germany*. D. Berg-Schlosser and R. Rytlewski (Eds.). London, Macmillan: 232-268.
- Halman, L. (2001). *The European Values Study: A Third Wave*. Source book of the 1999/2000 European Values Study surveys. Tilburg, EVS/WORC.
- Inglehart, R. (1997). *Modernization and postmodernization : cultural, economic, and political change in 43 societies*. Princeton, N.J., Princeton University Press.
- Inglehart, R. ir W. E. Baker (2000). "Modernization, Cultural Change, and the Persistence of Traditional Values." *American Sociological Review* 65(1): 19-51.
- Inglehart, R. ir H. D. Klingemann (2000). Genes, Culture, Democracy, and Happiness. In *Subjective Well-being across Cultures*. E. Diener and M. Eunkook (Eds.). Cambridge, MIT Press: 165-183.
- Inter-university Consortium for Political and Social Research. (2000). "World Values Surveys and European Values Surveys, 1981-1984, 1990-1993, and 1995-1997." ICPSR. <http://webapp.icpsr.umich.edu/cocoon/ICPSR-STUDY/02790.xml>
- Klingemann, H.-D. (1999). Mapping Political Support in the 1990s: A Global Analysis. In *Critical Citizens: Global Support for Democratic Government*. P. Norris (Ed.). Oxford, Oxford University Press: 31-56.
- Munck, G. L. ir J. Verkuilen (2002). "Conceptualizing and Measuring Democracy." *Comparative Political Studies* 35(1): 5-34.
- Parsons, T. (1967). *Sociological Theory and Modern Society*. New York, Free Press.
- Schwartz, S. H. (1992). "Universals in the content and structure of values: Theoretical advances and empirical tests in 20 countries." *Advances in experimental social psychology* 25: 1-65.
- Welzel, C., R. Inglehart ir H.-D. Klingemann (2003). "The Theory of Human Development: A Cross-Cultural Analysis." *European Journal Of Political Research* 42(3): 341-379.

**II DALIS. ĮVADINIO KURSO Į STATISTINĘ ANALIZĘ
MOKOMOJI MEDŽIAGA**

Autorius:

doc. dr. Vytautas Janilionis

(parašas)

KAUNAS, 2008

1. ĮVADAS

„Įvadinis kursas į statistinę analizę“:

- suteikia teorinių statistikos žinių ir praktinių įgūdžių, būtinų analizuojant HSM empirinius duomenis;
- supažindina su pagrindiniais statistiniais duomenų analizės metodais ir jų sprendžiamais uždaviniais;
- ugdo gebėjimus parinkti tinkamus statistikos metodus duomenų analizei;
- ugdo gebėjimus interpretuoti statistinės analizės rezultatus, gautus naudojant šiuolaikinę duomenų analizės programinę įrangą SPSS arba STATA;
- padeda suprasti ir interpretuoti profesinėje literatūroje pateikiamus statistinių tyrimų rezultatus.

Kurso medžiagą sudaro devyni skyriai:

Aprašomoji statistika.

Pasikliautinieji intervalai.

Hipotezių tikrinimas.

Vidurkių palyginimas. Dispersinė analizė.

Požymių priklausomumo tyrimas. Koreliacinė analizė.

Regresinė analizė.

Klasterinė analizė.

Faktorinė analizė.

Daugiamačių skalių metodas (MDS).

Kiekviename skyriuje pateikta įvadinė teorinė medžiaga ir uždavinių sprendimo su SPSS arba STATA pavyzdžiai, iliustruojantys nagrinėjamos temos klausimų sprendimo būdus ir gautų rezultatų interpretavimą.

Mokomosios medžiagos bendra apimtis 145 puslapiai.

1. APRAŠOMOJI STATISTIKA

Aprašomoji statistika tai duomenų sisteminimo ir grafinio vaizdavimo būdai. Pagrindinis **aprašomosios statistikos tikslas** - glaustas ir vaizdus surinktų duomenų apibūdinimas. Duomenims sisteminti naudojamos skaitinės charakteristikos: dažniai, santykiniai dažniai, vidurkis, mediana, standartinis nuokrypis ir t.t. Grafiškai apibūdinti duomenis galima stulpelinėmis diagramomis, skritulinėmis diagramomis, žiedinėmis diagramomis, histogramomis ir t.t. Yra daugybė grafinio vaizdavimo būdų. Kuris būdas yra geriausias konkrečioje situacijoje, priklauso nuo daugybės faktorių. Duomenų grafiniame vaizdavime nepaprastai daug vietos kūrybai.

1.1. Dažniai

Tarkime, kad $\mathbf{X}$ yra stebimas atsitiktinis dydis (kintamasis), o $x=(x_1, \dots, x_n)$ yra šio atsitiktinio dydžio n stebėjimų rezultatai – paprastoji imtis. Surašę šią imtį nemažėjančia tvarka $x_{(1)} \leq \dots \leq x_{(n)}$ turėsime vadinamąją **variacinę seką**. Skirtumas tarp didžiausios ir mažiausios reikšmių

$$\mathbf{IP} = x_{(n)} - x_{(1)}$$

yra vadinamas **imties pločiu (range)**, čia $x_{(1)} = x_{\min}$, $x_{(n)} = x_{\max}$.

Pavyzdžiui, tarkime duota imtis $x=(1, 3, 2, 2, 4, 1, 2, 1, 3, 2)$. Imties didumas $n=10$, jos variacinė seka bus

$$1, 1, 1, 2, 2, 2, 2, 3, 3, 4,$$

o imties plotis $\mathbf{IP}=4-1=3$.

Matome, kad kai kurie imties duomenys gali pasikartoti. Skirtingus imties duomenis žymėsime raide $\mathbf{x}$ su indeksais. Tarkime, $\mathbf{x}_1, \dots, \mathbf{x}_r$ yra visi skirtingi imties duomenys (paprastai juos surašo didėjančia tvarka: $\mathbf{x}_1 < \dots < \mathbf{x}_r$). Pažymėkime m_i reikšmės $\mathbf{x}_i$ pasikartojimų imtyje skaičių. Skaičius m_i yra vadinamas reikšmės $\mathbf{x}_i$ **dažniu (frequency)**, $i=1, 2, \dots, r$. Suprantama $\sum_{i=1}^r m_i = n$. Reikšmių ir jų dažnių sąrašas

$\mathbf{x}_i$	$\mathbf{x}_1$	$\mathbf{x}_2$	...	$\mathbf{x}_r$
m_i	m_1	m_2	...	m_r

yra vadinamas **dažnių lentele (frequency table)** arba **dažnių skirstiniu (frequency distribution)**.

Paprastai iš anksto aišku, ar tiriamas atsitiktinis dydis $\mathbf{X}$ yra **diskretusis** ar **tolydusis**. Tarkime, $\mathbf{X}$ yra diskretusis atsitiktinis dydis. Tada $\mathbf{x}_1, \dots, \mathbf{x}_r$ yra galimos atsitiktinio dydžio $\mathbf{X}$ reikšmės. Nagrinėjamoje imtyje yra 4 skirtingos reikšmės: 1, 2, 3 ir 4. Užrašysime dažnių skirstinį

$\mathbf{x}_i$	1	2	3	4
m_i	3	4	2	1

Galimų $\mathbf{X}$ reikšmių gali būti ir daugiau, tačiau jos paprasčiausiai nepateko į imtį.

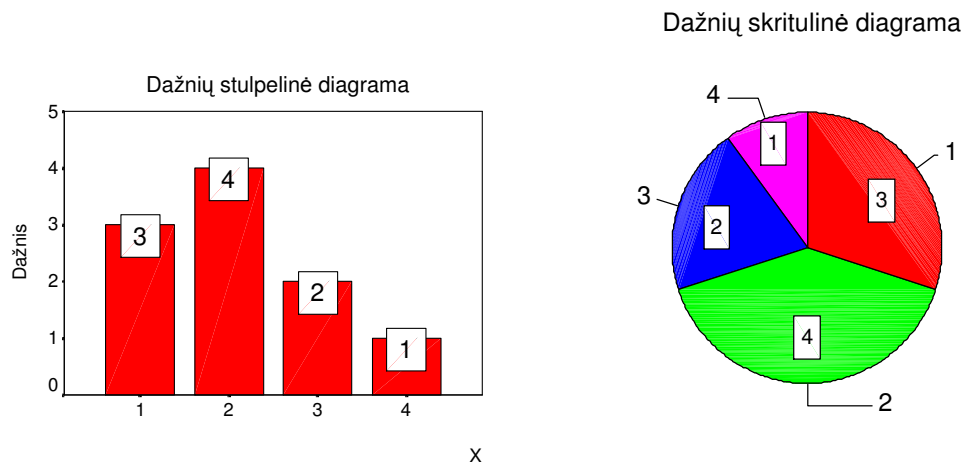
Reikšmės $\mathbf{x}_i$ **santykinis dažnis (relative frequency)** $W_i = \frac{m_i}{n}$ yra šios reikšmės tikimybės $p_i = P(\mathbf{X} = \mathbf{x}_i)$ įvertis, $i=1, 2, \dots, r$. Vadinasi, yra prasmė sudaryti **santykinių dažnių skirstinį**,

x_i	x_1	x_2	...	x_r
w_i	w_1	w_2	...	w_r

kuris paprastai yra vadinamas **empiriniu santykinų dažnių skirstiniu**. Suprantama, atsitiktinio dydžio X empirinis santykinų dažnių skirstinys yra tikrojo (populiacijos) skirstinio aproksimacija. Mūsų pavyzdyje empirinis dažnių skirstinys bus

x_i	1	2	3	4
w_i	0,3	0,4	0,2	0,1

Skirstinys dažnai vaizduojamas daugiakampiu, stulpeline diagrama arba skrituline diagrama. 1.1 pav. pateikti dažnių skirstinio grafinio vaizdavimo pavyzdžiai.



1.1 pav. Dažnių skirstinys pateiktas stulpeline ir skrituline dažnių diagramomis (SPSS)

Kai tiriamas atsitiktinis dydis X yra **tolydusis** imtis sutvarkoma į vadinamąją **intervalinę variacinę seką**. Ši seka dažnai naudojama ir tuo atveju, kai atsitiktinis dydis X yra diskretusis bet įgyja daug skirtingų reikšmių. Dalinių intervalų skaičių r rekomenduojama pasirinkti artimą reikšmei $1+3,32 \cdot \lg(n)$, čia n – imties didumas.

Suprantama, a_0 ir a_r gali būti parinkti taip, kad dalinio intervalo ilgis $h = \frac{a_r - a_0}{r}$ taip pat būtų „gražus“ skaičius. Dalinius intervalus pažymėkime $(a_{i-1}, a_i]$, o m_i imties duomenų, patekusių į intervalą $(a_{i-1}, a_i]$, skaičių $i=1, 2, \dots, r$. Suprantama, $\sum_{i=1}^r m_i = n$. Šių apskaičiavimų rezultatas

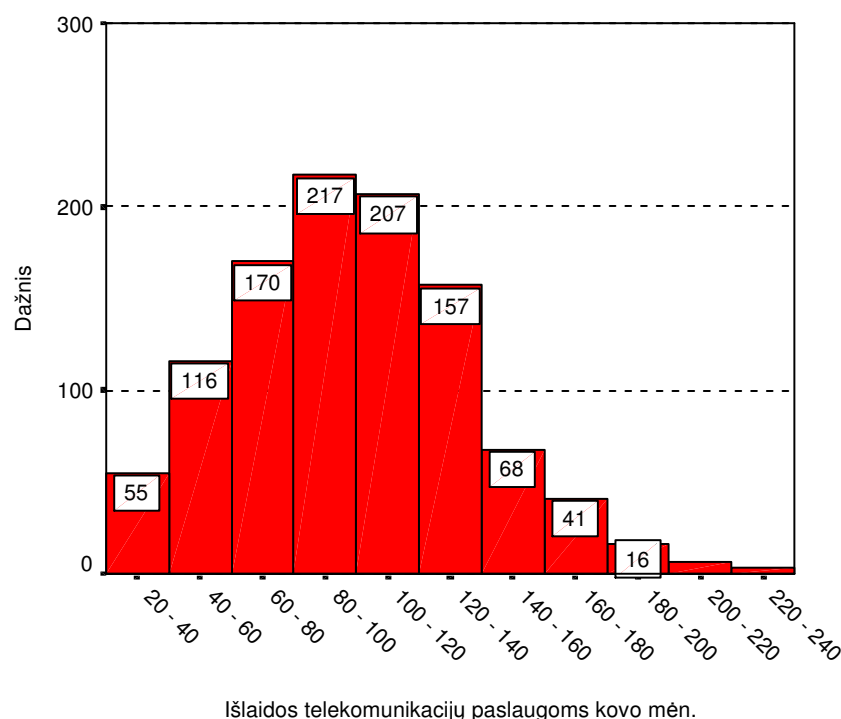
x	$(a_0; a_1]$	$(a_1; a_2]$	...	$(a_{r-1}; a_r]$
m_i	m_1	m_2	...	m_r

yra vadinamas **intervalinių dažnių skirstiniu**. Analogiškai sudarome **intervalinių santykinų dažnių skirstinį**

$\mathbf{x}$	$(a_0; a_1]$	$(a_1; a_2]$	...	$(a_{r-1}; a_r]$
$\mathbf{W_i}$	$\mathbf{W_1}$	$\mathbf{W_2}$	...	$\mathbf{W_r}$

čia W_i imties duomenų, patekusių į intervalą $(a_{i-1}, a_i]$, santykinis dažnis $i=1,2,\dots,r$, tenkinantis sąlygą $\sum_{i=1}^r W_i = 1$.

Intervalinių dažnių skirstinį galima pavaizduoti histograma. 1.2 pav. pateiktas intervalinių dažnių skirstinys rodo, kad 217 respondentų išlaidos telekomunikacijų paslaugoms kovo mėnesį yra didesnės už 80 ir neviršija 100 litų.



1.2. pav. Intervalinių dažnių skirstinys pateiktas dažnių histograma (SPSS)

Įvairiose ataskaitose statistiniai duomenys vaizduojami dažnių, santykinų dažnių arba santykinų dažnių išreikštų procentais $W_i \cdot 100$, $i=1,2,\dots,r$ histogramomis, kurios braižomos SPSS paketu keičiant ordinačių ašies (Y ašies) mastelio daugiklio (**Scaling factor**) reikšmę. Jeigu jo reikšmė lygi:

1 - nubraižoma dažnių histograma;

$\frac{n}{100}$ - nubraižoma procentinių dažnių histograma;

n - nubraižoma santykinų dažnių histograma.

1.2. Empirinė skirstinio funkcija

Empirinė skirstinio funkcija vadinama funkcija

$$\hat{F}_n(x) = \frac{n_x}{n},$$

čia n_x - imties reikšmių x_i nedidesnių už x skaičius; n -imties tūris. Didelių imčių empirinė skirstinio funkcija mažai skiriasi nuo tikrosios (populiacijos) funkcijos $F(x)$, t.y. kai $n \rightarrow \infty$, tai $\hat{F}_n(x) \rightarrow F(x)$.

Empirinės skirstinio funkcijos reikšmė taške a apytiksliai lygi tikimybei, kad atsitiktinis dydis X įgys reikšmę nedidesnę už a , t.y. $\hat{F}_n(a) \approx P(X \leq a)$.

Diskrečiuoju atveju iš imties visuomet galima užrašyti empirinį skirstinį

x_i	x_1	x_2	...	x_r
W_i	W_1	W_2	...	W_r

taip, kad būtų $x_1 < \dots < x_r$ - čia $W_i = \frac{m_i}{n}$, m_i yra reikšmės x_i dažnis ($i=1, \dots, r$). Empirinės skirstinio funkcijos išraiška:

$$\hat{F}_n(x) = \begin{cases} 0, & \text{kai } x < x_1; \\ W_1, & \text{kai } x_1 \leq x < x_2; \\ W_1 + W_2, & \text{kai } x_2 \leq x < x_3; \\ \dots & \dots \\ W_1 + \dots + W_{r-1}, & \text{kai } x_{r-1} \leq x < x_r; \\ 1, & \text{kai } x \geq x_r. \end{cases}$$

Pavyzdžiui, diskrečiojo atsitiktinio dydžio X empiriniam skirstiniui

x_i	1	2	3	4
W_i	0,3	0,4	0,2	0,1

užrašysime empirinę skirstinio funkciją

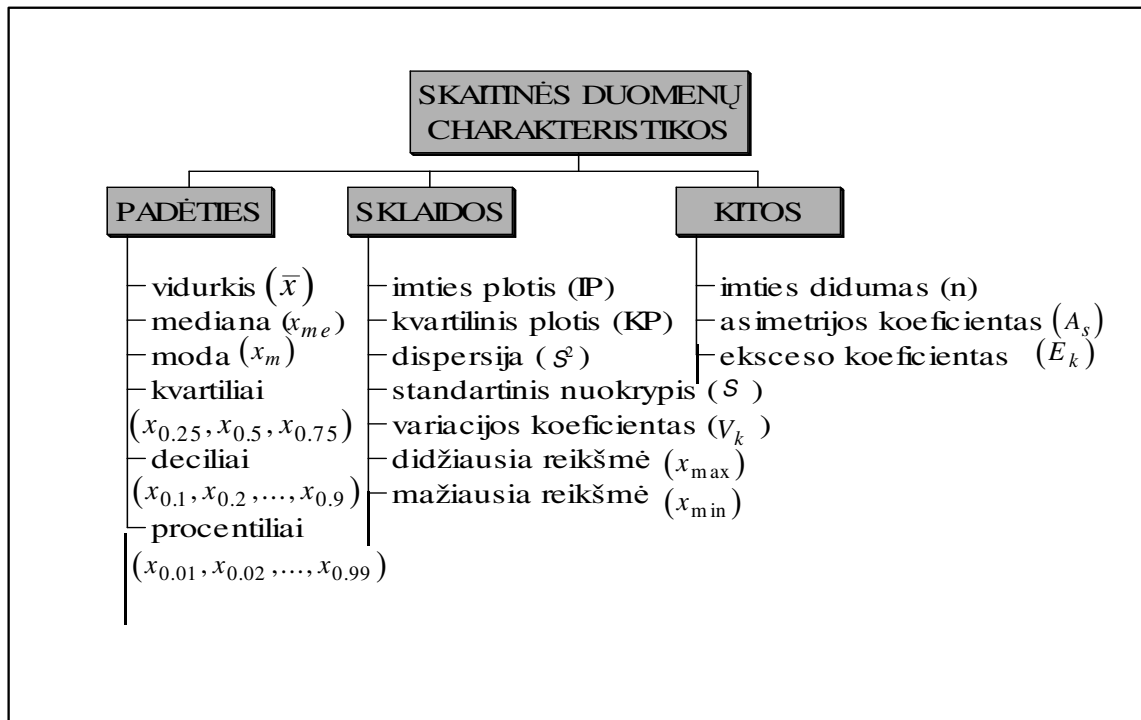
$$\hat{F}_n(x) = \begin{cases} 0, & , \text{kai } x < 1; \\ 0,3 & , \text{kai } 1 \leq x < 2; \\ 0,7 & , \text{kai } 2 \leq x < 3; \\ 0,9 & , \text{kai } 3 \leq x < 4; \\ 1 & , \text{kai } x \geq 4. \end{cases}$$

Žinodami, kad empirinės pasiskirstymo funkcijos reikšmė taške 2 lygi 0,7 ($\hat{F}_n(2) = 0,7$), galime sakyti, kad tikimybė atsitiktiniam dydžiui X įgyti reikšmę nedidesnę už 2 apytiksliai lygi 0,7 ($\hat{F}_n(2) \approx P(X \leq 2)$).

1.3. Skaitinės duomenų charakteristikos

Skaitinės duomenų charakteristikos (1.3 pav.) naudojamos kai norima glaustai apibūdinti ir palyginti tarpusavyje dideles surinktų duomenų aibes. Charakteristikos skirstomos į dvi pagrindines grupes. **Padėties charakteristikos** – apibūdina duomenų reikšmių didumą. Svarbiausios padėties charakteristikos yra vidurkis, mediana, moda ir kvartilai. **Sklaidos charakteristikos** – apibūdina duomenų reikšmių išsisklaidymą. Svarbiausios sklaidos charakteristikos yra dispersija, standartinis nuokrypis, kvartilinis plotis ir variacijos koeficientas. Be charakteristikų priklausančiu šioms

dviems pagrindinėms grupėms dar yra ir kitų (asimetrijos koeficientas, eksceso koeficientas ir t.t.), kurios naudojamos duomenų glaustam apibūdinimui.



1.3 pav. Skaitinės duomenų charakteristikos

SPSS aprašomajai statistikai skirti šie meniu:

Analyze → Descriptive Statistics
 → Frequencies...
 → Descriptives...
 → Explore ...

Graphs →
 → Bar ...
 → Pie ...
 → Histogram ...
 → Boxplot...

STATA aprašomajai statistikai skirti šie meniu:

Statistics → Summaries, tables, and tests
 → Summary and descriptive statistics...
 → Tables...

Graphs →
 → Bar chart
 → Pie chart
 → Histogram ...
 → Box plot...

1.3.1. Padėties charakteristikos

Svarbiausios padėties charakteristikos yra vidurkis, mediana, moda ir kvartilai. Paaiškinsime kaip apskaičiuojamos empirinės skaitinės charakteristikos. Tarkime, kad stebdami atsitiktinį dydį X n

kartų, gavome imtį $x = (x_1, x_2, \dots, x_n)$. Surašydami imties elementus didėjimo tvarka gausime variacinę eilutę $x^* = (x_1^*, x_2^*, \dots, x_n^*)$.

Imties (*empirinis*) **vidurkis (mean, average)** apskaičiuojamas

$$\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i,$$

čia n – **imties didumas (sample size, count)**. Imties vidurkis apibūdina imties vidutinę reikšmę.

Imties (empirine) moda (mode) vadinama tokia atsitiktinio dydžio X reikšmė, kurios įgijimo tikimybė diskrečiuoju atveju yra maksimali, o tolydžiuoju atveju tame taške tankio funkcijos reikšmė yra maksimali. Imties modą žymėsime x_{mo} .

Imties (empirinis) q-kvantilis dalija variacinę eilutę į $q \cdot 100$ ir $(1-q) \cdot 100$ procentinių dalių ($0 < q < 1$), t.y. $q \cdot 100$ procentai imties reikšmių yra ne didesnės už kvantilį ir $(1-q) \cdot 100$ procentai – ne mažesnės už kvantilį x_q .

Pavyzdžiui, kintamojo **išlaidos** 0,45-kvantilis $x_{0,45} = 56$ Lt. Išvada: 45% respondentų išlaidos telekomunikacijų paslaugoms yra ne didesnės už 56 Lt, o 55% - ne mažesnės už 56 Lt. Kai $p = 0.5$, turime imties (empirinę) **medianą (median)**. Ją žymėsime x_{me} . Empirinė mediana yra dažnai naudojama skaitinė duomenų charakteristika. Norėdami ją rasti, pirma turime išrikiuoti skaičius pagal dydį. Kitaip tariant, mes turime perrašyti imtį didėjimo tvarka iš kairės į dešinę. Po to imame skaičių, esantį rikiuotės “viduryje” (ar du tokius skaičius). Galimi du atvejai:

- skaičius n nelyginis. Vidurio skaičiaus numeris yra $(n+1)/2$, todėl tas skaičius ir yra mediana;
- skaičius n lyginis, turime du vidurinius skaičius, kurių numeriai yra $n/2$ ir $(n/2)+1$, mediana yra tų dviejų skaičių vidurkis.

Rasti medianą nėra sunku. Pavyzdžiui, jei duota imtis

$$x = (1, 3, 2, 2, 5, 1, 2, 1, 3, 2), \quad n = 10,$$

tai jos variacinė seka bus

$$1, 1, 1, 2, 2, 2, 2, 3, 3, 4,$$

o mediana bus lygi 2, t.y. 50% imties reikšmių yra mažesnės arba lygios už 2 ir 50% - didesnės arba lygios 2.

Gana dažna klaida yra medianos (vidurio) ir vidurkio painiojimas. Šios dvi sąvokos giminingos, todėl net ir tie, kurie gerai jas skiria, dažnai klysta ir mano, kad mediana ir vidurkis turėtų būti artimi vienas kitam. Nors iš tikrųjų taip dažnai ir būna, bet bendru atveju tai nėra teisinga. Imkime, pavyzdžiui, skaičius 3, 3, 3 ir 91. Šių skaičių mediana yra 3, o vidurkis yra daug didesnis ir lygus 25. Kita vertus, skaičių 3, 3, 100, 102 ir 102 mediana (100) yra daug didesnė už jų vidurkį (62). Taigi klaidinga būtų manyti, jog mediana ir vidurkis yra „maždaug tas pats“. Mediana perskiria surūšiuotus duomenis į dvi dalis: *apatinę pusę* ir *viršutinę pusę*. Išsiaiškinę skaičių aibės “apatinės pusės” ir “viršutinės pusės” prasmę, galime apibrėžti kvartilius.

Pirmasis kvartilis (apatinis) yra apatinės pusės mediana; **trečiasis kvartilis** (viršutinis) yra viršutinės pusės mediana, t.y., kai $p = 0,25$ ir $p = 0,75$, turime **empirinius apatinį ir viršutinį kvartilius (lower quartile – $x_{0,25}$ ir upper quartile – $x_{0,75}$)**. Mediana padalija duomenis į dvi puses, o kvartilių paskirtis – dalyti duomenis į ketvirčius. Pavyzdžiui, jei apatinis kvartilis yra 10,2 tai 25% imties reikšmių yra ne didesnės už 10,2 ir 75% ne mažesnės už 10,2. Analogiškai, jei viršutinis kvartilis yra 100, tai 75% reikšmių yra ne didesnės už 100 ir 25% ne mažesnės už 100. Pati mediana yra antrasis kvartilis, todėl apie 50% skaičių yra ne didesniu už medianą. Nesunku suvokti, kodėl

pirmasis kvartilis, mediana ir trečiasis kvartilis yra kartais atitinkamai vadinami 25-uju, 50-uju ir 75-uju procentiliu. Kvantiliai $x_{0,01}, x_{0,02}, \dots, x_{0,99}$ vadinami **empiriniais procentiliais** (percentiles).

1.3.2. Sklaidos charakteristikos

Tarkime, kad stebėdami atsitiktinį dydį X n kartų, gavome imtį $x = (x_1, x_2, \dots, x_n)$.

Apibūdinant duomenis skaičiais, svarbu nustatyti, kaip duomenų reikšmės yra pasklidusios. Nagrinėsime dvi imtis:

pirma imtis = (45,46,47,48,49,51,52,53,54,55),

antra imtis = (1,11,21,31,41,59,69,79,89,99).

Siūlome patikrinti, kad abiejų imčių empiriniai vidurkiai lygūs 50, o medianos taip pat lygios 50. Kita vertus, akivaizdu, kad šios imtys labai skiriasi, ir antros imties reikšmės yra daug plačiau pasklidusios už pirmos imties reikšmes. Kaip apibūdinti sklaidos dydį? Paprasčiausia būtų imti skirtumą tarp didžiausios ir mažiausios imties reikšmių x_{min} ir x_{max} . Skirtumas $IP = x_{min} - x_{max}$ vadinamas **imties pločiu (range)**. Pirmos imties plotis yra 10, o antros imties – 98. Kaip sklaidos charakteristika, plotis yra naudingas tik tada, kai nėra išskirčių, kadangi jos gali visiškai iškreipti pločio prasmę.

Išskirčių sukeltam iškrypimui eliminuoti įprasta naudoti **kvartilinį plotį (KP)**. Kvartilinis plotis lygus viršutinio (trečiojo) ir apatinio (pirmojo) kvartilių skirtumui $KP = x_{0,75} - x_{0,25}$, ir jis **apibūdina vidurinių 50% duomenų reikšmių sklaidą**. Kvartilinis plotis dažniausiai yra patikima sklaidos charakteristika.

Imties (empirinė) dispersija (variance):

$$s^2 = \mu_2 = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2 = \overline{x^2} - \bar{x}^2,$$

apibūdina atsitiktinio dydžio X reikšmių išsisklaidymo apie empirinį vidurkį laipsnį. Iš empirinės dispersijos apibrėžimo išplaukia, kad ji matuojama kvadratiniais vienetais. Todėl išsisklaidymo laipsniui apibūdinti dažnai vartojama kita charakteristika $s = \sqrt{s^2}$, kuri vadinama **imties standartiniu nuokrypiu (standard deviation)**.

Atsitiktinio dydžio reikšmių išsisklaidymo laipsnį apibūdina ir **kitimo (variacijos) koeficientas (coeff. of variation)**

$$V_K = \frac{s}{\bar{x}}.$$

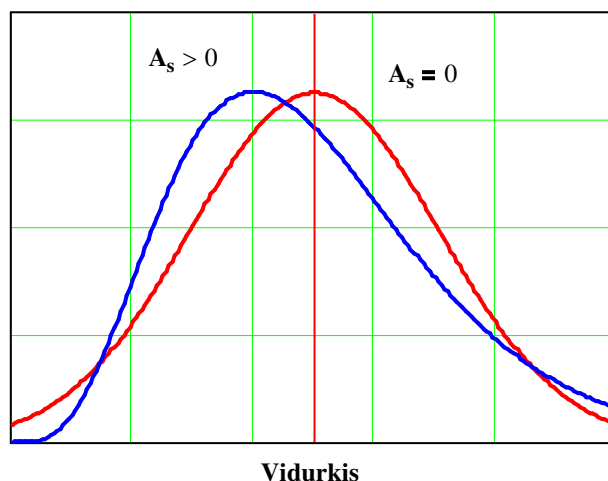
Jis vartojamas lyginant kelis skirtingos eilės atsitiktinius dydžius (pvz., vieno dydžio reikšmės yra dešimčių eilės, o kito tūkstančių eilės) pagal išsisklaidymo apie empirinį vidurkį laipsnį.

Dydis $s_{\bar{x}} = \frac{s}{\sqrt{n}}$ vadinamas **standartine vidurkio įverčio paklaida (standart error of mean)**.

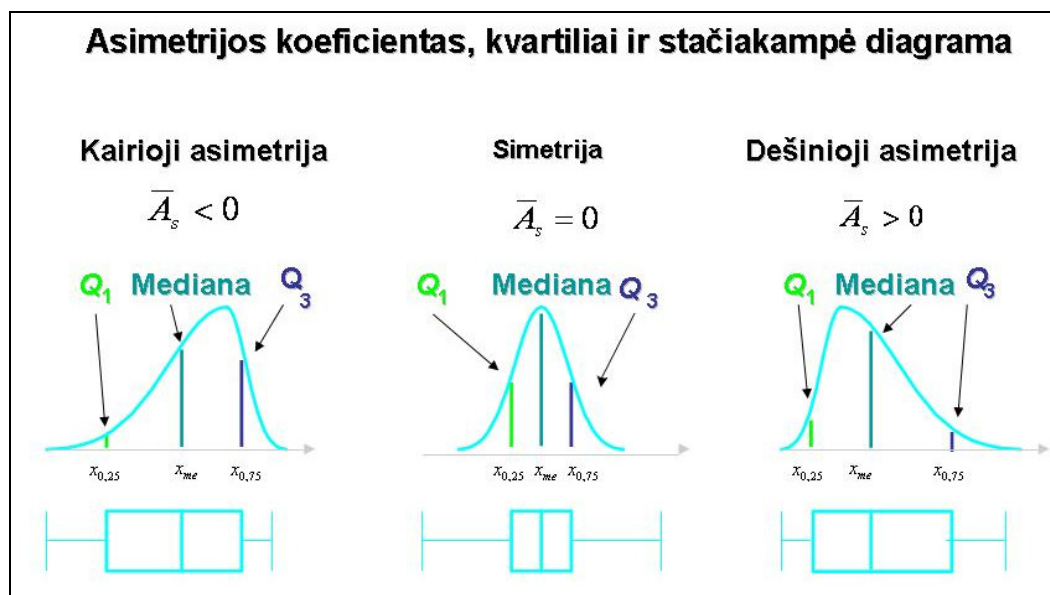
Kuo didesnis imties didumas n , tuo $\bar{x}$ įverčio paklaida mažesnė.

1.3.3. Asimetrijos ir eksceso koeficientai

Imties (empirinis) asimetrijos koeficientu (skewness) $\bar{A}_s$ apibūdina empirinio skirstinio asimetriškumą. Jeigu skirstinys simetriškas vidurkio atžvilgiu, tai $\bar{A}_s = 0$. **Dešinėsios** asimetrijos atveju $\bar{A}_s > 0$, **kairiosios** - $\bar{A}_s < 0$. **Normaliojo skirstino** atveju $\bar{A}_s = 0$.



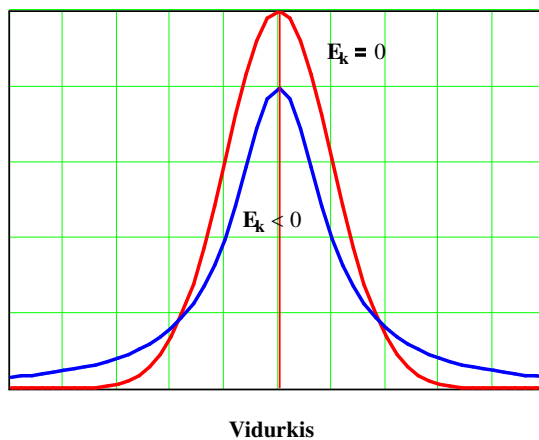
1.4 pav. Skirstinio tankio funkcija, kai asimetrijos koeficientas yra teigiamas ir lygus nuliui



1.5 pav. Skirstinio tankio funkcijos, stačiakampės diagramos ir kvantilių išsidėstymas prie skirtingų asimetrijos koeficiento reikšmių.

Imties (empirinis) eksceso koeficientu (kurtosis) apibūdina empirinio skirstinio tankio (histogramos) **smailumą** ($\bar{E}_k > 0$) arba **lėkštumą** ($\bar{E}_k < 0$) lyginant su normaliuoju skirstiniu. Normaliojo skirstinio atveju $\bar{E}_k = 0$. Jeigu empirinio skirstinio tankio funkcijos maksimumas yra

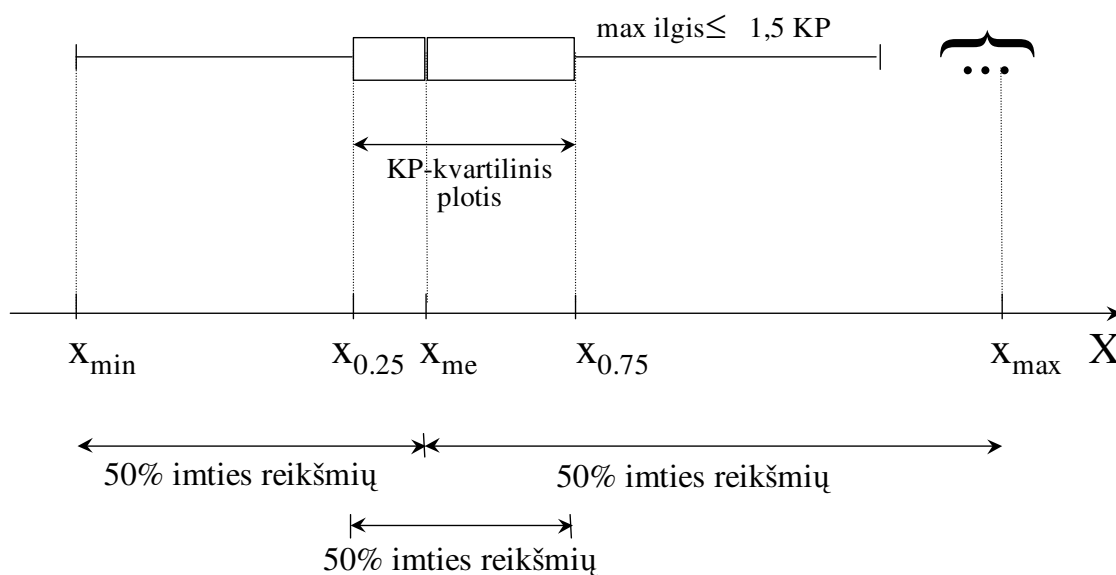
aukščiau (žemiau) nei normaliojo dėsnio, tai sakoma, kad empirinis skirstinys turi teigiamą (neigiamą) ekscesą.



1.6 pav. Skirstinio tankio funkcijos, kai asimetrijos koeficientas yra neigiamas ir lygus nuliui

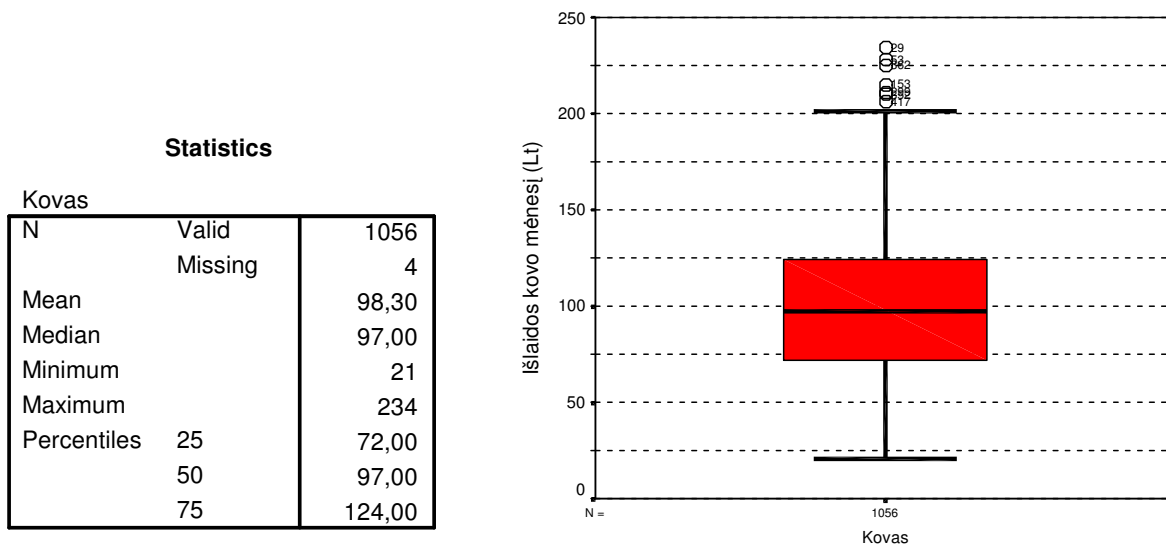
1.4. Stačiakampė diagrama

Šia diagramą pasiūlė amerikiečių mokslininkas Džonas Tjukis. Ji vaizduoja duomenų skaitines charakteristikas. Dabar yra naudojami keli šios diagramos variantai, kurie skiriasi vienas nuo kito tik vaizduojamų skaitinių charakteristikų skaičiumi. 1.7 paveiksle pavaizduota tipiška stačiakampė diagrama. Ją sudaro stačiakampis, einantis nuo apatinio (pirmojo) kvartilio $x_{0,25}$ iki viršutinio (trečiojo) kvartilio $x_{0,75}$. Linija stačiakampio viduje atitinka medianą x_{me} . Nuo stačiakampio šonų išeina „ūšai“, besitęsiantys link mažiausios reikšmės x_{min} ir didžiausios reikšmės x_{max} . Maksimalus ūso ilgis lygus 1,5 kvartilinio pločio. Reikšmės kurių nedengia „ūšai“ vadinamos išskirtimis. **Išskirtis** -tai duomenų reikšmė, išsiskirianti iš bendros visumos, t.y. reikšmė, kuri yra nenatūraliai didesnė arba mažesnė už kitas duomenų reikšmes.



1.6 pav. Stačiakampė diagrama

1.8 pav. pateikta stačiakampė diagrama nubraižyta pagal atsakymų į klausimą „Kiek išleidote pinigų telekomunikacijų paslaugoms kovo mėnesį?“ Iš 1056 apklaustų respondentų, 4 neatsakė į pateiktą klausimą. Kovo mėnesį vidutinės respondentų išlaidos telekomunikacijų paslaugoms buvo 98,30 Lt., 25% respondentų telekomunikacijų paslaugoms išleido ne daugiau kaip 72 litus, 20% respondentų telekomunikacijų paslaugoms išleido ne daugiau kaip 97 litus ir t.t.



1.8 pav. Išlaidų telekomunikacijų paslaugoms kovo mėn. stačiakampė diagrama.

1.5. Diagrama STIEBAS ir LAPAS

Amerikiečių mokslininkas **Džonas Tjukis** (John W. Tukey) sukūrė tiriamąją duomenų analizę (Exploratory Data Analysis, sutrumpintai EDA) pasiūlydamas daug naujų ir vaizdžių duomenų grafinio vaizdavimo būdų. **Diagrama „Stiebas ir lapas“ (Steam and Leaf)** - vienas iš jų. Ji naudojama, kai stebimas atsitiktinis dydis yra tolydusis arba diskretusis įgyjantis daug skirtingų reikšmių ir kai norima atvaizduoti ne tik intervalinius dažnius, bet ir reikšmių pasiskirstymą intervalų viduje.

Nubraižysime JAV universitetų pilno profesoriaus metinės algos duomenų diagramą Stiebas ir Lapas. Kad būtų lengviau suprasti ką vaizduoja diagrama, sudarykime pilno profesoriaus metinės algos reikšmių variacinę seką.

```

55,90  56,10  56,70  61,00  62,40  63,00  63,50  64,90  65,30  65,50
65,70  66,50  66,70  67,10  67,10  67,10  68,80  68,80  69,50  70,20
70,20  70,30  70,96  71,40  71,60  72,40  72,90  73,00  73,10  75,00
75,40  75,50  76,80  76,90  77,90  78,90  79,30  79,80  82,60  83,00
83,40  85,10  86,90  87,00  90,20  90,50  91,20  92,70  93,30  96,50

```

Žemiau pateikta diagrama „Stiebas ir lapas gauta naudojant paketą SPSS.

```

Stem-and-Leaf Plot
Frequency    Stem & Leaf
(Dažnis)      (Stiebas) (Lapas)

      3,00      5 .  566
      5,00      6 . 12334
     11,00      6 . 55566777889
     10,00      7 . 0000112233
      9,00      7 . 555667899
      3,00      8 . 233

```

3,00	8 . 567
5,00	9 . 00123
1,00	9 . 6

Stem width: 10,00
Each leaf: 1 case(s)

Diagramą sudaro trys stulpeliai. Pirmame stulpelyje vaizduojami dažniai, antrame - stiebo reikšmės, trečiame - lapų reikšmės. Kiekvienas lapas gali vaizduoti vieną arba kelis stebinius. Diagrama turi du parametrus:

- stiebo plotis;
- vienu diagramos lapu vaizduojamų stebinių skaičius.

Priklausomai nuo stebimų reikšmių didumo, diagramoje gali būti vaizduojami du, trys ar daugiau kiekvieno stebinio reikšmės vyriausiųjų skilčių skaitmenys. Visi skaitmenys vaizduojami stiebe išskyrus paskutinį, kuris vaizduojamas lape. Stiebas vaizduoja sveikąją skaičiaus dalį, o lapas dešimtasias skaičiaus dalis. Vaizduojamo stebinio suapvalinta reikšmė gaunama padauginus stiebe ir lape vaizduojamą skaičių iš stiebo pločio.

Pateiktoje diagramoje stiebo plotis yra 10 ir kiekvienas lapas vaizduoja vieną stebinį. Pavyzdžiui, pirmoje diagramos eilutėje pavaizduoti trys stebiniai. Kiekviena stebinio reikšmė vaizduojama dviem skaitmenimis. Pirmas skaitmuo (skaičiaus sveikoji dalis) pateikta stiebe, o antrasis skaitmuo (skaičiaus dešimtosios dalys) - lape. Dviejų skaitmenų skaičių vaizduojamą diagramoje padauginę iš stiebo pločio gauname suapvalintą stebinio reikšmę. Pirmoje eilutėje pavaizduotų stebinių reikšmės:

$5,5 \cdot 10 = \mathbf{55}$ (tikroji reikšmė 55,90) ;
 $5,6 \cdot 10 = \mathbf{56}$ (tikroji reikšmė 56,10);
 $5,6 \cdot 10 = \mathbf{56}$ (tikroji reikšmė 56,70).

tik tam tikra jo charakteristika. Statistinėms išvadoms naudojama kokia nors atsitiktinės imties funkcija $f(X_1, X_2, \dots, X_n)$, kuri vadinama statistika. Pavyzdžiui, imties vidurkis $\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i$.

Terminas **statistika** yra **tradicinis**. Jis turi **dvi** prasmes: statistika-mokslas ir statistika – atsitiktinės imties funkcija. Kadangi **statistika** yra atsitiktinių dydžių funkcija, todėl ji taip pat yra atsitiktinis dydis. Taigi galima kalbėti apie statistikos, arba vadinamąjį statistikos imties, skirstinį (2.2 pav.). Statistikos (atsitiktinio dydžio), realizaciją žymėsime mažosiomis raidėmis, o statistiką – didžiosiomis. **Statistikos** yra **atsitiktiniai dydžiai**, todėl **galima apibrėžti jų vidurkius, dispersijas bei kitas charakteristikas**.

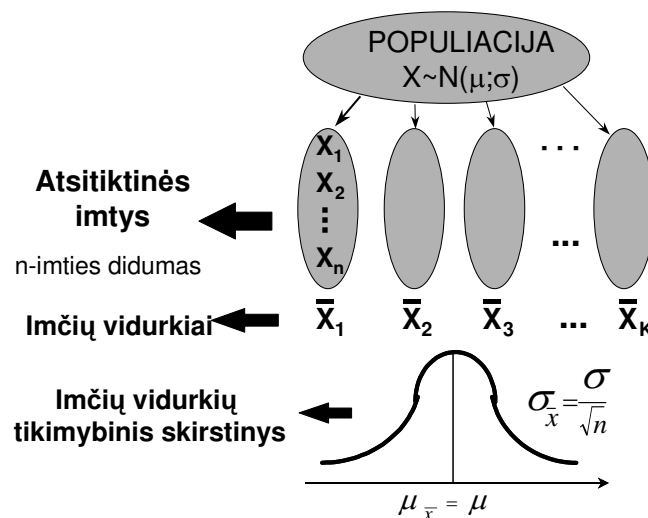
Pavyzdžiui,

Atsitiktinė imtis $X = (X_1, \dots, X_n)$.

Atsitiktinės imties realizacija $x = (x_1, \dots, x_n)$, $x = (1, 3, 1, 3, 2, 2)$.

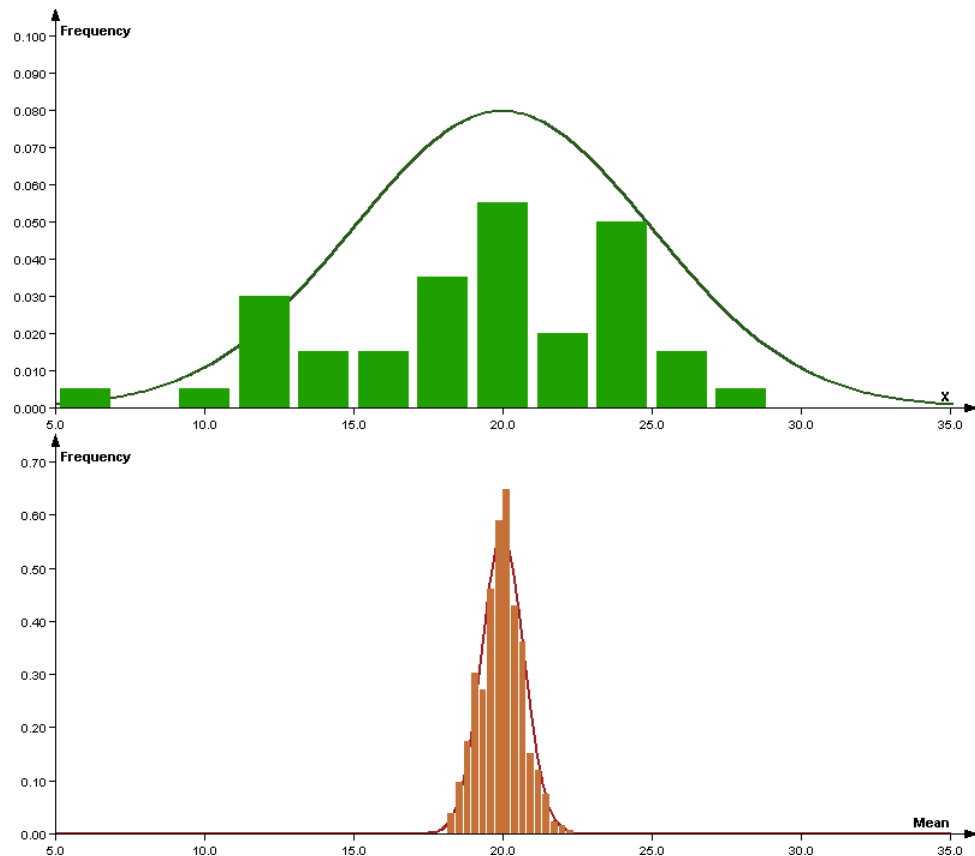
Statistika $\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i$ - imties vidurkis (atsitiktinis dydis).

Statistikos realizacija $\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$, imties vidurkis $\bar{x} = 2$ (skaičius).



2.2 pav. Statistikos (vidurkio) skirstinys

Sprendžiant konkretų uždavinį, statistinės išvados remiasi tik viena (konkrečia) imties realizacija. Kad ir kokia ta realizacija būtų, pasirinktos sprendimo taisyklės turi tikti. Kitaip tariant, pasirinkę tikimybių teorijos metodus, turime įvertinti statistikos atsitiktinumo lygį ir pasirinkti tokią sprendimo procedūrą, kad klaidingos išvados tikimybė būtų maža.



2.3 pav. Vidurkių (statistikos) skirstinio dažnių histograma gauta iš populiacijos $X \sim N(20; 5)$ atrinkus 500 atsitiktinių dydžio 50 imčių.

Standartinė statistikos paklaida. Statistikos standartinis nuokrypis vadinamas standartine statistikos paklaida. Kuo standartinė paklaida mažesnė, tuo statistika daugiau koncentruota apie vidurkį. Vidurkio $\bar{X}$ standartinis nuokrypis vadinamas standartine vidurkio paklaida ir lygi:

$$\sigma_{\bar{X}} = \sigma / \sqrt{n}.$$

Didėjant n , standartinė vidurkio paklaida nyksta. Realiai tikrosios standartinės paklaidos dažniausiai neįmanoma rasti, todėl vietoje jos skaičiuojama, **empirinė standartinė vidurkio paklaida**:

$$S_{\bar{X}} = \frac{S}{\sqrt{n}}.$$

2.2. Parametrų taškiniai įverčiai

Pavyzdžiui, norime įvertinti įmonės akcijų kurso svyravimą, vidutines klientų išlaidas telekomunikacijų paslaugoms per mėnesį, vidutinę vietinio telefoninio pokalbio trukmę ir pan. Trumpai kalbant, remdamiesi imties duomenimis, norime įvertinti populiacijos parametą. Matematiškai tokią situaciją aprašo parametrinis modelis: stebimas atsitiktinis dydis X , kurio skirstinys $F(\Theta)$ priklauso nuo nežinomo parametro θ . Žymima $X \sim F(\Theta)$.

Nežinomo populiacijos parametro Θ taškinio įverčio radimo uždavinys – tai apytikslės formulės $\Theta \approx \mathcal{G}(X_1, \dots, X_n)$ radimo uždavinys. Statistika, kuri naudojama nežinomam parametrai θ įvertinti, vadinama θ **taškiniu įverčiu** ir žymima $\hat{\Theta} = \mathcal{G}(X_1, \dots, X_n)$

Atkreipiame dėmesį, kad:

- nežinomas parametras θ yra **skaičius**;

- parametro **taškinis įvertis** $\hat{\theta}$ yra **atsitiktinis dydis**;
- taškinio **įverčio realizacija** yra skaičius, randamas konkrečiai imties realizacijai.

a) Koks taškinis įvertis yra geras ?

Dažniausiai reikalaujama, kad įvertis būtų:

- ♦ suderintas;
- ♦ nepaslinktas, $E\hat{\theta} = \Theta$;
- ♦ efektyvus, $D\hat{\theta} \rightarrow \min$;

Įvertis yra suderintas, jeigu jo realizacija $\hat{\theta}$ apskaičiuota didelėms imtims, beveik nesiskiria nuo paties parametro θ . Parametro θ įvertis $\hat{\theta}$ vadinamas suderintuoju, jeigu kiekvienam fiksuotam $\varepsilon > 0$

$$P\left(\left|\hat{\theta} - \theta\right| > \varepsilon\right) \rightarrow 0, \quad \text{kai } n \rightarrow \infty.$$

Įverčio nepaslinktumas – viena iš dažniausiai pasitaikančių įverčių savybių. Tam tikra prasme tai reikalavimas, kad įverčio realizacijų nuokrypiai nuo parametro būtų subalansuoti.

Parametro θ įvertis $\hat{\theta}$ vadinamas nepaslinktuoju, jeigu

$$E\hat{\theta} = \theta.$$

Baigtinės populiacijos atveju įverčio $\hat{\theta}$ nepaslinktumas reiškia, kad paėmus visas skirtingas dydžio n imtis ir suskaičiavus statistikos realizacijas $\hat{\theta}$, visų šių realizacijų aritmetinis vidurkis lygus θ .

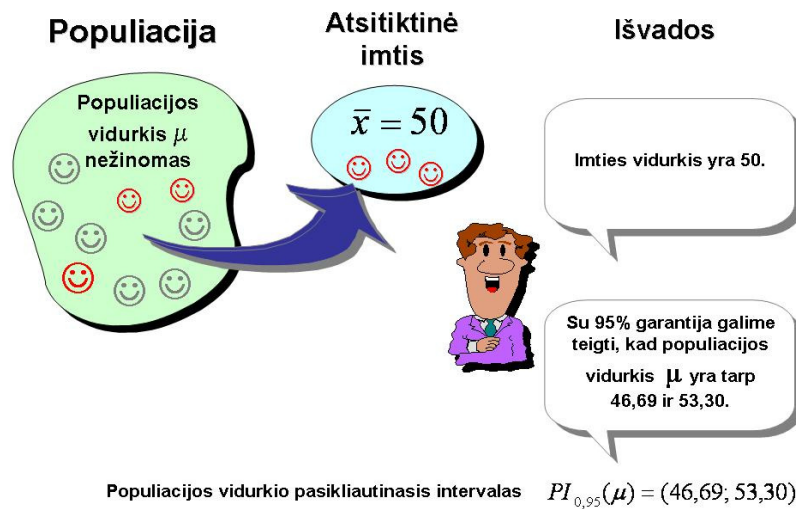
Efektyvusis įvertis. Tarkime, turime kelis n elementų imties nepaslinktuosius parametro θ įverčius. Natūralu geresniu laikyti tą įvertį, kuris daugiau „koncentruotas“ apie θ , t.y. kurio dispersija mažesnė.

Tegul $\hat{\theta}_1$ ir $\hat{\theta}_2$ yra nepaslinktieji θ įverčiai. Tuomet $\hat{\theta}_1$ yra efektyvesnis už $\hat{\theta}_2$, jeigu $D\hat{\theta}_1 < D\hat{\theta}_2$.

2.3. Pasikliautinųjų intervalų skaičiavimas

Kalbėdami apie atsitiktinį dydį mes dažnai norime žinoti intervalą, iš kurio praktiškai visada (su tikimybe artima 1) bus šio atsitiktinio dydžio reikšmės. Parametro Θ taškinis įvertis $\hat{\Theta}$ yra atsitiktinis dydis. Kad ir koks „geras“ būtų įvertis $\hat{\Theta}$ jo reikšmės išsisklaidę apie Θ . Todėl radus taškinį įvertį $\hat{\Theta}$, būtina atsakyti į klausimą, koks šio įverčio tikslumas ir patikimumas?

Sprendžiant praktinius uždavinius svarbu žinoti, kokios galimos paklaidos, jei nežinomą parametą Θ keičiam jo taškiniu įverčiu $\hat{\Theta}$. Kokia tikimybė, kad paklaidos neviršys tam tikrų rėžių? Šie klausimai ypač aktualūs, kai imties tūris yra nedidelis, nes tada taškinis įvertis pasižymi dideliu atsitiktinumu.



2.4 pav. Išvadų apie vidurkį formulavimas

Kadangi $\hat{\Theta}$ yra atsitiktinis dydis, pareikalaujame, kad nelygybė

$$|\Theta - \hat{\Theta}| < \varepsilon$$

galiotų su tam tikra **didele** tikimybe $(1 - \alpha) \in \{0,9; 0,95; 0,95\}$.

Tai trumpai galime užrašyti

$$P(|\Theta - \hat{\Theta}| < \varepsilon) = 1 - \alpha.$$

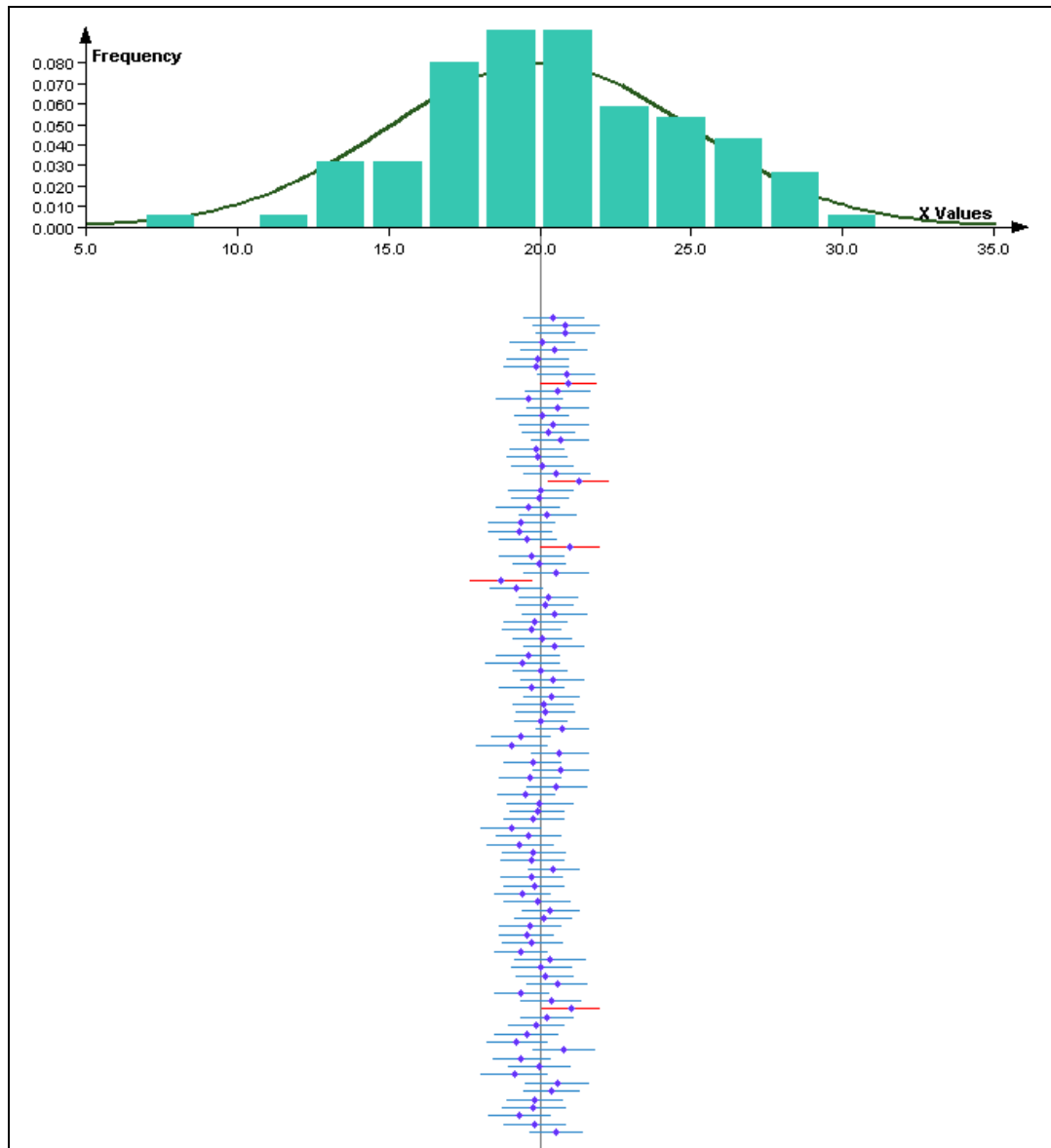
Ši lygybė ekvivalenti lygybei

$$P(\hat{\Theta} - \varepsilon < \Theta < \hat{\Theta} + \varepsilon) = 1 - \alpha.$$

Intervalas $(\hat{\Theta} - \varepsilon; \hat{\Theta} + \varepsilon)$ vadinamas **pasikliautiniu intervalu (Confidence interval)**, tikimybė $1 - \alpha$ **pasiklovimo tikimybe arba pasiklovimo lygmeniu**, skaičius ε - paklaida.

2.4 pav. pateiktas išvadų apie populiacijos ir imties vidurkį formulavimo pavyzdys.

Detaliau paaiškinsime bendrąjį algoritmą pasikliautiniams intervalams rasti. Tarkime, kad atsitiktinio dydžio X pasiskirstymo funkcijos $F(x, \Theta_1, \dots, \Theta_n)$ analizinė išraiška žinoma, tačiau ji priklauso nuo nežinomų parametrų $\Theta_1, \dots, \Theta_n$, kuriuos reikia įvertinti. Pavyzdžiui, normalusis skirstinys turi du parametrus: μ - vidurkis ir σ - standartinis nuokrypis.



2.5 pav. Populiacijos vidurkio 0,95 pasikliautiniai intervalai, gauti iš populiacijos $X \sim N(20; 5)$ atrinkus 500 atsitiktinių dydžio 50 imčių.

Bendras pasikliautinio intervalo sudarymo algoritmas:

1. Maksimalaus tikėtumo, momentų ar kitu metodu randamas stebimo atsitiktinio dydžio X skirstinio nežinomo parametro Θ taškinis įvertis $\hat{\Theta}$.
2. Parenkama statistika $U(\Theta)$, priklausanti nuo parametro Θ , kurios skirstinys žinomas.
 Statistika $U(\Theta)$ parenkama priklausomai nuo :
 - X tikimybių skirstinio tipo;
 - nežinomų parametrų skaičiaus;
 - imties didumo;

Pastaba. Visų žinomų atsitiktinių dydžių tikimybių skirstinių parametrų taškinis įverčius ir pasikliautinių intervalų apskaičiavimui naudojamas statistikas galima rasti statistikos žinynuose ir vadovėliuose. Jūs turite tik nesuklysti pasirenkant !!!

Pavyzdžiui, kai $\bar{X}$ skirstinys yra normalusis $\bar{X} \sim N(\mu, \sigma)$, parametro μ pasikliautinojo intervalo radimui naudojamos šios statistikos:

a) kai σ žinomas

$$Z = \frac{\bar{X} - \mu}{\sigma} \sqrt{n}, \quad Z \sim N(0,1);$$

b) kai σ nežinomas

$$t = \frac{\bar{X} - \mu}{\bar{s}} \sqrt{n}, \quad t \sim St(n-1);$$

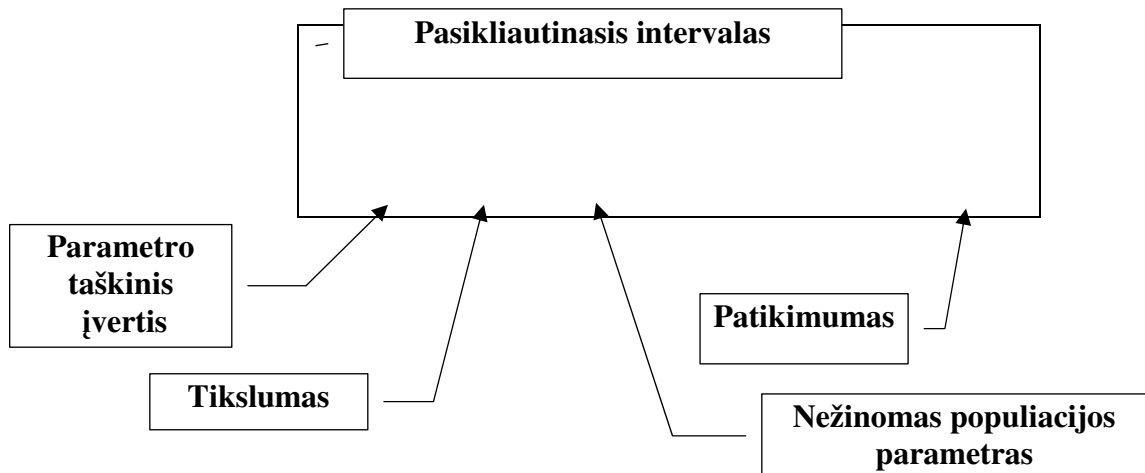
3. Parenkamas pasiklovimo lygmuo

$$1 - \alpha \in \{0,9; 0,95; 0,99\}$$

4. Randami kvantiliai $u_{\alpha/2}$ ir $u_{1-\alpha/2}$ tenkinantys lygtį

$$P(u_{\alpha/2} < U(\Theta) < u_{1-\alpha/2}) = 1 - \alpha \quad (*)$$

5. Lygtis (*) pertvarkoma į jai ekvivalenčią lygtį



Pastaba. Bendru atveju pasikliautinis intervalas gal būti nesimetrinis vidurkio atžvilgiu.

2.4. Normaliojo skirstinio vidurkio pasikliautinojo intervalo radimo pavyzdys

Duota atsitiktinė imtis $x = (x_1; x_2; \dots; x_n)$. Žinoma, kad stebimo atsitiktinio

dydžio $\bar{X}$ skirstinys yra normalusis $\bar{X} \sim N(\mu, \sigma)$. Abu parametrai μ ir σ nežinomi.

Rasti nežinomo parametro μ pasikliautinąjį intervalą.

Sprendimas

1. Geriausias parametro μ taškinis įvertis yra empirinis vidurkis $\bar{x}$ (rastas taikant momentų ir maksimalaus tikėtimumo metodus)

$$\hat{\mu} = \bar{x} = \frac{1}{n} \sum_{i=1}^n x_i.$$

2. Pasikliautinojo intervalo sudarymui pasirenkame statistiką

$$t = \frac{\bar{x} - \mu}{\bar{s}} \sqrt{n}, \quad t \sim St(n-1).$$

3. Pasirenkame pasiklovimo lygmenį $1 - \alpha$.

4. Randame Stjudento skirstinio kvantilius $t_{\alpha/2; n-1}$ ir $t_{1-\alpha/2; n-1}$.

5. Pertvarkę lygtį

$$P\left(t_{\alpha/2; n-1} < \frac{\bar{x} - \mu}{\bar{s}} < t_{1-\alpha/2; n-1}\right) = 1 - \alpha,$$

gauname

$$P\left(\bar{x} - \frac{\bar{s}}{\sqrt{n}} t_{1-\alpha/2; n-1} < \mu < \bar{x} - \frac{\bar{s}}{\sqrt{n}} t_{\alpha/2; n-1}\right) = 1 - \alpha.$$

Tarkime, kad $\bar{x} = 50$, $\bar{s} = 20$.

Panaudoję išvestą formulę apskaičiuosime nežinomo parametro μ pasikliautinuosius intervalus prie skirtingų pasiklovimo lygmenų $1 - \alpha$ ir imčių didumų n . Apskaičiavimo rezultatai pateikti 2.1 lentelėje.

2.1 lentelė

Pasiklovimo lygmuo $1 - \alpha$	$\frac{\alpha}{2}$	Nežinomo parametro μ pasikliautinis intervalas			
		n=100		n=10	
		$t_{\alpha/2; 99}$		$t_{\alpha/2; 9}$	
0.8	0.1	-1.29	50 ± 2.58	-1.38	50 ± 3.90
0.9	0.05	-1.66	50 ± 3.32	-1.83	50 ± 5.18
0.95	0.025	-1.98	50 ± 3.96	-2.26	50 ± 6.39
0.99	0.005	-2.63	50 ± 5.28	-3.25	50 ± 9.19

Pasikliautinųjų intervalų radimo formulės.

Normaliojo skirstinio $X \sim N(\mu, \sigma^2)$ vidurkio pasikliautinis intervalas:

$$PI_{1-\alpha}(\mu) = \left(\bar{X} - \frac{\bar{S}}{\sqrt{n}} \cdot t_{1-\alpha/2; n-1}; \bar{X} - \frac{\bar{S}}{\sqrt{n}} \cdot t_{\alpha/2; n-1} \right).$$

Normaliojo skirstinio $X \sim N(\mu, \sigma^2)$ dispersijos pasikliautinis intervalas

$$PI_{1-\alpha}(\sigma^2) = \left(\frac{\bar{S}^2 \cdot (n-1)}{\chi^2_{1-\frac{\alpha}{2}; n-1}} ; \frac{\bar{S}^2 \cdot (n-1)}{\chi^2_{\frac{\alpha}{2}; n-1}} \right).$$

čia $\bar{S}^2 = \frac{1}{n-1} \cdot \sum_{i=1}^n (X_i - \bar{X})^2$, nepaslinktas populiacijos dispersijos σ^2 taškinis įvertis.

Įvykio tikimybės pasikliautinis intervalas

$$PI_{1-\alpha}(p) = \left(\hat{p} - z_{1-\frac{\alpha}{2}} \sqrt{\frac{\hat{p} \cdot (1-\hat{p})}{n}} ; \hat{p} + z_{\frac{\alpha}{2}} \sqrt{\frac{\hat{p} \cdot (1-\hat{p})}{n}} \right),$$

čia $\hat{p} = \frac{m}{n}$.

Kitų skirstinių parametrų pasikliautinių intervalų radimo formules galima rasti statistikos vadovėliuose ir žinynuose. Pasikliautinajam intervalui sudaryti naudojami trys tarpusavyje susiję dydžiai: pasiklovimo lygmuo, imties didumas n ir tikslumas. Žinoma norisi turėti kuo didesnę pasiklovimo lygmenį, tačiau tuomet labai padidėja pasikliautinojo intervalo ilgis ir sumažėja informacijos tikslumas. Pavyzdžiui, kažin ar kam reikalinga išvada, kad 100% vidutinis žmonių gyvenimo trukmės pasikliautinis intervalas yra (0, 120) metų. Kita vertus, labai mažindami pasikliautinojo intervalo ilgį, kartu sumažiname ir pasiklovimo lygmenį. Lygmenys 0,9; 0,95 ir 0,99 yra kompromisinis sprendimas, leidžiantis išsaugoti pakankamą pasiklovimo lygį, kartu garantuojant ne per didelius intervalų ilgius. Didinant imtį, mažinamas pasikliautinojo intervalo ilgis. Kuo didesnis imties didumas n , tuo trumpesnis pasikliautinis intervalas.

2.5. Normaliojo skirstinio vidurkio pasikliautinojo intervalo radimo pavyzdys

Tarkime, kad duota išlaidų telekomunikacijų paslaugoms kovo mėnesį imtis. Apskaičiuosime klientų išlaidų vidurkio 0,95 pasikliautinąjį intervalą (2.4 pav.).

SPSS meniu: Analyze → Descriptive statistics → Explore.

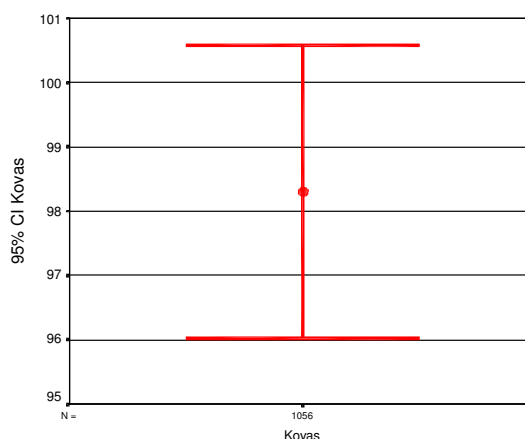
Gavome, kad populiacijos vidurkio μ pasikliautinis intervalas yra $PI_{0,95}(\mu) = (96,02 ; 100,58)$.

Išvada: su 95% garantija galime teigti, kad klientų vidutinės išlaidos telekomunikacijų paslaugoms kovo mėnesį buvo intervale nuo 96,02 iki 100,58 Lt.

Descriptives			Statistic	Std. Error
Kovas	Mean		98,30	1,16
	95% Confidence Interval for Mean	Lower Bound	96,02	
		Upper Bound	100,58	
	5% Trimmed Mean		97,28	
	Median		97,00	
	Variance		1426,642	
	Std. Deviation		37,77	
	Minimum		21	
	Maximum		234	
	Range		213	
	Interquartile Range		52,00	
	Skewness		,363	,075
	Kurtosis		,053	,150

2.6 pav. Skaitinės charakteristikos ir vidurkio pasikliautinis intervalas (SPSS)

Nubraižysime klientų išlaidų kovo mėnesį vidurkio 0,95 pasikliautinio intervalo grafiką (1.11 pav). **SPSS meniu: Graphs→ Error bar (Simple, Summary of separate variables)**

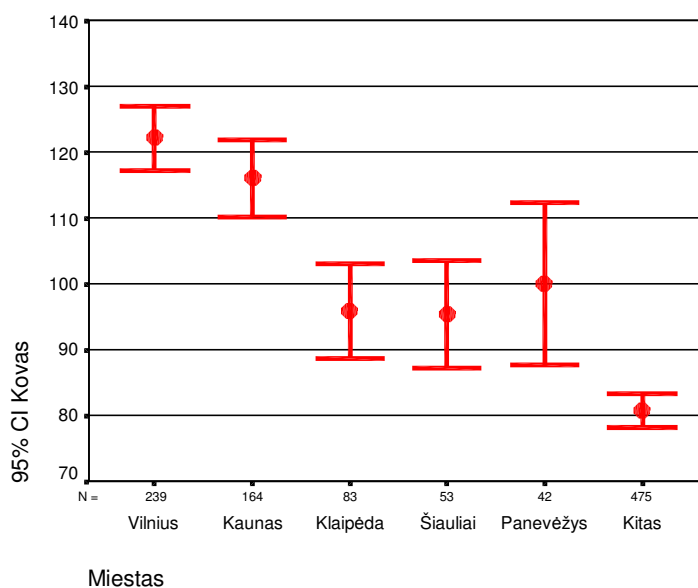


2.7 pav. Klientų išlaidų kovo mėnesį vidurkio 0,95 pasikliautinio intervalo grafikas

Nubraižysime klientų išlaidų kovo mėnesį vidurkio 0,95 pasikliautinio intervalo grafikus kiekvienai kintamojo **miestas** kategorijai (2.7 pav.).

SPSS meniu: Graphs→ Error bar (**Simple, Summaries for groups of cases**)

Išvada. Įmonės klientų vilniečių vidutinės išlaidos telekomunikacijų paslaugoms kovo mėnesį buvo didžiausios ir statistiškai reikšmingai skyrėsi nuo kitų miestų klientų, išskyrus Kauno miesto klientus. Įmonės klientų vilniečių ir kauniečių vidutinės išlaidos telekomunikacijų paslaugoms kovo mėnesį buvo didesnės už klaipėdiečių ir šiauliečių (skyrėsi statistiškai reikšmingai) ir t.t.



2.8 pav. Klientų išlaidų kovo mėnesį vidurkio 0,95 pasikliautinojo intervalo grafikai kiekvienai kintamojo **miestas** kategorijai

2.6. Įvykio tikimybės pasikliautinojo intervalo radimo pavyzdys

Apklausus 100 atsitiktinai atrinktų įmonės darbuotojų, paaiškėjo, kad prieš įmonėje vykdomas reformas pasisakė 35 darbuotojai. Raskite darbuotojų, nepritariančių įmonėje vykdomoms reformoms, dalies 90% pasikliautinąjį intervalą.

Tarkime, kad X yra darbuotojo atsakymas į klausimą. Tegul X įgyja 1 (jeigu darbuotojas nepritaria įmonėje vykdomoms reformoms) arba 0 (priešingu atveju). Tuomet X skirstinys yra binominis $X \sim B(1; p)$, čia p -dalis įmonės darbuotojų nepritariančių reformoms. Pasikliautinąjį intervalą apskaičiuosime pagal formulę

$$PI_{1-\alpha}(p) = \left(\hat{p} - z_{1-\frac{\alpha}{2}} \sqrt{\frac{\hat{p} \cdot (1-\hat{p})}{n}} ; \hat{p} + z_{\frac{\alpha}{2}} \sqrt{\frac{\hat{p} \cdot (1-\hat{p})}{n}} \right).$$

Mūsų atveju $\hat{p} = \frac{35}{100}$, $1-\alpha=0,9$, $\alpha=0,1$, $z_{1-\frac{\alpha}{2}} = -z_{\frac{\alpha}{2}} = 1,64$. Įstatę reikšmes į formulę,

gauname

$$PI_{0,90}(p) = (0,24 ; 0,46)$$

Išvada: su 90% garantija galime teigti, kad įmonėje vykdomoms reformoms nepritaria nuo 24 iki 46 procentų dirbančiųjų.

3. HIPOTEZIŲ TIKRINIMAS

3.1. Pagrindinės sąvokos

Dažnai atliekamų tyrimų tikslas – ne tik įvertinti tiriamųjų objektų skaitines charakteristikas, bet ir palyginti: įvairias metodikas pagal jų efektyvumą; įvairių žmonių grupių nuostatas, politinių partijų reitingus, paslaugų vartotojų grupių charakteristikas arba jų poreikius ir t.t. Matematinėje kalboje tokie palyginimo uždaviniai formuluojami kaip hipotezių tikrinimo uždaviniai.

Hipoteze statistikoje vadinamas teiginys apie nežinomus populiacijų požymių (kintamųjų) skirstinius. Pavyzdžiui, statistinėmis hipotezėmis bus šie teiginiai:

- stebimo atsitiktinio dydžio skirstinys yra normalusis;
- atsitiktinio dydžio vidurkis lygus 100;
- dviejų atsitiktinių dydžių vidurkiai yra lygūs;
- vieno atsitiktinio dydžio dispersija yra didesnė negu kito ir t.t.

Tikrinamoji hipotezė vadinama **nuline** ir žymima H_0 . Kartu nagrinėjama ir jai priešinga hipotezė H_a . Ji vadinama **alternatyviaja**. Hipotezės skirstomos į parametrines ir neparametrines. Jei statistinė hipotezė tikrinama nežinomų pasiskirstymo dėsnio parametrų atžvilgiu, tai ji vadinama **parametrine hipoteze** (populiacijos parametras lyginamas su koku nors skaičiumi, arba tarpusavyje lyginami kelių populiacijų analogiški parametrai). Pavyzdžiui, jei **vieno** stebimo atsitiktinio dydžio skirstinys normalusis $X \sim N(\mu, \sigma^2)$ galima tikrinti šias hipotezes:

$$1) \quad H_0 : \mu = \mu_0, \quad H_a : \mu \neq \mu_0; (H_a : \mu < \mu_0; H_a : \mu > \mu_0)$$

$$2) \quad H_0 : \sigma^2 = \sigma_0^2, \quad H_a : \sigma^2 \neq \sigma_0^2; (H_a : \sigma^2 < \sigma_0^2; H_a : \sigma^2 > \sigma_0^2)$$

$$3) \quad \begin{array}{ll} H_0 : \mu = \mu_0 & H_a : \mu \neq \mu_0 \\ \sigma^2 = \sigma_0^2 & \sigma^2 \neq \sigma_0^2 \end{array};$$

Tarkime, kad **dviejų** stebimų atsitiktinių dydžių skirstiniai yra normalieji $X_1 \sim N(\mu_1, \sigma_1^2)$ ir $X_2 \sim N(\mu_2, \sigma_2^2)$. Lyginant jų parametrus, dažnai tenka tikrinti hipotezes:

$$1) \quad H_0 : \mu_1 = \mu_2, \quad H_a : \mu_1 \neq \mu_2; (H_a : \mu_1 < \mu_2; H_a : \mu_1 > \mu_2),$$

$$2) \quad H_0 : \sigma_1^2 = \sigma_2^2, \quad H_a : \sigma_1^2 \neq \sigma_2^2; (H_a : \sigma_1^2 < \sigma_2^2; H_a : \sigma_1^2 > \sigma_2^2).$$

Parametrinė hipotezė vadinama **paprastąja**, jeigu kokio nors parametro atžvilgiu iškelta tik viena prielaida, pavyzdžiui, $H_0 : \Theta = \Theta_0$; priešingu atveju ji vadinama **sudėtingąja**, pavyzdžiui, $H_a : \Theta < \Theta_0$.

Pateiksime keletą hipotezių pavyzdžių:

1 pavyzdys. Atsitiktinai atrinkus 1000 įmonės klientų apklausta, kiek vidutiniškai minučių per savaitę jie kalba telefonu su savi šeimos nariais. Ar turint šios apklausos duomenis, galima teigti, kad klientai kalba su savo šeimos nariais daugiau kaip 20 min. Tiriamos populiacijos pokalbių su savo šeimos nariais per savaitę trukmės vidurkį pažymėjus μ , uždavinį galime suformuluoti taip:

$$\begin{cases} H_0 : \mu = 20, \\ H_a : \mu > 20. \end{cases}$$

2 pavyzdys. Firmos vadovai nori sužinoti ar mokymo programa padidina pardavimų apimtis. Atsitiktinai atrinkti 100 pardavėjų ir surinkti duomenys apie kiekvieno pardavėjo per savaitę parduotų paslaugų apimtį (litas) prieš mokymą ir po. Firmos pardavėjų per savaitę parduodamų prekių apimtį vidurkį prieš mokymą ir po pažymėję atitinkamai μ_1 ir μ_2 , uždavinį galime užrašyti taip:

$$\begin{cases} H_0 : \mu_1 = \mu_2 \\ H_a : \mu_1 < \mu_2 \end{cases},$$

3 pavyzdys. Prieš pradėdant ruošti naujos paslaugos reklaminį plakatą, vadybininkas nori sužinoti, ar moterų mėgstančių jo siūlomą prekę, procentas pirkėjų populiacijoje yra toks pat kaip ir vyrų. Iš pirkėjų populiacijos atsitiktinai atrinkus 70 moterų ir 50 vyrų paaiškėjo, kad 30 respondentų moterų ir 25 respondentams vyrams patinka siūloma paslauga. Ar turint šiuos duomenis galime priimti sprendimą, kad tarp vyrų siūloma paslauga yra populiareesnė? Populiacijos moterų dalį, kurioms patinka siūloma paslauga pažymėję p_1 , o vyrų dalį, kuriems patinka siūloma paslauga - p_2 , problemą galime užrašyti taip:

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

3.2. Hipotezių tikrinimo klaidos

Statistikoje nagrinėjami atsitiktiniai dydžiai, todėl ir apie hipotezės atmetimą arba priėmimą galime kalbėti tik su tam tikra tikimybe. Taigi, galimos hipotezės tikrinimo klaidos. Pažymėkime α tikimybę, kad teisinga hipotezė H_0 bus atmesta. Tuomet $1 - \alpha$ bus tikimybė, kad teisinga hipotezė H_0 bus priimta. Tikimybė α vadinama **reikšmingumo lygmeniu**. Jis parodo mūsų pasirinktą teisės suklysti laipsnį. Taisyklė, pagal kurią (remiantis imties duomenimis) hipotezė H_0 priimama arba atmetama, vadinama **statistiniu kriterijumi**.

Tikrindami hipotezę H_0 , galime padaryti dviejų rūšių klaidas:

- atmesti hipotezę H_0 , kai ji yra teisinga (pirmos rūšies klaida).
- priimti hipotezę H_0 , kai ji yra klaidinga (antros rūšies klaida).

Pažymėkime α - pirmos rūšies klaidos tikimybę, o β - antros rūšies klaidos tikimybę. Galimi atvejai nurodyti 3.1 lentelėje.

3.1 lentelė. Hipotezių tikrinimo klaidos

		a. Populiacija	
		H_0 teisinga	H_0 klaidinga
Sprendimas pagal imties duomenis	Atmesti H_0	I rūšies klaida su tikimybe α (reikšmingumo lygmuo)	Teisingas sprendimas su tikimybe $1 - \beta$ (kriterijaus galia)
	Neatmesti H_0	Teisingas sprendimas su tikimybe $1 - \alpha$	II rūšies klaida su tikimybe β

Tradicionis nematematinis pavyzdys, motyvuojantis pirmosios rūšies klaidos tikimybės fiksavimą, yra toks; teismas sprendžia, ar teisiamasis kaltas. Hipotezė H_0 - nekaltas, alternatyva H_a - kaltas. Pirmosios rūšies klaida – nekaltą pripažinti kaltu. Antrosios rūšies klaida – kaltą išteisinti. Manoma,

kad pirmosios rūšies klaida pavojingesnė (beje, žinoma ir kitokių požiūrių). Todėl stengiamasi pirmosios rūšies klaidos tikimybę fiksuoti, t.y. imti ją pakankamai mažą.

3.3. Bendroji hipotezės tikrinimo schema

Tikrinant parametrines hipotezes pirmiausia turime **parinkti tikimybinį modelį**. Tai esminis etapas, kuriuo nustatoma **koks yra matuojamojo kintamojo tikimybinis skirstinys**. Šis skirstinys turi vieną arba kelis nežinomus parametrus. Pasirinkdami tinkamą tikimybinį skirstinį, sąmoningai ignoruojame neatitikimus, kurių tikimybės labai mažos. Pavyzdžiui, nusprendžiame, kad klientų išlaidos firmos teikiamoms paslaugoms turi normalųjį skirstinį, nors iš tikro normalusis skirstinys gali įgyti ir neigiamas reikšmes.

Parinkus tikimybinį modelį užrašome **statistinę hipotezę**. Tradiciškai hipotezė H_0 reiškia, kad nėra statistiškai reikšmingo skirtumo tarp jūsų turimų duomenų ir nulinės hipotezės teiginio, o alternatyva H_a - skirtumas yra. Hipotezę H_0 laikome teisinga ir atmetame ją tik tuomet, kai duomenys rodo, kad ji labai mažai tikėtina. Ką parinkti hipoteze, lemia pirmosios ir antrosios rūšies klaidos.

Hipotezei tikrinti parenkama statistika. Ji parenkama atsižvelgiant į populiacijos skirstinį, hipotezes H_0 ir H_a . Statistika sudaroma taip, kaip tuo atveju, kai H_0 teisinga, ji turėtų žinomą skirstinį (standartinį normalųjį, χ^2 , Stjudento, Fišerio ir pan.). Formuluoju hipotezę, kartu pasirenkamas reikšmingumo lygmuo α . Šis pasirinkimas – subjektyvus dalykas. Pagal nusistovėjusią tyrimų tradiciją $\alpha = 0,05$ arba $\alpha = 0,01$.

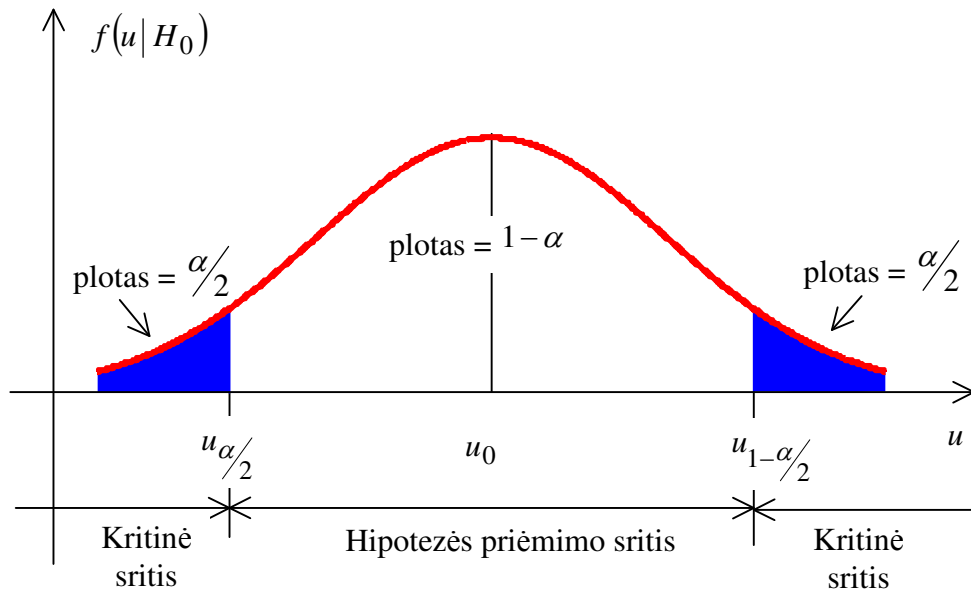
Geriausių statistikų parinkimo klausimai čia nenagrinėjami. Dažniausiai naudojamų statistikų lenteles galima rasti matematinės statistikos vadovėliuose ir žinynuose.

Parinkus **statistiką**, jos galimų reikšmių sritis suskaidoma į dvi nesikertančias sritis: **hipotezės priėmimo sritį ir kritinę sritį**. Taškai, skirianys vieną sritį nuo kitos, vadinami **kritinėmis reikšmėmis** (kritiniais taškais). Jei pagal imties duomenis apskaičiuota statistikos reikšmė (stebėta statistikos reikšmė) patenka į hipotezės kritinę sritį, tai hipotezė H_0 yra nesuderinama su stebėjimo duomenimis ir todėl atmetama, priešingu atveju ji neatmetama (priimama). Pateiksime **bendrąją hipotezės tikrinimo schemą**:

- 1⁰.Formuluojamos nulinė H_0 ir alternatyvioji H_a hipotezės.
- 2⁰.Parenkamas reikšmingumo lygmuo $\alpha \in \{0,1; 0,05; 0,01; \}$.
- 3⁰.Hipotezei H_0 tikrinti parenkama statistika U , kurios tankio funkcija, kai H_0 , teisinga, yra $f(u|H_0)$.
- 4⁰.Parinktam reikšmingumo lygmeniui α randamos kritinės reikšmės, hipotezės priėmimo sritis U_{H_0} ir kritinė sritis U_K . Hipotezės kritinė sritis gali būti dvipusė ir vienpusė. Tai priklauso nuo tikrinamos hipotezės. Jeigu tikrinama hipotezė $H_0 : \Theta = \Theta_0$, o alternatyvi hipotezė yra $H_a : \Theta \neq \Theta_0$, tai **kritinė sritis yra dvipusė**. Šiuo atveju yra dvi kritinės reikšmės - kvantiliai $u_{\alpha/2}$ ir $u_{1-\alpha/2}$, kurie tenkina lygtis:

$$\text{Kritinė sritis yra intervalas } U_K = (-\infty, u_{\alpha/2}] \cup [u_{1-\alpha/2}, +\infty).$$

Hipotezės H_0 priėmimo sritis yra intervalas $U_{H_0} = (u_{\alpha/2}, u_{1-\alpha/2})$.



3.1 pav. Statistikos U tankio funkcija, kai H_0 teisinga (dvipusė kritinė sritis).

Jeigu tikrinama hipotezė $H_0 : \Theta = \Theta_0$, o alternatyvi hipotezė yra $H_a : \Theta < \Theta_0$ ($H_a : \Theta > \Theta_0$), tai **kritinė sritis yra vienpusė kairė (dešinė)**. Kriterijaus su kaire kritine sritimi, kritinė reikšmė yra α kvantilis u_α , o kritinė ir priėmimo sritys

$$U_K = (-\infty, u_\alpha] , \quad U_{H_0} = (u_\alpha, +\infty) .$$

Analogiškai apibrėžiamos kriterijus su dešine kritine sritimi hipotezės H_0 priėmimo ir kritinės sritys:

$$U_{H_0} = (-\infty, u_{1-\alpha}) , \quad U_K = [u_{1-\alpha}, +\infty) .$$

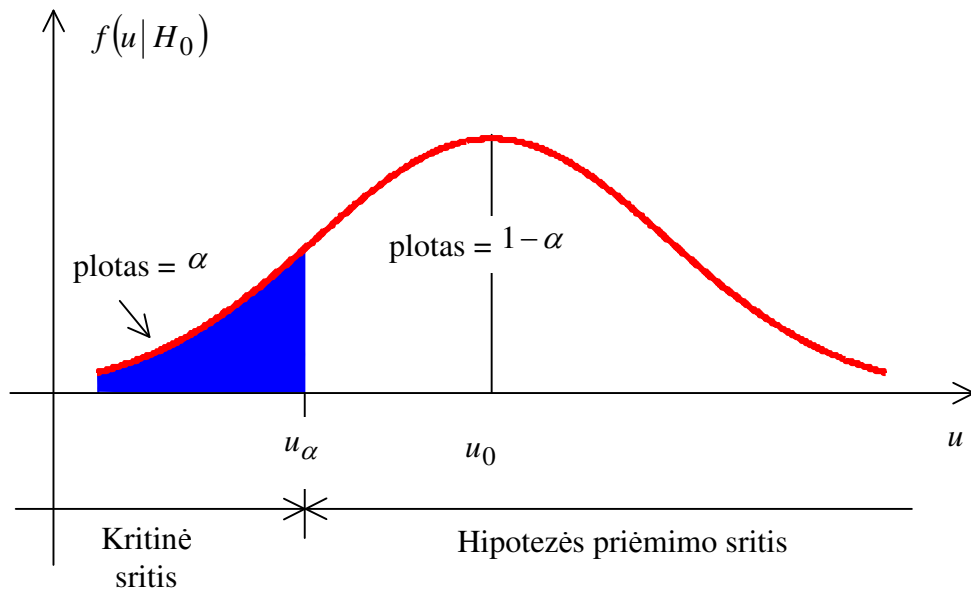
Pagal imties duomenis apskaičiuojama stebėtoji statistikos U reikšmė u_{imt} ir priimamas sprendimas. Jeigu $u_{\text{imt}} \in U_{H_0}$, tai hipotezė H_0 neatmetama, priešingu atveju ji atmetama ir priimama H_a . Jeigu priimama H_a , tai galima teigti, kad jos teisingumas įrodytas. Jeigu neatmetama H_0 , tai reiškia, kad hipotezė H_0 yra suderinama su stebėjimo rezultatais, tačiau gali egzistuoti ir kitos nulinės hipotezės, turinčios tą pačią savybę. Pavyzdžiui, jeigu neatmetama hipotezė $H_0 : \mu = 3$, tai pačiai imčiai galimos ir kitos hipotezės (pavyzdžiui, $H_0 : \mu = 3.3$ arba $H_0 : \mu = 2.5$), kurios taip pat bus suderinamos su stebėjimo rezultatais.

3.4. p-reikšmė ir kriterijaus galia

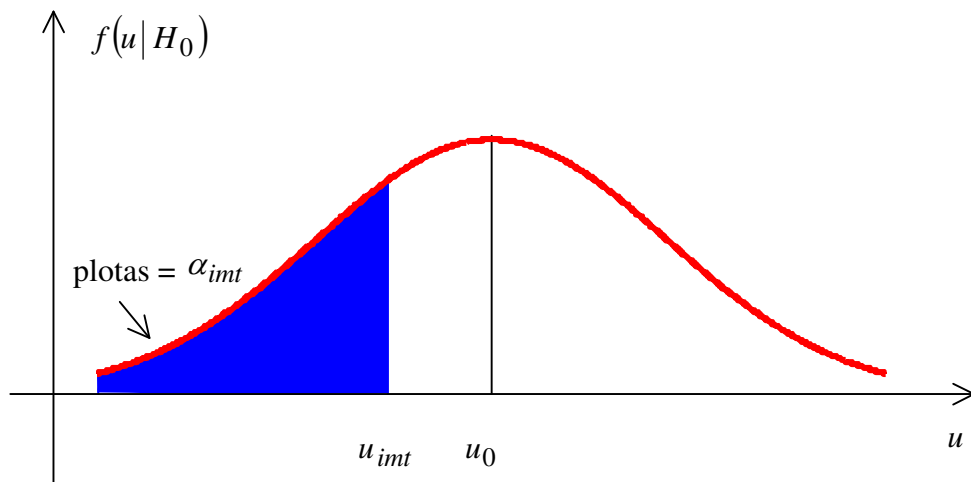
Statistikos paketuose naudojama modifikuota hipotezės tikrinimo schema. Pagal imties duomenis apskaičiuojama tikimybė

$$\alpha_{imt} = \begin{cases} P(U \geq u_{imt}), & \text{kai kritinė sritis yra dešinė,} \\ P(U \leq u_{imt}), & \text{kai kritinė sritis yra kairė,} \\ 2 * \min(P(U \geq u_{imt}), P(U \leq u_{imt})), & \text{kai kritinė sritis yra dvipusė.} \end{cases}$$

t.y. skaičiuojamas stebėtas reikšmingumo lygmuo (mažiausias reikšmingumo lygmuo, su kuriuo teisinga H_0 gali būti atmesta turimiems duomenims). Tikimybė α_{imt} vadinama statistikos u_{imt} **stebėtu reikšmingumo lygmeniu** arba **p-reikšme** (statistikos paketų pateikiamose rezultatų lentelėse žymima **Sig. Level** arba **p-value**). Hipotezė H_0 atmetama, kai $\alpha \geq \alpha_{imt}$, priešingu atveju ji neatmetama. Reikšmės α_{imt} geometrinė interpretacija kriterijui su kaire kritine sritimi pateikta paveiksle.



3.2 pav. Statistikos U tankio funkcija, kai H_0 teisinga (dešinė kritinė sritis).



3.3 pav. Statistikos U tankio funkcija, kai H_0 teisinga stebėta statistikos U reikšmė.

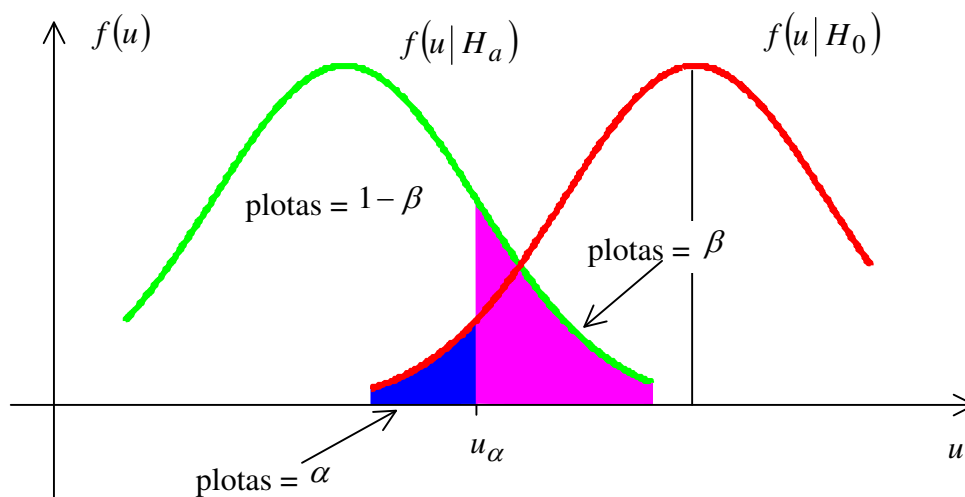
Pateiktu atveju hipotezė H_0 priimama, nes $\alpha < \alpha_{int}$. Tegul α yra reikšmingumo lygmuo, o $p = \alpha_{int}$ **p-reikšmė**. Galime suformuluoti **bendrą sprendimo priėmimo taisyklę**:

Jeigu $p < \alpha$, tai hipotezė H_0 atmetama.

Jeigu $p \geq \alpha$, tai hipotezė H_0 neatmetama.

Pavyzdžiui, jeigu $\alpha = 0,05$, o $p = 0,015$, tai H_0 atmetama. Patogiau yra naudotis p-reikšmėmis nei kritine sritimi, nes jos nėra susietos su konkrečiais reikšmingumo lygmenimis. Pavyzdžiui, $p = 0,002$ galima lyginti ir su $\alpha = 0,05$, ir su $\alpha = 0,01$. Beveik visi statistikos paketai (SPSS, SAS, Statgraphics ir kt.) skaičiuoja p-reikšmes. Jos naudojamos ir formuluojant statistines išvadas– “į butelius įpilto gėrimo kiekis statistiškai reikšmingai skiriasi nuo 0,5 litro ($p < 0,01$)” ir t.t.

Pirmos rūšies klaidos tikimybė α vadinama **reikšmingumo lygmeniu**. Ji parodo mūsų pasirinktą teisės suklysti laipsnį. **Kriterijaus galia** vadinama tikimybė atmesti klaidingą nulinę hipotezę H_0 ir priimti paprastą alternatyvą $H_a : \Theta = \Theta_0$, kai paprastoji hipotezė teisinga. Ši tikimybė lygi $1 - \beta$. Kriterijų stengiamasi sudaryti taip, kad abiejų rūšių klaidų tikimybės būtų minimalios, tačiau tai ne visada įmanoma. I ir II rūšies klaidų geometrinė interpretacija kriterijui su kaire kritine sritimi pateikta paveiksle.



3.4 pav. I ir II rūšies klaidų grafinė interpretacija

Nesunku pastebėti, kad, praplėtus kritinę sritį (stumiant kvantilį u_α į dešinę). II rūšies klaidos tikimybė sumažės, bet padidės I rūšies klaidos tikimybė, ir atvirkščiai. Todėl paprastai parenkama reikšmė, kurios neturėtų viršyti pirmos rūšies klaidos tikimybė, ir stengiamasi minimizuoti antros rūšies klaidos tikimybę. Kuo mažesnis α , tuo sunkiau atmesti H_0 . Kai $\alpha = 0$ priimamos visos hipotezės H_0 , taip pat ir neteisingos. Dažniausiai pasirenkama $\alpha = 0,05$, tačiau nereikia suabsoliutinti 0,05 reikšmingumo lygmens svarbos. Jis tik parodo mūsų pasirinktą teisės suklysti laipsnį, o tai priklauso nuo sprendžiamo uždavinio.

Antros rūšies klaidos tikimybė β priklauso nuo imties didumo n , reikšmingumo lygmens α , alternatyvios hipotezės H_a ir kriterijaus

3.5. Išvadų formulavimas

Tarkime, tikrinamame hipotezę:

$$\begin{cases} H_0 : \theta = \theta_0, \\ H_a : \theta \neq \theta_0. \end{cases}$$

Jeigu nulinę hipotezę atmetame, tai formuluodami išvadą, sakome “statistiškai reikšmingai skiriasi”. Vadinasi, parametro taškinio įverčio $\hat{\theta}$ ir populiacijos spėjamos parametro reikšmės θ_0 skirtumas didelis ir labai mažai tikėtina, kad to priežastis yra imties atsitiktinumas. Jeigu imtis tokia, kad H_0 neatmetame, tai sakome, kad gauto ir spėjamo vidurkių skirtumas statistiškai nereikšmingas. Mes neteigiame, kad imties vidurkis ir spėjamasis nesiskiria; sakome, kad nėra pagrindo manyti, jog populiacijos vidurkis skiriasi nuo spėjamojo, o imties ir spėjamojo vidurkių skirtumą galima paaiškinti atsitiktinumu. Primename, kad **statistinės išvados daromos su tam tikra tikimybe**. Tam tarnauja ir pasirenkamas reikšmingumo lygmuo α . Jeigu H_0 neatmetame, tai dažniausiai išvada taip ir formuluojama: “ H_0 neatmetama” (tarytum rekomenduojama susilaikyti nuo galutinio sprendimo), o ne “ H_0 teisinga”.

Reikia skirti **statistiškai reikšmingą skirtumą** nuo **tyrimo prasme reikšmingo skirtumo**. Jeigu matavimų skaičius mažas, tai H_0 retai atmetama. Pavyzdžiui, tarkime, ištyrus 5 privačių ir 5 valstybinių darbuotojų algas, gautos vidutinės algos yra atitinkamai 4000 ir 3000 Lt, bet statistiškai reikšmingo skirtumo nėra. Nerastas skirtumas šiuo atveju yra per mažų imčių pasekmė – iš 5 stebėjimų negalima daryti išvados apie visas populiacijas. Galima ir prieštaringa situacija – didelėms imtims statistiškai reikšmingas pripažintas net menkiausias skirtumas (tokia statistikų skirstinių prigimtis). Tačiau nebūtinai skirtumas iš tikro reikšmingas (tyrimo prasme). Pavyzdžiui, jeigu didelėje įmonėje 1000 vyrų ir 2000 moterų pradinių atlyginimų vidurkiai skiriasi 5 litais ir šis skirtumas statistiškai reikšmingas, tai dar nereiškia, kad tie 5 litai iš tikro yra reikšmingi tyrimo prasme.

3.6. Hipotezės apie populiacijos vidurkį tikrinimas

Pavyzdys. Stebime atsitiktinį dydį X . Tarkime, kad jo skirstinys yra normalusis $X \sim N(\mu, \sigma^2)$. Parametrai μ ir σ^2 nežinomi. Pagal imties $x = (x_1, \dots, x_n)$ duomenis tikrinsime hipotezę $H_0: \mu = \mu_0$. Nagrinėsime tą atvejį, kai alternatyva yra $\mu \neq \mu_0$.

1⁰.Formuluojamos hipotezės H_0 ir H_a

$$H_a : \mu = \mu_0, \quad H_a : \mu \neq \mu_0.$$

2⁰.Parenkamas reikšmingumo lygmuo α .

3⁰.Hipotezei H_0 tikrinti parenkama statistika

$$T = \frac{\bar{X} - \mu_0}{S} \sqrt{n},$$

turinti Stjudento skirstinį su $n-1$ laisvės laipsnių.

4⁰. Randame dvi kritines reikšmes – du Stjudento kvantilius $t_{\alpha/2; n-1}$, $t_{1-\alpha/2; n-1}$ ir hipotezės H_0 priėmimo sritį $\mathbf{T}_{H_0}$ bei kritinę sritį $\mathbf{T}_K$:

$$\mathbf{T}_{H_0} = (t_{\alpha/2; n-1}, t_{1-\alpha/2; n-1}),$$

$$\mathbf{T}_K = (-\infty, t_{\alpha/2; n-1}] \cup [t_{1-\alpha/2; n-1}, +\infty).$$

5⁰. Pagal imties duomenis apskaičiuojama statistikos reikšmė t_{imt} ir priimamas statistinis sprendimas. Jeigu $t_{imt} \in T_{H_0}$ hipotezė H_0 neatmetama, priešingu atveju ji atmetama ir priimama H_a .

Jeigu alternatyvioji hipotezė yra $H_a : \mu < \mu_0$, tai naudojamas kriterijus su kaire kritine sritimi:

$$T_K = (-\infty, t_{\alpha, n-1}], \quad T_{H_0} = (t_{\alpha, n-1}, +\infty).$$

Jeigu alternatyvioji hipotezė yra $H_a : \mu > \mu_0$, tai naudojamas kriterijus su dešine kritine sritimi:

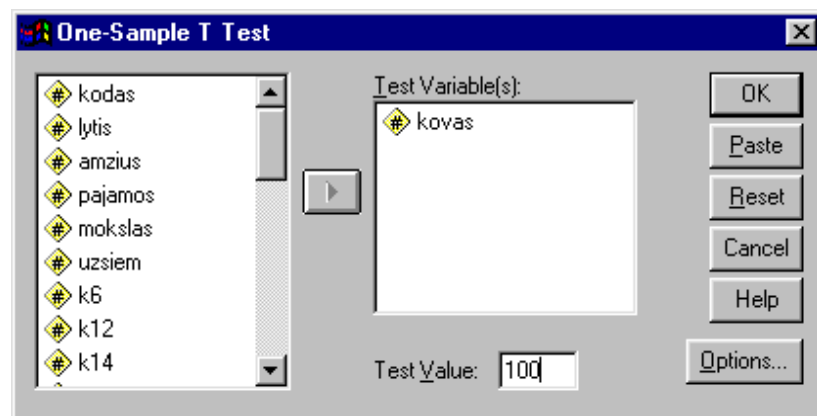
$$T_{H_0} = (-\infty, t_{1-\alpha, n-1}], \quad T_K = (t_{1-\alpha, n-1}, +\infty).$$

Pavyzdys. Ar pagal turimus apklausos duomenis galime teigti, kad firmos klientų vidutinės išlaidos telekomunikacijų paslaugoms kovo mėnesį neviršija 100 litų per mėnesį ($\alpha = 0,05$).

Sprendimas: Formuluoame statistinę hipotezę:

$$\begin{cases} H_0 : \mu = 100, \\ H_a : \mu < 100. \end{cases}$$

SPSS paketu (Analyze→Compare Means→One-sample T-Test)



One-Sample Statistics

	N	Mean	Std. Deviation	Std. Error Mean
Kovas	1056	98,30	37,77	1,16

One-Sample Test

Test Value = 100						
	t	df	Sig. (2-tailed)	Mean Difference	95% Confidence Interval of the Difference	
					Lower	Upper
Kovas	-1,459	1055	,145	-1,70	-3,98	,58

3.5 pav. Hipotezės apie vidurkį tikrinimo rezultatai

Gavome, kad $n = 1055$, $\bar{x} = 98,30$, $s = 37,77$. Stebėta statistikos reikšmė $t_{imt} = -1,45$. Skirtumas tarp imties vidurkio ir spėjamo populiacijos vidurkio lygus $\bar{x} - 100 = -1,7$. Skirtumo tarp tikrosios ir populiacijos vidurkio reikšmės 0,95 pasikliautinis intervalas

$$PI_{0,95}(\mu - 100) = (-3,98; 0,58).$$

Su 95% garantija galime teigti, kad $-3,98 < \mu - 100 < 0,58$ arba $96,02 < \mu < 100,58$.

PASTABA: dažnai tikrinant hipotezes statistikos paketai skaičiuoja p reikšmę tik vienai alternatyvai, o kitoms alternatyvoms Jūs turite apskaičiuoti patys!!!

Pavyzdžiui, **t kriterijui** paketas SPSS skaičiuoja **p reikšmę** tik dvipusei kritinei sričiai

(**Sig 2-tailed**), t.y. hipotezei
$$\begin{cases} H_0 : \theta = \theta_0, \\ H_a : \theta \neq \theta_0. \end{cases}$$

Vienpusėms alternatyvoms p reikšmės Jūs turite apskaičiuoti patys:

- jeigu alternatyva yra $H_a : \theta < \theta_0$ ir stebėta Stjudento statistikos reikšmė $t_{imt} \leq 0$, taip p reikšmė $= p^* / 2$, čia p^* - dvipusės kritinės srities p reikšmė, jeigu $t_{imt} > 0$, tai p reikšmė $= 1 - p^* / 2$;
- jeigu alternatyva yra $H_a : \theta > \theta_0$ ir $t_{imt} > 0$, taip p reikšmė $= p^* / 2$, jeigu $t_{imt} \leq 0$, tai p reikšmė $= 1 - p^* / 2$.

Tos pačios taisyklės tinka ir standartinio normaliojo skirstinio statistikai Z.

Mūsų atveju $p^* = 0,145$. Kadangi $t_{imt} = -1,459 < 0$, tai $p = p^* / 2 = 0,145 / 2 = 0,0725 > 0,05$ ir hipotezę H_0 **neatmetame**.

3.7. Hipotezės apie proporciją tikrinimas

Tarkime stebimas atsitiktinis dydis X įgyja tik dvi reikšmes 0 ir 1 (taip gali būti koduojami atsakymai į anketos klausimus, kuriuose nurodytos dvi kategorijos: prieš/už, nesutinku/sutinku, nepritariu/pritariu, nepatinka/patinka ir t.t.). Tarkime, norime patikrinti hipotezę, kad apklausos teiginiui pritaria p_0 populiacijos dalis:

$$\begin{cases} H_0 : p = p_0 \\ H_a : p \neq p_0 \end{cases}$$

Kai imtis didelė ($n > 30$) hipotezės tikrinimui naudojama statistika

$$Z = \frac{\hat{p} - p_0}{\sqrt{p_0(1 - p_0)}} \cdot \sqrt{n},$$

turinti standartinį normalųjį skirstinį $Z \sim N(0,1)$, čia $\hat{p} = \frac{m}{n}$, n- imties didumas, m - „vienetukų“ skaičius imtyje.

Pavyzdys. Prieš pradėdami teikti naują paslaugą, reklamos skyriaus darbuotojai paprašė 100 atsitiktinai atrinktų įmonės klientų įvertinti naują paslaugą. Teigiamai naują paslaugą įvertino 63 klientai. Ar šie duomenys neprieštarauja naujos paslaugos autorių teiginiui, kad paslauga patiks bent dviem iš trijų klientų? ($\alpha = 0,01$).

Sprendimas. Formuluojuame statistinę hipotezę:

$$\begin{cases} H_0 : p = 2/3, \\ H_a : p < 2/3. \end{cases}$$

Apskaičiuojame $Z = (0.63 - 2/3) / (\sqrt{(2/3)(1/3)/100}) = -0.777$.

Kadangi $z_{imt} = -0.777 > -2.326 = z_{0,01}$ tai hipotezės H_0 neatmetame. Imties duomenys neprieštarauja paslaugos autorių teiginiui.

SPSS paketu (**Analyze→Noparametric Tests→Binomial**) gauname tą patį rezultatą:

Binomial Test					
	Category	N	Observed Prop.	Test Prop.	Asymp. Sig. (1-tailed)
MESAINIS	Group 1 Patinka	63	,63	,667	,249^a
	Group 2 Nepatinka	37	,370		
	Total	100	1,000		

a. Alternative hypothesis states that the proportion of cases in the first group < ,667.

b. Based on Z Approximation.

3.6 pav. Hipotezės apie proporciją tikrinimo rezultatai (SPSS)

p-reikšmė lygi 0,249>0,01 hipotezė H_0 neatmetama.

3.8. Suderinamumo hipotezių tikrinimas

Nagrinėjant parametrinių hipotezių tikrinimo uždavinius, stebimo atsitiktinio dydžio skirstinys nusakomas **žinomos** analizės išraiškos skirstinio funkcija $F(x, \Theta_1, \dots, \Theta_h)$, priklausančia nuo h nežinomų parametrų. Dabar kalbėsime apie hipotezių tikrinimo uždavinius, kai **skirstinio funkcijos analizinė išraiška nežinoma**, t.y. kai mes nežinome koks yra stebimo atsitiktinio dydžio skirstinys.

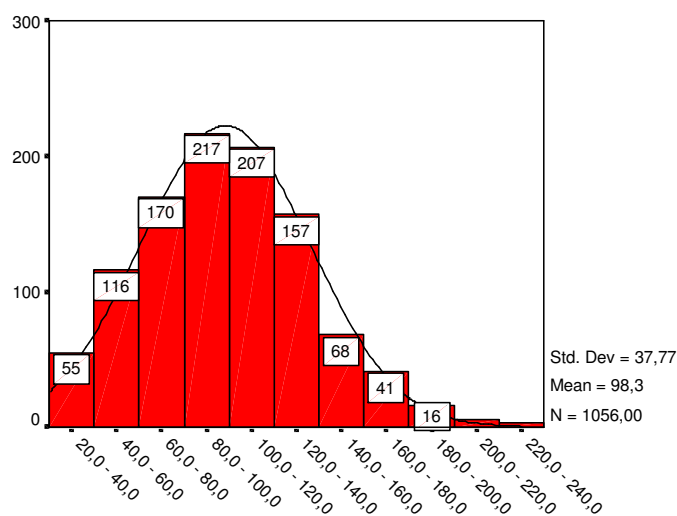
Tarkime, kad $X=(X_1, \dots, X_n)$ yra dydžio X atsitiktinė imtis. Suformuluosime suderinamumo hipotezę: „ X skirstinys yra normalusis”.

Pavyzdžiui, **ar klientų išlaidos telekomunikacijų paslaugoms kovo mėnesį aprašomos normaliuoju skirstiniu?**

Tolydžiojo atsitiktinio dydžio empirinį skirstinį su teoriniu galima palyginti grafiškai, nubraižant histogramą ir teorinio skirstinio tankio funkciją. Pavyzdžiui, duota išlaidų telekomunikacijų paslaugoms kovo mėnesį imtis. Nubraižysime histogramą ir normaliojo skirstinio tankio funkciją.

SPSS meniu: Graphs→Histogram... (Display normal curve).

Imties tūris $n=1056$. Didžiausios respondentų išlaidos yra 234 Lt, o mažiausios - 21 Lt. Imties plotis $IP=234-21=213$ Lt. Apskaičiuojame rekomenduojamą dalinių intervalų skaičių $r=[1+3,32*\lg(1056)]=[11,02]=11$, čia $[x]$ pažymėtas apvalinimo veiksmas.



Išlaidos telekomunikacijų paslaugoms kovo mėnesį

3.7 pav. Histograma ir normaliojo skirstinio tankio funkcija

Parenkame a_0 ir a_r taip, kad dalinio intervalo ilgis

$$h = \frac{a_r - a_0}{r}$$

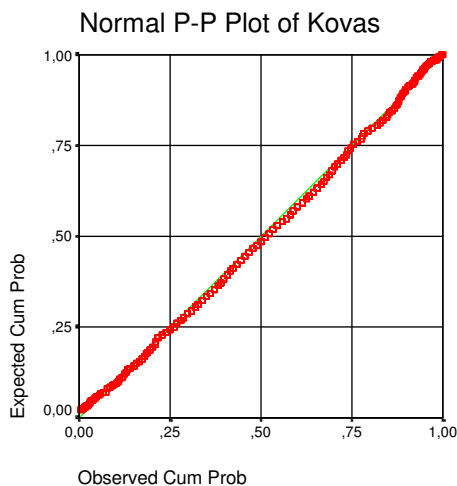
būtų “gražus” skaičius. Mūsų atveju geriausia pasirinkti

$$a_0 = 20, a_r = 240, \text{ nes } h = \frac{240 - 20}{11} = 20.$$

Nubraižome histogramą su parinktais parametrais ir matome, kad yra geras empirinio skirstinio suderinamumas su teoriniu (šiuo atveju normaliuoju skirstiniu).

Grafiškai empirinio **skirstinio funkcijos** $F_n(x)$ ir teorinio skirstinio funkcijos $F(x)$ skirtumą galima pavaizduoti braižant **P-P grafiką** (empirinio ir teorinio skirstinių funkcijų palyginimo grafiką).

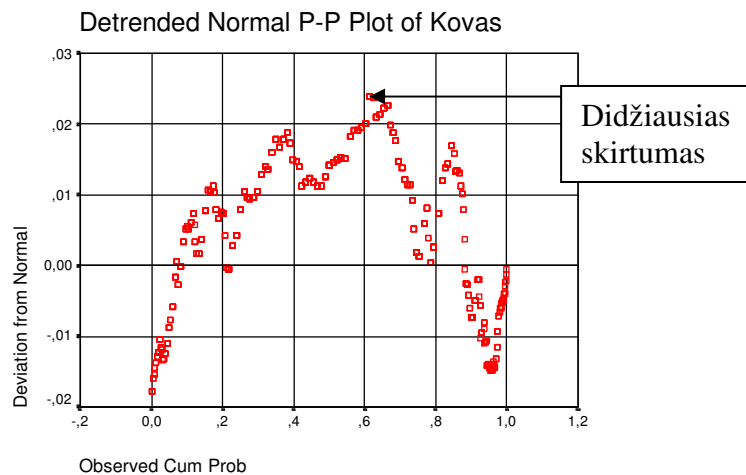
SPSS meniu: Graphs→P-P (Test Distribution: **normal**, Distribution parameters: **Estimate from data**)...



3.8 pav. P-P grafikas (empirinio ir teorinio skirstinių funkcijų palyginimo grafikas)

Kuo arčiau tiesės išsidėsto taškai, tuo geresnis empirinio skirstinio suderinamumas su teoriniu (šiuo atveju normaliuoju skirstiniu).

Sekantis grafikas vaizduoja skirtumą $\hat{F}_n(x) - F(x)$.



3.9 pav. Empirinės ir teorinės skirstinio funkcijų skirtimo grafikas $\hat{F}_n(x) - F(x)$

3.8.1. Kolmogorovo Smirnovio suderinamumo kriterijus

Suderinamumo hipotezės rekomenduojama tikrinti įvairiais būdais. Remiantis skirtingais kriterijais, atstumas tarp teorinio ir empirinio skirstinio išmatuojamas įvairiais matais. Todėl, taikant keletą kriterijų, gaunama daugiau informacijos.

Kolmogorovo-Smirnovio kriterijus taikomas tikrinant suderinamumo hipotezę:

$$H_0 : \text{Atsitiktinio dydžio } X \text{ skirstinio funkcija yra } F(x).$$

Apibrėšime Kolmogorovo-Smirnovio kriterijų. Tarkime, kad atsitiktinio dydžio X skirstinio funkcija yra $F(x)$. Pagal imties duomenis randama empirinė skirstinio funkcija $F_n(x)$. Kolmogorovas ir Smirnovas suderinamumo hipotezės tikrinimui pasiūlė statistiką, kuri matuojant atstumą tarp empirinio ir teorinio skirstinių įvertina skirtumą tarp empirinio skirstinio funkcijos $F_n(x)$ ir teorinio skirstinio funkcijos $F(x)$:

$$Z = \sqrt{n} \max_x |F_n(x) - F(x)|,$$

čia n – imties didumas.

Taikant paketą SPSS suderinamumo hipotezę galima patikrinti naudojant procedūrą NPAR TEST.

SPSS meniu: Analyze → Nonparametric Tests → 1-Sample-K-S

Mūsų anksčiau nagrinėto pavyzdžio SPSS rezultatai pateikti lentelėje.

One-Sample Kolmogorov-Smirnov Test		
		Kovas
N		1056
Normal Parameters ^{a,b}	Mean	98,30
	Std. Deviation	37,77
Most Extreme Differences	Absolute	,031
	Positive	,031
	Negative	-,020
Kolmogorov-Smirnov Z		1,011
Asymp. Sig. (2-tailed)		,258

a. Test distribution is Normal.

b. Calculated from data.

3.10 pav. Suderinamumo hipotezės tikrinimo rezultatai (Kolmogorovo Smirnov suderinamumo kriterijus, SPSS)

Joje nurodyti: imties didumas ($n=1056$); normaliojo skirstinio parametrų taškiniai įverčiai apskaičiuoti pagal negrupuotus duomenis (vidurkis 98,30 Lt. ir standartinis nuokrypis 37,77 Lt.); didžiausio skirtumo tarp empirinio skirstinio funkcijos $F_n(x)$ ir teorinio skirstinio funkcijos $F(x)$ absoliutinis didumas $\max_x |F_n(x) - F(x)|$ (Most extreme absolute difference 0,031); didžiausias teigiamas ir didžiausias neigiamas skirtumai (Most extreme positive and negative differences: 0,031 ir -0,020); Kolmogorovo ir Smirnov statistikos realizacija (Kolmogorov-Smirnov $Z = 1,011$) ir p -reikšmė 0,258. Kadangi $p=0,258 > 0,05$ tai atmesti H_0 nėra pagrindo.

Išvada. Galime teigti, kad išlaidos telekomunikacijos paslaugoms kovo mėnesį turi normalųjį skirstinį.

3.8.2. Chi kvadratu suderinamumo kriterijus

Vienas iš populiariausių suderinamumo kriterijų yra χ^2 (chi kvadratu) kriterijus. Jis naudojamas hipotezėms apie kintamojo skirstinį (normalųjį, binominį ir t.t.) populiacijoje tikrinti. χ^2 kriterijus parodo, ar empirinio ir teorinio skirstinių skirtumas yra reikšmingas, t.y. tikrinama, ar turimas empirinis skirstinys suderinamas su teoriniu modeliu. χ^2 kriterijus taikomas grupuotiems duomenims, todėl imtis turi būti gana didelė.

Bendroji kriterijaus sudarymo schema yra tokia:

1. Jei stebimas diskretusis kintamasis, tai iš pradžių apskaičiuojami imties reikšmių (kategorijų dažniai).
2. Jei stebimas tolydusis kintamasis, tai reikšmių sritis suskaidoma į nesikertančius intervalus ir apskaičiuojami intervaliniai dažniai.

3. Tarkime, kategorijų (intervaliniai) dažniai yra $O_1, O_2, \dots, O_k$, čia k – kategorijų (intervalų) skaičius. Naudodamiesi teorinio (nurodyto hipotezės H_0 formuluotėje) savybėmis, apskaičiuojame, kiek kintamojo reikšmių turėtų būti priskirta kiekvienai kategorijai (patektų į kiekvieną intervalą), jei hipotezė apie kintamojo skirstinį būtų teisinga, t.y. randame tikėtinius dažnius $E_1, E_2, \dots, E_k$.
4. Įvertiname tikėtinų ir stebimų dažnių skirtumus. Kuo šie skirtumai didesni, tuo labiau abejotina, kad hipotezė apie skirstinį teisinga. Sprendimo priėmimo taisyklės grindžiamos šiuo tikėtinų ir stebimų dažnių skirtumų didumu.

Aprašysime, kaip taikomas suderinamumo kriterijus χ^2 , kai stebimas diskretusis kintamasis. Tarkime, kad pagal stebimą kintamąjį populiaciją galima suskirstyti į k kategorijų. Populiacijos, priklausančios j -ajai kategorijai, dalį pažymime p_j , $j = 1, \dots, k$ (ekvivalenti formuluotė: stebimas atsitiktinis dydis, su tikimybe p_j įgyjantis j -ąją reikšmę). Jei hipotezė apie skirstinį H_0 teisinga, tai stebimojo kintamojo skirstinys yra žinomas ir tikimybė priklausyti j -ajai kategorijai yra p_j^0 . Suformuluosime hipotezę apie empirinio ir teorinio skirstinių suderinamumą:

$$\begin{cases} H_0 : p_1 = p_1^0, & p_2 = p_2^0, \dots, p_k = p_k^0, \\ H_a : p_j \neq p_j^0, & \text{bent vienam iš } j = 1, \dots, k. \end{cases}$$

Kai H_0 teisinga, iš n stebėjimų imties $E_j = np_j^0$ stebėjimų turėtų priklausyti j -ajai kategorijai, jei iš tikrųjų priklauso O_j , tai skirtumas $O_j - E_j$ rodo, ar hipotezė H_0 tikėtina. Diskrečiojo skirstinio atveju, empirinio ir teorinio skirstinių suderinamumui tikrinti naudojama statistika

$$\chi^2 = \sum_{j=1}^k \frac{(O_j - E_j)^2}{E_j}.$$

Jei hipotezė H_0 yra teisinga, tai kriterijaus statistika turi χ^2 skirstinį su $k-1$ laisvės laipsnių. Tikrinant hipotezę H_0 naudojamas kriterijus su dešine kritine sritimi.

$$P(\chi^2 \geq \chi_{1-\alpha, k-1}^2) = \alpha.$$

H_0 neatmetimo sritis $\chi_{H_0}^2 = [0; \chi_{1-\alpha, k-1}^2)$, kritinė sritis $\chi_K^2 = [\chi_{1-\alpha, k-1}^2, \infty)$. Jei stebėta statistikos reikšmė priklauso hipotezės H_0 neatmetimo sričiai $\chi_{imt}^2 \in \chi_{H_0}^2$, tai H_0 neatmetama, priešingu atveju – atmetama.

Pastaba. Taikant χ^2 kriterijų, reikalaujama, kad kiekvienos kategorijos tikėtinas (teorinis) dažnis būtų ne mažesnis už 5, t.y. $np_j^0 \geq 5$, $j=1, 2, \dots, k$.

Pavyzdys. Atsitiktinai parinktų 60 klientų atsakė į klausimą, kokia paslauga geriausia (iš 6 firmos teikiamų paslaugų). Rezultatai pateikti lentelėje.

Geriausia paslauga

Paslauga	1	2	3	4	5	6
Stebėtas dažnis O_j	5	8	10	12	12	13

Ar remiantis šiais duomenimis galima sakyti, kad visų **paslaugų reitingai vienodi**?

Sprendimas. Tarkime, atsitiktinis dydis (kintamasis) X yra numeris paslaugos, kurią renkasi atsitiktinis klientas. Tada teorinis X tikimybių skirstinys aprašomas lentele.

X	1	2	3	4	5	6
P	p_1^0	p_2^0	p_3^0	p_4^0	p_5^0	p_6^0

Joje p_j^0 žymi tikimybę, kad klientas geriausia pripažins j -ąją, $j=1,\dots,6$, paslaugą. Kitais žodžiais tariant p_j^0 žymi populiacijos dydį, kuri geriausia laiko j -ąją paslaugą. Mus domina, ar tikrai j -ąją paslaugą geriausia laiko p_j^0 populiacijos dalis. Formuluojant tikimybių teorijos terminais, mums reikia patikrinti hipotezę, kad turima duomenų aibė gauta iš populiacijos, kurios skirstinys apibrėžtas aukščiau pateikta lentele. Stebimieji dažniai o_j (O_j realizacijos) pateikti pirmoje lentelėje, tikėtinus dažnius e_j (E_j realizacijas) randame imdami p_j^0 iš antros lentelės.

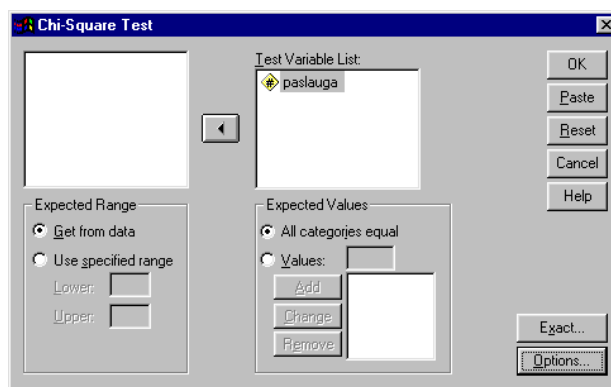
Klientų populiacijoje, kurioje visos paslaugos vertinamos vienodai, kiekvienai iš paslaugų pirmenybę atiduoda $1/6$ klientų. Tada $p_j^0 = 1/6$, $j=1,\dots,6$. Kadangi $n=60$, o visi p_j^0 lygūs tai ir visi $e_j = np_j^0 = 60 \cdot (1/6) = 10$. Pagal formulę apskaičiuojame statistikos realizaciją

$$\chi^2 = \frac{(5-10)^2}{10} + \frac{(8-10)^2}{10} + \frac{(10-10)^2}{10} + \frac{(12-10)^2}{10} + \frac{(12-10)^2}{10} + \frac{(13-10)^2}{10} = 4,6.$$

Kadangi, $\chi^2 = 4,6 < 11,070 = \chi_{0,95;5}^2$, hipotezės H_0 atmesti negalima.

Išvada. Duomenys neprieštarauja hipotezei, kad paslaugų reitingai yra vienodi.

Taikant paketą SPSS, suderinamumo hipotezę galima patikrinti naudojant procedūrą NPAR TEST arba meniu **Analyze**→**Noparametric Tests**→**Chi -Square**.



Mūsų anksčiau nagrinėto pavyzdžio SPSS rezultatai pateikti lentelėje, kurioje nurodytos stebimos reikšmės o_i , tikėtinosios reikšmės e_i ir jų skirtumai $(o_i - e_i)$.

Paslauga

	Observed N	Expected N	Residual
1	5	10,0	-5,0
2	8	10,0	-2,0
3	10	10,0	,0
4	12	10,0	2,0
5	12	10,0	2,0
6	13	10,0	3,0
Total	60		

Kitoje SPSS lentelėje pateikiama statistikos realizacija $Chi-Square = \chi^2_{int} = 4,600$, laisvės laipsnių skaičius $df = (k - 1) = 6 - 1 = 5$ ir p-reikšmė 0,467. Kadangi $0,467 \geq 0,05$, tai H_0 atmesti nėra pagrindo.

Test Statistics

	Paslauga
Chi-Square ^a	4,600
df	5
Asymp. Sig.	,467

a. 0 cells (,0%) have expected frequencies less than 5. The minimum expected cell frequency is 10,0.

Išvada. Duomenys neprieštaruoja hipotezei, kad paslaugų retinai yra vienodi.

4. VIDURKIŲ PALYGINIMAS. DISPERSINĖ ANALIZĖ

Šiame skyriuje nagrinėsime dviejų ir daugiau populiacijų vidurkių skirtumus. Tikrieji populiacijų vidurkiai dažnai yra nežinomi, todėl apie jų skirtumus spėsime pagal vidurkių įverčių skirtumus (imčių vidurkių skirtumus). Kai imčių vidurkių skirtumai dideli, labai mažai tikėtina, kad tai atsitiktinumo pasekmė. Tuomet sakome, kad imčių vidurkiai statistiškai reikšmingai skiriasi ir didelė tikimybė, jog šia prasme skiriasi ir pačių populiacijų vidurkiai. Tarkime, kad mus domina:

- ar naktinės pamainos darbo našumas mažesnis nei dieninės;
- ar trijų prekių vertinimo vidurkiai pirkėjų populiacijoje skiriasi statistiškai reikšmingai;
- ar penkių vertybinių popierių kainų stabilumas biržoje yra vienodas.

Visais šiais atvejais reikia lyginti du ir daugiau nepriklausomų kintamųjų. Pavyzdžiui, X-dieninės pamainos darbininko pagamintų detalių skaičius ir Y - naktinės pamainos darbininko pagamintų detalių skaičius.

Hipotezės apie populiacijų vidurkių skirtumus dviems nepriklausomoms arba dviems porinėms imtims tikrinamos naudojant Stjudento statistikas. Šiame skyrelyje nagrinėsime hipotezių apie dviejų vidurkių skirtumą tikrinimą tik dviems nepriklausomoms imtims. Kai reikia patikrinti hipotezę apie trijų ir daugiau vidurkių lygybę naudojama dispersinė analizė. Pirmuosius darbus iš dispersinės analizės paskelbė Fišeris (1918, 1925, 1935 m.) Dispersinė analizė yra atsitiktinių dydžių skirstinių priklausomybės nuo tam tikrų faktorių tyrimas. Faktoriai gali būti kiekybiniai ar kokybiniai Vienas iš pagrindinių dispersinės analizės tikslų – ištirti ar priklausomo kintamojo Y, išmatuoto skirtingose populiacijose, vidurkiai skiriasi. Šiame kurse apsiribosime dispersinės analizės su pastoviais faktoriais modeliais.

4.1. Hipotezės dėl dviejų populiacijų vidurkių skirtumo tikrinimas

4.1.2. Dviejų nepriklausomų imčių T-kriterijus

Dažnai tenka lyginti dviejų atsitiktinių dydžių nežinomus vidurkius. 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 $X \sim N(\mu_X, \sigma_X^2)$ ir $Y \sim N(\mu_Y, \sigma_Y^2)$, kurių vidurkiai μ_X ir μ_Y nežinomi. Abiejų dydžių dispersijos taip pat nežinomos. Norime patikrinti hipotezę:

$$H_0 : \mu_X = \mu_Y,$$

$$H_a : \mu_X \neq \mu_Y, \text{ (arba } H_a : \mu_X > \mu_Y \text{ arba } H_a : \mu_X < \mu_Y \text{)}.$$

Hipotezės H_0 tikrinimui **kai dispersijos lygios** $\sigma_X^2 = \sigma_Y^2 = \sigma^2$, naudojama Stjudento statistika

$$T = \frac{\bar{X} - \bar{Y}}{\sqrt{S_p^2(1/n + 1/m)}},$$

turinti Stjudento skirstinį su $(n+m-2)$ laisvės laipsnių, čia $S_p^2 = \frac{(n-1)S_X^2 + (m-1)S_Y^2}{n+m-2}$ yra jungtinės populiacijos dispersijos σ^2 įvertis. Įstatę S_p^2 išraišką į statistikos T formulę gauname:

$$T = \frac{\bar{X} - \bar{Y}}{\sqrt{(n-1)S_Y^2 + (m-1)S_Y^2}} \sqrt{\frac{nm(n+m-2)}{n+m}}.$$

Statistika T, taip pat naudojama vidurkių skirtumo $1-\alpha$ pasikliautinajam intervalui konstruoti:

$$(\bar{X} - \bar{Y}) \pm t_{1-\alpha/2; n+m-2} \sqrt{S_p^2 (1/n + 1/m)}.$$

Kai stebimų kintamųjų **dispersijos nelygios** $\sigma_X^2 \neq \sigma_Y^2$ susiduriame su vadinamąja Berenso-Fišerio problema. Žinomi keli apytiksliai šios problemos sprendimai. Pateiksime vieną iš jų, kuris realizuotas sistemoje SPSS. Hipotezės H_0 tikrinimui naudojama Stjudento statistika

$$T = \frac{\bar{X} - \bar{Y}}{\sqrt{S_x^2/n + S_y^2/m}},$$

turinti Stjudento skirstinį su $\nu = \frac{1}{\nu_x + \nu_y}$ laisvės laipsnių,

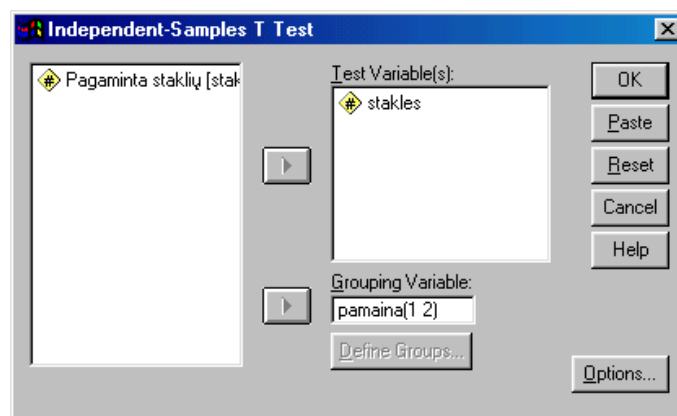
$$\text{čia } \nu_x = \left(\frac{\frac{S_x^2}{n}}{\frac{S_x^2}{n} + \frac{S_y^2}{m}} \right)^2 \cdot \frac{1}{n-1}, \quad \nu_y = \left(\frac{\frac{S_y^2}{m}}{\frac{S_x^2}{n} + \frac{S_y^2}{m}} \right)^2 \cdot \frac{1}{m-1}.$$

Pavyzdys Firmos vadovai nori patikrinti, ar naktinės pamainos darbo našumas mažesnis nei dieninės. Per 20 darbo dienų dieninėje pamainoje pagaminta 20; 19; 22; 18; 22; 24; 25; 23; 22; 22; 19; 21; 26; 25; 23; 17; 18; 17; 22; 20 gręžio staklių. Per tą patį laikotarpį naktinėje pamainoje pagaminta 20; 19; 19; 17; 20; 23; 23; 20; 20; 21; 19; 21; 22; 20; 15; 16; 17; 17; 16; 17 gręžio staklių. Tarkime, kad reikšmingumo lygmuo $\alpha = 0,05$.

Sprendimas. Formuluojuame statistinę hipotezę:

$$\begin{cases} H_0 : \mu_X = \mu_Y, \\ H_a : \mu_X > \mu_Y. \end{cases}$$

SPSS paketu (**Analyze**→**Compare Means**→**Independent-Samples T-Test**) gauname:



Group Statistics

	Pamaina	N	Mean	Std. Deviation	Std. Error Mean
Pagaminta staklių	Dieninė	20	21,25	2,69	,60
	Vakarinė	20	19,10	2,34	,52

4.2 pav. Aprašomosios statistikos lentelė

$\bar{x} = 21,25$; $s_x = 2,69$, $n = 20$; $\bar{y} = 19,1$; $s_y = 2,34$, $m = 20$. SPSS paketo rezultatų lentelėje pateikiamos dvi Stjudento statistikos realizacijos (žiūr. sekančioje lentelėje): viena kai dispersijos lygios, kita kai dispersijos nelygios. Todėl tikrinant hipotezę apie vidurkių lygybę, kartu patikrinama hipotezė apie dispersijų lygybę. Kaip rasti t kriterijaus p-reikšmę? Prieš tikrinant hipotezę apie vidurkių lygybę, reikia nuspręsti, ar populiacijų dispersijas galima laikyti lygiomis. Grafoje 'Levene's Test for Equality of Variances' pateikiama statistikos stebėta Levene statistikos reikšmė (0,571) ir p-reikšmė (0,455). Kai p-reikšmė ne mažesnė už pasirinktą reikšmingumo lygmenį α , dispersijos statistiškai reikšmingai nesiskiria. Jeigu p-reikšmė mažesnė už α , tai populiacijų dispersijos statistiškai reikšmingai skiriasi. Nagrinėjamojo pavyzdžio atveju $0,455 > 0,05$, todėl dispersijas galima laikyti nesiskiriančiomis.

Išvadoms apie populiacijos vidurkių lygybę skirta likusioji lentelės dalis ('t-test for Equality of Means'). Stulpelyje 't' pateiktos stebėtos statistikos t reikšmės t_{imt} , stulpelyje 'df' – laisvės laipsnių skaičius $n+m-2=38$, stulpelyje 'Sig. (2-tailed)' – p-reikšmės, kiti stulpeliai skirti vidurkių skirtumo įverčiams.

Independent Samples Test

		Levene's Test for Equality of Variances		t-test for Equality of Means						
		F	Sig.	t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	95% Confidence Interval of the Difference	
									Lower	Upper
Pagaminta staklių	Equal variances assumed	,571	,455	2,697	38	,010	2,15	,80	,54	3,76
	Equal variances not assumed			2,697	37,264	,010	2,15	,80	,53	3,77

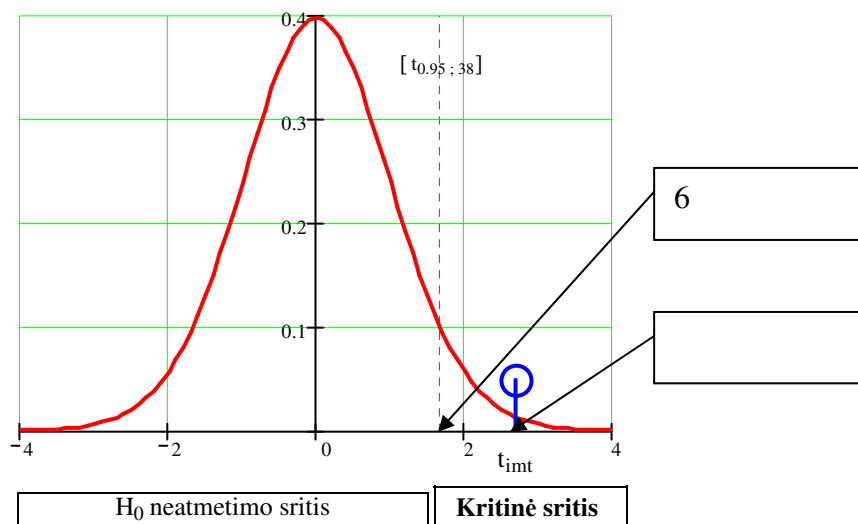
4.3 pav. Hipotezės apie dviejų vidurkių lygybę tikrinimo rezultatai

Visų stulpelių pirmoji eilutė yra lygių dispersijų atvejui ('Equal variances assumed'), antroji – nelygių dispersijų atvejui ('Equal variances not assumed'). Kadangi nagrinėjame pavyzdyje dispersijas galime laikyti lygiomis, tai išvadas formuluojame pagal pirmą eilutę. Stebėta t-kriterijaus reikšmė $t_{imt} = 2,697$, o jos p-reikšmė kriterijui su dvipuse kritine sritimi $p^* = 0,01$ ('Sig. (2-tailed)'). Kadangi mūsų alternatyva $H_a : \mu_x > \mu_y$, skaičiuojame p-reikšmę kriterijui su vienpuse dešine kritine sritimi $p = p^*/2 = 0,01/2 = 0,005$ (jeigu alternatyva yra $H_a : \mu_x > \mu_y$ ir $t_{imt} > 0$, taip p reikšmė $= p^*/2$, jeigu $t_{imt} \leq 0$, tai p-reikšmė $= 1 - p^*/2$). Kadangi gauta p-reikšmė yra mažesnė už 0,05, tai nulinę hipotezę atmetame ir darome **išvadą**, kad vidutinis naktinės pamainos darbo našumas mažesnis už dieninės pamainos darbo našumą.

Išspręsimė tą patį uždavinį nesinaudodami p-reikšme. Mūsų pavyzdyje alternatyva yra $H_a : \mu_X > \mu_Y$ todėl naudosime kriterijų su dešinė kritine sritimi:

$$\mathbf{T}_{H_0} = (-\infty; t_{1-\alpha; n+m-2}], \quad \mathbf{T}_K = (t_{1-\alpha; n+m-2}; +\infty),$$

čia $n + m - 2 = 20 + 20 - 2 = 38$, $\alpha = 0,05$. Apskaičiuojame Stjudento kvantilį $t_{0,95;38} = 1,686$ ir randame hipotezės kritinę sritį $\mathbf{T}_K = (1,686; +\infty)$. Stebėta statistikos reikšmė $t_{imt} = 2,697$ patenka į kritinę sritį, todėl hipotezė H_0 atmetama. Gauti rezultatai grafiškai pavaizduoti paveiksle.



4.4 pav. Hipotezės apie dviejų vidurkių lygybę tikrinimo rezultatai

Skirtumas tarp imties vidurkių lygus $\bar{x} - \bar{y} = 2,15$. Skirtumo tarp populiacijos vidurkių 0,95 pasikliautinis intervalas

$$PI_{0,95}(\mu_X - \mu_Y) = (0,54; 3,76).$$

Su 95% garantija galime teigti, kad dieninė pamaina pagamina vidutiniškai nuo 0,54 iki 3,76 staklių daugiau už naktinę pamainą ($0,54 < \mu_X - \mu_Y < 3,76$).

4.1.2. Mano-Vitnio-Vilkoksono rangų sumų kriterijus dviems nepriklausomoms imtims.

Tarkime, turime dvi imtis. Pažymėkime:

- n_1 – pirmos imties narių skaičių,
 - n_2 – antros imties narių skaičių,
 - R_1 – pirmos imties narių rangų sumą,
 - R_2 – antros imties narių rangų sumą.
- Čia n_1 ir n_2 nebūtinai turi sutapti.

Reikia patikrinti hipotezę:

H_0 : **Kintamųjų X ir Y skirstiniai vienodi,**

H_a : **Kintamųjų X ir Y skirstiniai nėra vienodi.**

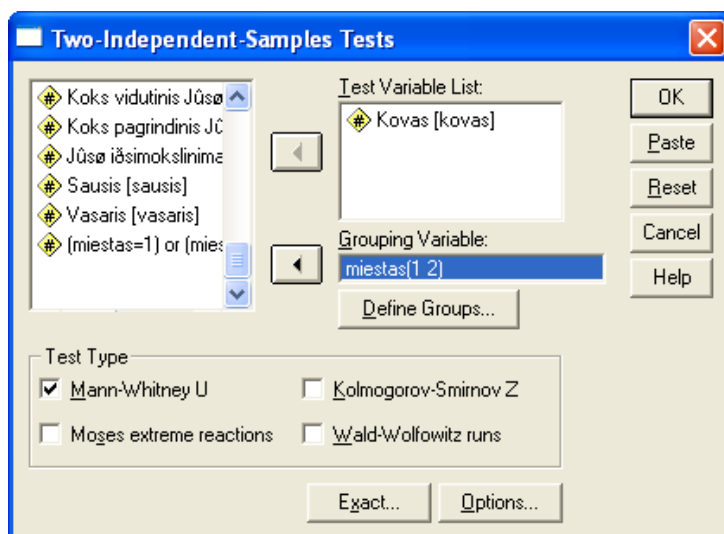
Kai imtys didelės ($n_1 > 20$, $n_2 > 20$, SPSS) hipotezei apie skirstinių sutapimą dviejose populiacijose tikrinti naudojama statistika

$$U_1 = n_1 n_2 + \frac{n_1(n_1 + 1)}{2} - R_1.$$

Jeigu nulinė hipotezė, kad visi $n_1 + n_2$ stebėjimai turi tą patį skirstinį, yra teisinga, tai statistika U turi normalųjį skirstinį su vidurkiu $\mu = \frac{n_1 n_2}{2}$ ir standartiniu nuokrypiu $\sigma = \sqrt{\frac{n_1 n_2 (n_1 + n_2 + 1)}{12}}$. Tuomet

statistika $Z = \frac{U_1 - \mu}{\sigma}$ turi standartinį normalųjį skirstinį su vidurkiu lygiu 0 ir standartiniu nuokrypiu lygiu 1.

Pavyzdys. Patikrinsime hipotezę „Vilniaus ir Kauno abonentų išlaidų telekomunikacijų paslaugoms kovo mėnesį skirstiniai yra vienodi“. Hipotezės tikrinimo rezultatai su SPSS pateikti 4.5 pav. Nulinė hipotezė neatmesta, nes $p = 0,079 > 0,05$.



Mann-Whitney Test

Ranks

	Miestas	N	Mean Rank	Sum of Ranks
Kovas	Vilnius	239	210,44	50295,50
	Kaunas	164	189,70	31110,50
	Total	403		

Test Statistics ^a	
	Kovas
Mann-Whitney U	17580,500
Wilcoxon W	31110,500
Z	-1,756
Asymp. Sig. (2-tailed)	,079

a. Grouping Variable: Miestas

4.5 pav. Hipotezės tikrinimo rezultatai

4.2. Vienfaktorinė dispersinė analizė

4.2.1. Vienfaktorinės dispersinės analizės modeliai

Pilnas randomizuotas vienfaktorinės dispersinės analizės modelis (Completely randomized factorial ANOVA). Tarkime, kad atsitiktinio dydžio Y skirstinys gali priklausyti nuo faktoriaus A , kuris įgyja I skirtingų reikšmių $a_1, a_2, \dots, a_I$. Pagal faktorių A skiriame I nepriklausomų populiacijų. Kiekvienoje populiacijoje matuojame tą patį priklausomą kintamąjį Y (išmatuotą intervalų arba santykių skalėje). Populiacijos, kurioje faktorius A įgyja reikšmę $A = a_i$, kintamąjį žymėsime Y_i , o jo imties vidurkį $\bar{Y}_i$. 4.6 pav. pateiktas pavyzdys duomenų, kurie gauti, kai faktorius A įgyja I reikšmių ir iš I nepriklausomų populiacijų atsitiktinai atrinkta po tris objektus bei išmatuotos kintamojo Y reikšmės.

Faktorius A	Kintamasis Y
a_1	R_1
	R_2
	R_3
a_2	R_1
	R_2
	R_3
...	...
a_I	R_1
	R_2
	R_3

4.6 pav. Pilnai randomizuoto vienfaktorinės dispersinės analizės modelio duomenų pavyzdys.

Blokuotųjų duomenų vienfaktorinė dispersinė analizės modelis. Blokuotieji duomenys dažniausiai gaunami, kai tų pačių objektų požymi matuojame keletą kartų (repeated measures). Blokuotųjų duomenų dispersinė analizė tinka ne tik pakartotiniams matavimams. Svarbu, kad bloką sudarančius duomenis vienyty kokia nors bendra informacija [17-19].

Faktorius B			
b ₁	b ₂	...	b _J
R ₁	R ₁		R ₁
R ₂	R ₂		R ₂
R ₃ $\bar{Y}_1$	R ₃ $\bar{Y}_2$	...	R ₃ $\bar{Y}_J$
R ₄	R ₄		R ₄
R ₅	R ₅		R ₅

4.7 pav. Blokuotų duomenų vienfaktorinės dispersinės analizės modelio duomenų pavyzdys.

Šiame kurse apsiribosime tik pilnai randomizuotais dispersinės analizės modeliais su pastoviais faktoriais.

4.2.2. Pilnas randomizuotas vienfaktorinės dispersinės analizės modelis

Tarkime, kad stebimo atsitiktinio dydžio Y skirstinys priklauso nuo faktoriaus A , kuris yra I skirtinguose lygmenyse. Taigi turime I imčių, kurios kiekvienos didumas yra $n_i, i = \overline{1, I}, n = \sum_{j=1}^I n_i$.

Kiekvieną stebėjimą y_{ij} išskaidome į du dėmenis:

$$y_{ij} = \beta_i + e_{ij}, \quad j = 1, \dots, n_i, i = 1, \dots, I,$$

čia β_i - nežinomi populiacijų vidurkiai $M(Y_i) = \beta_i$, o e_{ij} - nepriklausomi atsitiktiniai dydžiai, kurių skirstinys standartinis normalusis $N(0, \sigma^2)$.

Nulinė vienfaktorinės dispersinės analizės hipotezė teigia, kad visi populiacijų kintamųjų vidurkiai lygūs.

$$H_0 : \beta_1 = \dots = \beta_I,$$

Nulinės hipotezės tikrinimui naudojama Fišerio statistika

$$F = \frac{\overline{SS_A}}{\overline{SS_e}}, \quad \text{čia}$$

$SS_A = \sum_{i=1}^I (\bar{y}_i - \bar{y}_{..})^2 n_i$ - nuokrypių kvadratų suma, apibūdinanti faktoriaus A poveikį stebimo atsitiktinio dydžio Y vidurkiui;

$SS_e = \sum_{i=1}^I \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_i)^2$ - nuokrypių kvadratų suma, apibūdinanti atsitiktinių klaidų faktoriaus E poveikį stebimo atsitiktinio dydžio Y vidurkiui, kurį modelyje nusako atsitiktinis dydis e_{ij} ;

$\bar{y}_{..}$ - imties Y empirinis vidurkis;

$\bar{y}_i$ - imties Y_i empirinis vidurkis;

$SS_p = \sum_{i=1}^I \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_{..})^2$ - bendroji nuokrypių kvadratų suma;

$$SS_p = SS_A + SS_e;$$

$\overline{SS_A} = \frac{1}{I-1} SS_A$, $\overline{SS_e} = \frac{1}{n-I} SS_e$, – nuokrypių kvadratų vidurkiai (imties faktorinė ir liekamoji dispersijos).

Jei H_0 teisinga, tai $F \sim F(\nu_1, \nu_2)$, čia $\nu_1 = I - 1$, $\nu_2 = n - I$.

Hipotezės tikrinimui naudojamas Fišerio kriterijus su dešine kritine sritimi

$$F_{H_0} = (0, F_{1-\alpha; \nu_1; \nu_2}) \quad F_K = [F_{1-\alpha; \nu_1; \nu_2}, \infty).$$

H_0 atmetama, kai stebėtoji kriterijaus statistikos reikšmė patenka į kritinę sritį $F_{imt} \in F_K$, priešingu atveju stebėjimų duomenys nulinei hipotezei neprieštarauja, t.y. galime laikyti, jog faktorius A neveikia stebimojo (tiriamąjo) atsitiktinio dydžio Y vidurkį.

SPSS dispersinės analizės rezultatus pateikia lentelėje [7,9].

4.1 lentelė

Vienfaktorinės dispersinės analizės rezultatų lentelė

Nuokrypių šaltinis	Nuokrypių kvadratų suma	Laisvės laipsniai ν	Nuokrypių kvadratų vidurkis	Stebėta Fišerio statistikos reikšmė F_{imt}	p-reikšmė $\alpha_{imt} = P(F > F_{imt})$
	Sum of Squares	df	Mean Squares	F	Sig.
Between Groups Faktorius A	SS_A	$I - 1$	$\overline{SS_A}$	$\frac{\overline{SS_A}}{\overline{SS_e}}$	α_{imt}
Within Groups Atsitiktinių klaidų faktorius E	SS_e	$n - I$	$\overline{SS_e}$		
Total Visi faktoriai	SS_p	$n - 1$			

Jeigu stebėjimai neprieštarauja nulinei hipotezei H_0 apie vidurkių lygybę, tai analizę galima tuo ir užbaigti. Tokiu atveju visus stebėjimus galime sujungti į vieną didumo n imtį, gautą stebint normalųjį atsitiktinį dydį su vidurkiu β_0 ir dispersija σ_e^2 .

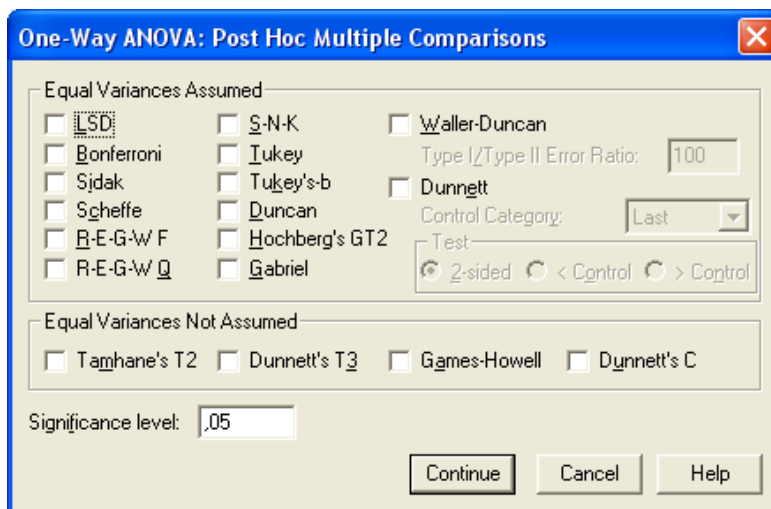
Priešinga išvada retai kada tenkina tyrinėtoją, nes natūraliai kyla klausimas, kaip Y priklauso nuo faktoriaus A lygmenų. Kokie faktoriaus lygiai yra nehomogeniškumo priežastimi? Ar galima faktoriaus A lygmenis apjungti į grupes taip, kad grupių viduje vidurkių skirtumas būtų nereikšmingas? Šio uždavinio sprendimui galima naudoti daugkartinius palyginimus.

Taikant Fišerio kriterijų populiacijų dispersijos turi būti lygios. SPSS hipotezės apie dispersijų lygybę tikrinimui naudojamas Levene kriterijus. Jis yra ne toks jautrus normalumo prielaidos pažeidimui. Jeigu hipotezė apie dispersijų lygybę atmetama, tuomet hipotezės apie vidurkių lygybę tikrinimui, vietoje Fišerio kriterijaus, reikia naudoti Welch arba Brown-Forsythe kriterijus [9,10].

4.2.3. Daugkartiniai lyginimai

Daugkartiniai lyginimų kriterijai skirstomi į **apriorinius** ir **aposteriorinius (post hoc)**. Aprioriniai lyginimai planuojami **prieš dispersinę analizę arba vietoje jos**. Aposterioriniai (post hoc) lyginimai vykdomi po to, kai jau žinomi analizės rezultatai (hipotezės apie kelių vidurkių lygybę tikrinimo rezultatai). Yra daug skirtingų daugkartinio lyginimo kriterijų. Dauguma kriterijų (tik su skirtingais reikšmingumo lygmenimis) galima naudoti ir kaip apriorinius ir kaip aposteriorinius daugkartinius lyginimus.

Aposterioriniai (Post Hoc) kriterijai. Yra daug skirtingų aposteriorinių (Post Hoc) daugkartinių lyginimų kriterijų. Vien SPSS pateikia 18 skirtingų kriterijų: LSD (Least Significant Difference - mažiausiai reikšmingo skirtumo), Benferonio (Bonferroni), Sidak'o, Šefe (Scheffe) R-E-G-W F (Ryan-Einot-Gabriel-Welsch F kriterijus), R-E-G-W Q (Ryan-Einot-Gabriel-Welsch Q kriterijus), S-N-K (Student-Newman-Keuls, Tjukio (Tukey), Tjukio b (Tukey b), Duncan, Hochberg GT2, Gabriel, Waller-Duncan, Dunnett kriterijai taikomi kai lyginamų populiacijų dispersijos yra lygios (Equal Variances Assumed). Tamhane T2, Dunnett T3, Games-Howell, Dunnett C. kriterijai, taikomi kai lyginamų populiacijų dispersijos nėra lygios (Equal Variances Not Assumed).



4.8 pav. SPSS aposterioriniai (Post Hoc) daugkartinio lyginimo kriterijai.

Kiekvienas iš jų turi pranašumų ir trūkumų. Vieni kriterijai taikomi, kai visų lyginamų imčių didumai vienodi, kitus galima taikyti kai imčių didumai nevienodi. Vieni kriterijai taikomi, kai visų populiacijų dispersijos yra lygios, kitus galima taikyti, kai dispersijos nelygios. Vieni kriterijai dažniau randa statistiškai reikšmingus skirtumus, kiti rečiau (konservatyvesni kriterijai) ir t.t. Aposteriorinio (Post Hoc) daugkartinio lyginimo kriterijaus parinkimas nėra lengvas uždavinys ir reikalauja gilesnių šių kriterijų studijų [10, 11, 17, 21]. Čia pateiksime tik kelis kriterijus.

LSD (Least Significant Difference) – Fišerio mažiausiai reikšmingo skirtumo kriterijus (1945 m.). Pagal Fišerio LSD kriterijų visos imčių poros lyginamos naudojant t-kriterijų. LSD yra liberaliausias aposteriorinis (Post Hoc) daugkartinio lyginimo kriterijus, t.y. jis dažniausiai randa statistiškai reikšmingus vidurkių skirtumus. Dėl šios priežasties nerekomenduojama jį taikyti praktikoje [10].

Benferonio (Bonferroni) kriterijus. Pagal Benferonio kriterijų pasirenkamas eksperimento reikšmingumo lygmuo α_E (t.y. tikimybė lyginant visas įmanomas poras nors kartą neteisingai nustatyti statistiškai reikšmingą dviejų vidurkių skirtumą) ir visos vidurkių poros lyginamos taikant

Stjudento t-kriterijų, esant reikšmingumo lygmeniui $\alpha = \alpha_E / C$, čia $C = I(I - 1)/2$. Bonferonio kriterijus **netaikytinas, kai lyginamų populiacijų vidurkių yra daug**, nes labai sumažėja α ir labai retai fiksuojamas statistiškai reikšmingas vidurkių skirtumas, nors tikrieji populiacijų vidurkiai ir skiriasi (t. y. labai išauga antrosios rūšies klaidos tikimybė).[].

Tjukio (Tukey) kriterijus. Tukey'o HSD (*Honestly Significant Difference* — ganėtinai statistiškai reikšmingo skirtumo) kriterijus — vienas dažniausiai taikomų kriterijų. Jis yra gera alternatyva Bonferoni kriterijui, kai imčių daug. Tjukio kriterijus yra labai konservatyvus, t.y. labiausiai nelinkęs atmesti nulinės hipotezės — vidurkių skirtumus pripažinti statistiškai reikšmingais. Be to SPSS Tjukio kriterijumi sudaro homogeniškumo grupes (į homogeniškumo grupės įrašomos populiacijos kurių vidurkių skirtumas yra statistiškai nereikšmingas). Jeigu iš populiacijų atrinktų imčių didumai skiriasi, tai gali skirtis ir homogeniškumo grupių bei HSD kriterijaus rezultatai, nes homogeniškumo grupėms sudaryti tuomet naudojami harmoniniai vidurkiai. Tjukio HSD kriterijaus galia prie prie didesnio imčių skaičiaus yra didesnė lyginant su Bonferoni kriterijaus galia ir atvirkščiai Bonferoni kriterijaus galia didesnė esant mažam imčių skaičiui.

4.2.4. Vienfaktorinės dispersinės analizės pavyzdys

Vienfaktorinė dispersinė analizė (One-way ANOVA) taikoma tada, kai populiacijos viena nuo kitos skiriamos tik pagal vieną požymį. Tarkime, kad trims atsitiktinai atrinktoms darbuotojų grupėms (po 20 žmonių kiekvienoje) buvo dėstomas tas pats profesijos dalykas. Pirmą grupę buvo mokoma intensyviai (1 dienos kursai), antrą grupę kursai vyko dvi dienas, o trečią grupę – tris dienas. Visų trijų kursų mokymo medžiagos apimtis ir valandų skaičius buvo tas pats. Po kursų visi darbuotojai laikė egzaminą, kuris buvo vertinamas 100 balų skalėje. Ar kursų dienų skaičius turi įtakos egzamino rezultatams?

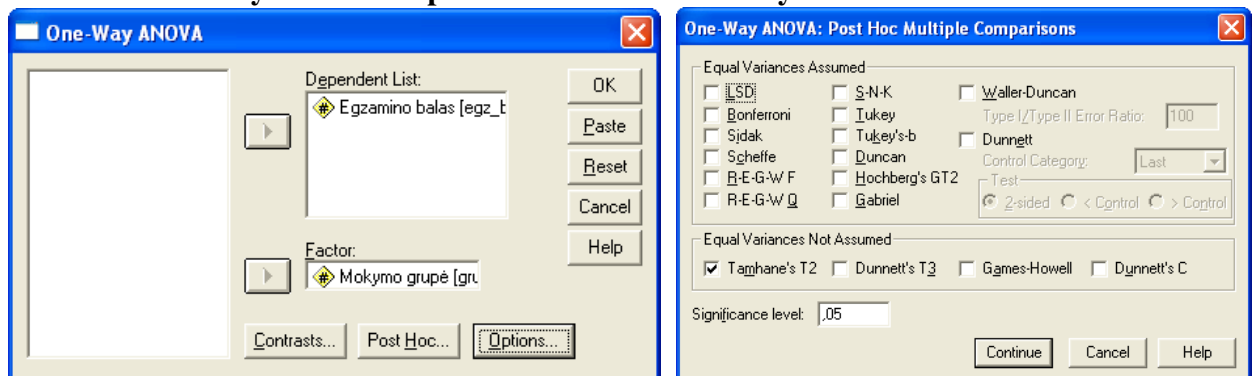
Duoti du kintamieji:

- **egz_balas** – egzamino balas. Kintamojo (Y) reikšmės išmatuotos intervalų skalėje.
- **grupe** – mokymo grupė (1-vienos dienos kursai, 2-dviejų dienų kursai, 3- trijų dienų kursai). Kintamojo (faktoriaus A) reikšmės išmatuotos vardų skalėje.

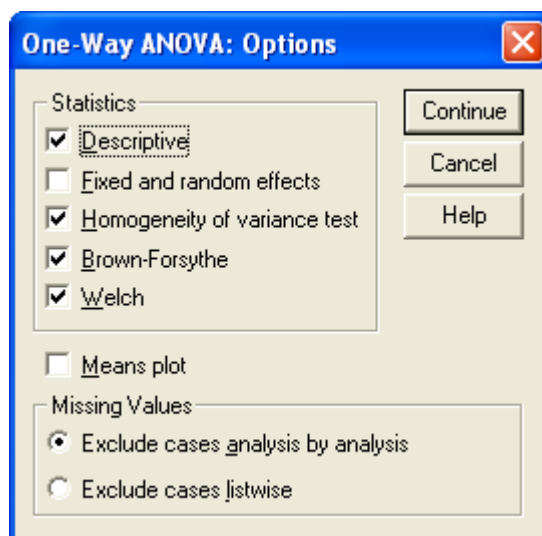
Uždavinio sprendimui taikysime vienfaktorinės dispersinės analizės metodą. Apskaičiuosime imties skaitines charakteristikas, patikrinsime hipotezę apie dispersijų lygybę, patikrinsime hipotezę apie vidurkių lygybę. Jeigu stebėjimai prieštarauja nulinėi hipotezei apie vidurkių lygybę, tai panaudojė daugiakartinius palyginimus išsiaiškinsime tarp kokių grupių vidurkių skirtumas yra reikšmingas?

4.9-4.10 pav. pateikti SPSS vienfaktorinės dispersinės analizės dialogo langai, kuriuose pateikti duomenys ir nurodyti kriterijai aukščiau išvardintų uždavinių sprendimui.

SPSS meniu: **Analyze ---> Compare Means ---> One-Way ANOVA.**



4.9 pav. SPSS vienfaktorinės dispersinės analizės langai



4.10 pav. SPSS vienfaktorinės dispersinės analizės langas (Options)

Descriptives

Egzamino balas								
	N	Mean	Std. Deviation	Std. Error	95% Confidence Interval for Mean		Minimum	Maximum
					Lower Bound	Upper Bound		
1	20	63,5798	13,50858	3,02061	57,2576	69,9020	32,68	86,66
2	20	73,5677	10,60901	2,37225	68,6025	78,5328	47,56	89,65
3	20	79,2792	4,40754	,98556	77,2165	81,3420	71,77	89,69
Total	60	72,1422	12,00312	1,54960	69,0415	75,2430	32,68	89,69

4.11 pav. Aprašomoji statistika

Išvada. Trečios grupės klausytojų vidutinis egzamino balas yra aukščiausias – 79,14 ir egzamino balų sklaida yra mažiausia (standartinis nuokrypis 4,4 balo). Blogiausiai egzaminą išlaikė pirma grupė (egzamino balų vidurkis 63,58). Šios grupės didžiausia yra ir egzamino balų sklaida (standartinis nuokrypis 13,5).

Test of Homogeneity of Variances Egzamino balas

Levene Statistic	df1	df2	Sig.
4,637	2	57	,014

4.12 pav. Hipotezės apie dispersijų lygybę tikrinimo rezultatai

Išvada. Hipotezė apie dispersijų lygybę atmesta ($p\text{-reikšmė}=0,014>0,05$), todėl hipotezės apie vidurkių lygybę tikrinimui, vietoje Fišerio kriterijaus, reikia naudoti Welch arba Brown-Forsythe kriterijus.

ANOVA

Egzamino balas

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	2525,691	2	1262,846	12,048	,000
Within Groups	5974,724	57	104,820		
Total	8500,415	59			

Robust Tests of Equality of Means

Egzamino balas

	Statistic(a)	df1	df2	Sig.
Welch	13,278	2	30,962	,000
Brown-Forsythe	12,048	2	40,540	,000

a. Asymptotically F distributed.

4.13 pav. Hipotezės apie vidurkių lygybę tikrinimo rezultatai

Išvada. Hipotezė apie vidurkių lygybę atmesta ($p\text{-reikšmė}=0,000<0,05$), t.y. bent dviejų grupių vidurkiai statistiškai reikšmingai skiriasi. (Pastaba. Kadangi hipotezė apie dispersijų lygybę atmesta, hipotezės apie vidurkių lygybę tikrinimui, vietoje Fišerio kriterijaus, turime naudoti Welch arba Brown-Forsythe kriterijus).

Multiple Comparisons

Dependent Variable: Egzamino balas

Tamhane

(I) Mokymo grupė	(J) Mokymo grupė	Mean Difference (I-J)	Std. Error	Sig.	95% Confidence Interval	
					Lower Bound	Upper Bound
1	2	-9,98789*	3,84079	,040	-19,6053	-,3705
	3	-15,69947*	3,17733	,000	-23,8792	-7,5198
2	1	9,98789*	3,84079	,040	,3705	19,6053
	3	-5,71158	2,56883	,102	-12,2771	,8539
3	1	15,69947*	3,17733	,000	7,5198	23,8792
	2	5,71158	2,56883	,102	-,8539	12,2771

*. The mean difference is significant at the .05 level.

4.14 pav. Daugiakartinių lyginimų rezultatai (Tamhane kriterijus)

4.14 pav. (Multiple Comparisons-daugkartiniai lyginimai) pateikti Tamhane kriterijaus taikymo rezultatai (naudojame šį kriterijų, nes grupių dispersijos nelygios). Lentelėje statistiškai reikšmingi vidurkių skirtumai, t.y. tie, kurių $p\text{-reikšmė}$ (Sig.) mažesnė už 0,05 lentelėje pažymėti žvaigždutėmis. Statistiškai reikšmingo skirtumo nėra tik tarp 2 ir 3 grupės vidurkių ($p\text{-reikšmė}=0,102>0,05$), t.y. dviejų ir trijų dienų kursų klausytojų egzamino balų vidurkiai skiriasi nereikšmingai. Skirtumas tarp trečios ir pirmos grupių egzaminų vidurkių yra reikšmingas ($p\text{-reikšmė}=0,000<0,05$). Trečios ir pirmos grupės klausytojų egzamino balų vidurkis skiriasi 15,7 balo. Skirtumo tarp populiacijos vidurkių 0,95 pasikliautinis intervalas

$$PI_{0,95}(\beta_3 - \beta_1) = (7,52; 23,88).$$

Su 95% garantija galime teigti, kad populiacijoje trijų dienų kursų klausytojų egzamino balų vidurkis gali būti didesnis nuo 7.52 iki 23,88 balo už vienos dienos kursų klausytojų egzamino vidurkį.

4.3. Kruskalo-Voliso kriterijus k-nepriklausomoms imtims

Tai apibendrintas Manno-Vitnio-Vilkoksono kriterijaus atvejis didesniai negu 2 imčių skaičiui. Reikia patikrinti hipotezę:

H_0 : Kintamųjų skirstiniai vienodi,

H_a : Kintamųjų skirstiniai nėra vienodi.

Pažymėkime:

n_j – j-osios imties narių skaičių,

R_j – j-osios imties narių rangų suma,

k – imčių skaičių,

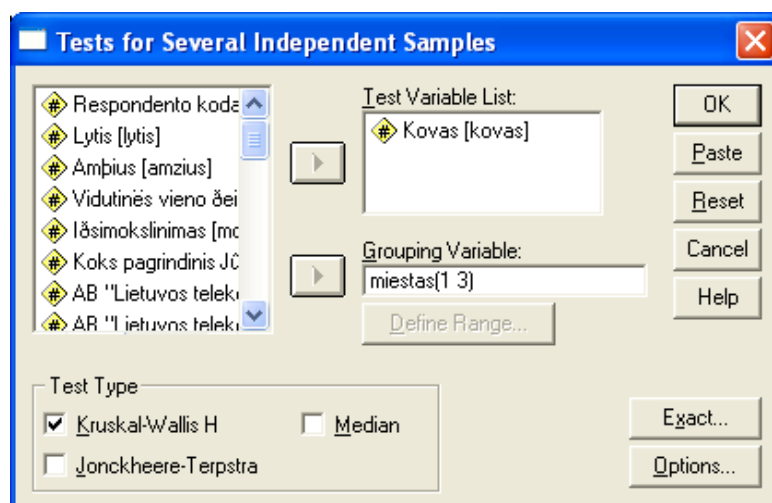
$n = n_1 + n_2 + \dots + n_k$ – bendrą visų imčių stebėjimų skaičių.

Statistika, naudojama tikrinant hipotezę apie tai, kad visos imtys yra iš visumų, kuriose tiriamas kintamasis turi tą patį vidurkį (skirstinį). apskaičiuojam taip:

$$H = \frac{12}{n(n+1)} \sum_{j=1}^k \frac{R_j^2}{n_j} - 3 \cdot (n+1).$$

Kai hipotezė yra teisinga, statistika H turi apytiksliai χ^2 skirstinį su $k-1$ laisvės laipsnių. Tačiau norint taikyti Kruskalo-Voliso kriterijų, reikia, kad visų imčių didumas būtų ne mažesnis už 5.

Pavyzdys. Patikrinsime hipotezę „Vilniaus, Kauno ir Klaipėdos abonentų išlaidų telekomunikacijų paslaugoms kovo mėnesį. skirstiniai yra vienodi“. Hipotezės tikrinimo rezultatai su SPSS pateikti 4.15 pav. Nulinė hipotezė atmesta, nes $p=0,000 < 0,05$.



Ranks

	Miestas	N	Mean Rank
Kovas	Vilnius	239	268,14
	Kaunas	164	243,33
	Klaipėda	83	172,89
	Total	486	

Test Statistics^{a,b}

	Kovas
Chi-Square	28,341
df	2
Asymp. Sig.	,000

a. Kruskal Wallis Test

b. Grouping Variable: Miestas

4.15 pav. Hipotezės apie trijų skirstinių vienodumą tikrinimo rezultatai

5. POŽYMIŲ PRIKLAUSOMUMO TYRIMAS. KORELIACINĖ ANALIZĖ

Dažnai reikia atsakyti į klausimą, ar stebimi kintamieji yra priklausomi, ar nepriklausomi? Pavyzdžiui:

- ar yra priklausomybė tarp politinių pažiūrų ir amžiaus;
- ar nusikalstamumo lygis priklauso nuo bedarbystės lygio;
- ar atlyginimo dydis priklauso nuo išsimokslinimo lygio;
- ar mokymosi rezultatai priklauso nuo mokymo metodikos;
- ar vaikų agresyvus elgesys priklauso nuo jų pomėgio žiūrėti smurtines TV laidas;
- ar užsienio politikos kursas priklauso nuo to, kokia partija yra valdžioje;
- ar dviejų ekspertų vertinimai yra suderinti;
- ar dvi žmonių grupės sutaria vienu ar kitu klausimu.

Kaip patikrinti hipotezę, kad kintamieji X ir Y yra nepriklausomi? Jeigu jie yra priklausomi, tai kaip įvertinti šio ryšio stiprumą?

Kintamųjų priklausomybės matas yra koreliacijos koeficientas. Tikrindami hipotezę apie populiacijos koreliacijos koeficiento lygybę nuliui, atsakome į klausimą apie kintamųjų priklausomybę (tam tikra prasme – tiesinę, monotoniškumą, suderinamumą ir t.t.).

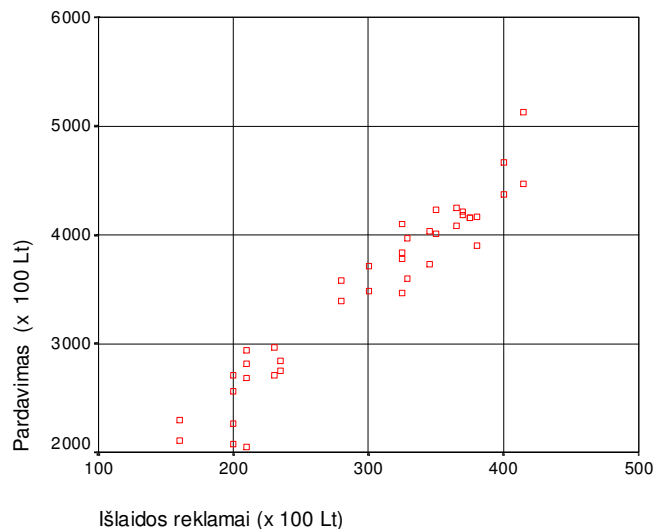
Iš koreliacijos koeficiento negalima nustatyti koreliacijos priežasties. Du kintamieji X ir Y gali stipriai koreliuoti dėl trijų priežasčių:

- kintamasis X daro poveikį kintamajam Y;
- kintamasis Y daro poveikį kintamajam X;
- abu kintamieji X ir Y yra veikiami trečio kintamojo.

Todėl koreliacinės analizės metu nustatytas ryšys negali būti interpretuojamas kaip priežastingumas, o tik kaip **asociacijos arba ryšio matas**.

5.1. Pirsono koreliacijos koeficientas

Pirsono koreliacijos koeficientas įvertina tiesinio ryšio stiprumą. Jis gali būti naudojamas, kai stebimų atsitiktinių dydžių X ir Y skirstiniai yra normalieji (reikšmės yra išmatuotos intervalų arba santykių skalėje). Tarkime, kad mus domina ryšys tarp prekių pardavimo ir išlaidų reklamai.



5.1 pav. Išlaidų reklamai ir pardavimų duomenų sklaidos diagrama (SPSS – Scatter plot)

Taškų sklaidos diagrama (5.1 pav.) rodo, kad didėjant išlaidoms reklamai tiesiškai didėja ir pardavimų apimtys. Taigi, taškų sklaidos diagramos parodo ryšio tarp dviejų kintamųjų pobūdį. Jos yra ypač naudingos, nes atskleidžia ir netiesinį ryšį, kuris gali būti neįvertintas skaičiuojant tiesinę koreliaciją. Tačiau dažnai reikia nustatyti ne tik ryšio pobūdį, bet ir jo stiprumą.

Dviejų atsitiktinių dydžių X ir Y tiesinio ryšio stiprumą įvertina **Pirsono koreliacijos koeficientas**. Jo taškinis įvertis (imties koreliacijos koeficientas)

$$\hat{\rho} = r = \frac{k_{xy}}{s_x s_y} = \frac{\overline{xy} - \bar{x} \cdot \bar{y}}{\sqrt{x^2 - (\bar{x})^2} \sqrt{y^2 - (\bar{y})^2}}.$$

Tiesinis ryšys bus stipresnis, kuo $|r|$ reikšmė bus arčiau 1. Jei $r > 0$, tai didėjant vieno atsitiktinio dydžio reikšmėms, kito reikšmės tiesiškai didėja. Jei $r < 0$, tai didėjant vieno atsitiktinio dydžio reikšmėms, kito reikšmės tiesiškai mažėja, r neparodo netiesinės priklausomybės. Kuo didesnė imtis, tuo r yra arčiau nežinomo populiacijos koreliacijos koeficiento ρ .

Hipotezės apie koreliacijos koeficientą tikrinimas. Kokį koreliacijos koeficiento ρ taškinio įverčio r didumą galime laikyti statistiškai reikšmingu? Prie kokio r didumo mes galime teigti, kad tarp stebėtų atsitiktinių dydžių X ir Y yra reikšmingas tiesinis ryšys? Empirinio koreliacijos koeficiento kaip ir visų taškinio įverčių tikslumas ir patikimumas mažėja, kai mažėja imties didumas.

Tarkime, kad stebime du atsitiktinius dydžius X ir Y, kurių koreliacijos koeficientas ρ yra nežinomas. Norint atsakyti į klausimą, ar šie dydžiai yra nekoreliuoti, tikrinama hipotezė:

$$H_0: \rho = 0;$$

$$H_a: \rho \neq 0.$$

Jeigu H_0 atmesime, kai tarp X ir Y yra statistiškai reikšmingas ryšys, kurio stiprumas gali svyruoti nuo labai silpno (ρ artimas nuliui) iki funkcinio ryšio ($|r|=1$).

Pavyzdys. Pirsono koreliacijos koeficientas įvertinantis tiesinio ryšio stiprumą tarp automobilio svorio ir benzino sąnaudų apskaičiuotas naudojant paketą SPSS. Imties didumas $n=392$.

SPSS meniu: Analyze-->Correlate-->Bivariate

		Svoris (kg)	Benzino sąnaudos 100 km (litrais)
Svoris (kg)	Pearson Correlation	1,000	,885
	Sig. (2-tailed)	,	,000
	N	392	392
Benzino sąnaudos 100 km (litrais)	Pearson Correlation	,885	1,000
	Sig. (2-tailed)	,000	,
	N	392	392

** Correlation is significant at the 0.01 level (2-tailed).

5.2 pav. Koreliacinės analizės lentelė (Pirsono koreliacijos koeficientas, SPSS)

Pateiktame pavyzdyje imties koreliacijos koeficientas $r=0.885$. Tarp automobilio svorio ir sunaudojamo benzino kiekio yra stiprus tiesinis ryšys. Nulinė hipotezė: „Populiacijos koreliacijos koeficientas lygus nuliui“, atmeta, nes p -reikšmė (Sig. 2 tailed) $=0,000 < 0,05$.

Išvada: Tarp automobilio svorio ir sunaudojamo benzino kiekio yra stiprus tiesinis ryšys ($r=0,885$).

5.2. Spirmeno ranginės koreliacijos koeficientas

Spirmeno ranginės koreliacijos koeficientas naudojamas ryšio stiprumo įvertinimui tarp kintamųjų išmatuotų santykių, intervalų ir tvarkos skalėse. Spręsdami praktinius uždavinius retai žinome stebimų atsitiktinių dydžių skirstinio tipą, arba žinome, kad jų skirstiniai nėra normalieji. Tokiu atveju reikia naudotis koeficientais nesusijusiais su skirstiniais.

Spirmeno ranginės koreliacijos koeficientas ρ_s apibūdina ryšio tarp X ir Y stiprumą **monotoniškumo prasme**, t.y. X didėjant, Y monotoniškai didėja (nebūtinai tiesiškai) arba mažėja. Hipotezės apie Spirmeno ranginės koreliacijos koeficiento reikšmingumą tikrinimui naudojama Stjudento statistika.

$$H_0: \rho_s = 0,$$

$$H_a: \rho_s \neq 0.$$

Pavyzdys. Įvertinsime ryšio stiprumą monotoniškumo prasme tarp atsakymų į apklausos anketos klausimus apie taisytinas sritis įmonės įvaizdžio gerinimui: **k1** – “įmonės bendravimas su plačiąja visuomene”; **k2** – “labdaros teikimas” ir **k3** – “pagrindinių paslaugų kokybė”.

SPSS meniu: Analyze→Correlate→Bivariate...

Correlations					
			Taisytinios sritys įvaizdžio gerinimui (Įmonės bendravimas su plačiąja visuomene)	Taisytinios sritys įvaizdžio gerinimui (Labdaros teikimas)	Taisytinios sritys įvaizdžio gerinimui (Pagrindinių paslaugų kokybė)
Spearman's rho	Taisytinios sritys įvaizdžio gerinimui (Įmonės bendravimas su plačiąja visuomene)	Correlation Coefficient	1,000	,405**	,217**
		Sig. (2-tailed)	,	,000	,000
		N	1010	1010	1010
	Taisytinios sritys įvaizdžio gerinimui (Labdaros teikimas)	Correlation Coefficient	,405**	1,000	,125**
		Sig. (2-tailed)	,000	,	,000
		N	1010	1010	1010
	Taisytinios sritys įvaizdžio gerinimui (Pagrindinių paslaugų kokybė)	Correlation Coefficient	,217**	,125**	1,000
		Sig. (2-tailed)	,000	,000	,
		N	1010	1010	1010

** . Correlation is significant at the .01 level (2-tailed).

5.3. pav. Ranginės koreliacijos koeficientai

Išvados. Į visus tris klausimus atsakė **1010** respondentų. Spirmeno ranginės koreliacijos koeficientas įvertinantis ryšio stiprumą tarp atsakymų į apklausos anketos klausimus **k1** – “įmonės bendravimas su plačiąja visuomene” ir **k2** – “labdaros teikimas” **$r_s = 0,405$** . Nulinė hipotezė: „Spirmeno ranginės koreliacijos koeficientas populiacijoje lygus nuliui“, atmesta (p -reikšmė $= 0,000 < 0,05$). Tarp klientų požiūrio į įmonės bendravimą su plačiąja visuomene ir labdaros teikimą yra vidutinio stiprumo ryšys monotoniškumo prasme. Vieno svarbai didėjant, kito taip pat didėja.

5.3. Požymių priklausomumo lentelės

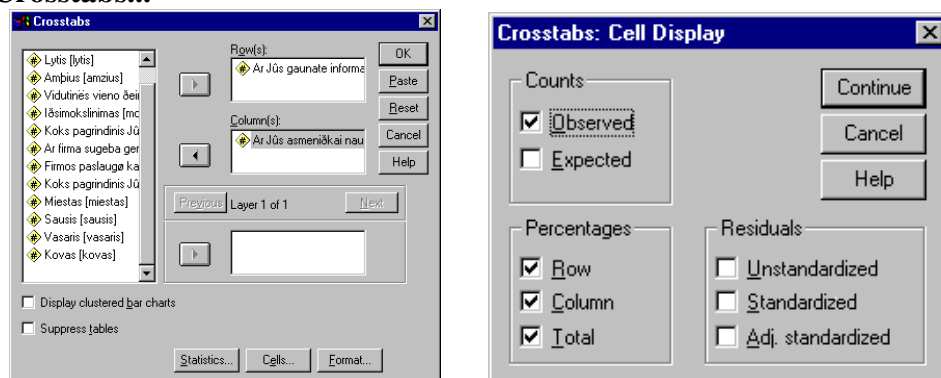
Tarkime, kad stebime atsitiktinių dydžių porą (X, Y) . Atsitiktinis dydis X įgyja I skirtingų reikšmių o atsitiktinis dydis Y įgyja J skirtingų reikšmių. Pažymėkime O_{ij} porų $(x_i; y_j)$ dažnį. Stebėjimo rezultatus galima pateikti požymių priklausomumo lentelę (2.1. lentelė).

2.1 lentelė. Požymių priklausomumo lentelė

$X \backslash Y$	y_1	y_2	...	y_J	Σ
------------------	-------	-------	-----	-------	----------

x_1	o_{11}	o_{12}	...	o_{1J}	$o_{1\bullet}$
x_2	o_{21}	o_{22}	...	o_{2J}	$o_{2\bullet}$
$\vdots$	$\vdots$	$\vdots$		$\vdots$	$\vdots$
x_I	o_{I1}	o_{I2}	...	o_{IJ}	$o_{I\bullet}$
Σ	$o_{\bullet 1}$	$o_{\bullet 2}$	...	$o_{\bullet J}$	n

SPSS požymių priklausomumo lentelių analizei skirtas SPSS meniu: **Analyze→Descriptive Statistics→Crosstabs...**



5.4 pav. SPSS požymių priklausomumo lentelių analizės dialogo langai

Be stebėtų (**observed**) dažnių, požymių priklausomumo lentelės kiekviename langelyje galima nurodyti tikėtinus (**expected**) dažnius, eilučių (**row**), stulpelių (**column**) ir visumos (**total**) procentus. Sudarysime atsakymų į klausimus (dvireikšmių kintamųjų): “**Ar Jūs gaunate informacijos apie firmos teikiamą paslaugą?**” ir “**Ar Jūs asmeniškai naudojātės firmos teikiama paslauga?**” dažnių lentelę.

			Ar Jūs asmeniškai naudojātės firmos teikiama paslauga?		Total
			Taip	Ne	
Ar Jūs gaunate informacijos apie firmos teikiamas paslaugas?	Taip	Count	94	417	511
		% within Ar Jūs gaunate informacijos apie firmos teikiamas paslaugas?	18,4%	81,6%	100,0%
		% within Ar Jūs asmeniškai naudojātės firmos teikiama paslauga?	63,5%	46,0%	48,4%
		% of Total	8,9%	39,5%	48,4%
	Ne	Count	54	490	544
		% within Ar Jūs gaunate informacijos apie firmos teikiamas paslaugas?	9,9%	90,1%	100,0%
		% within Ar Jūs asmeniškai naudojātės firmos teikiama paslauga?	36,5%	54,0%	51,6%
		% of Total	5,1%	46,4%	51,6%
Total	Count	148	907	1055	
	% within Ar Jūs gaunate informacijos apie firmos teikiamas paslaugas?	14,0%	86,0%	100,0%	
	% within Ar Jūs asmeniškai naudojātės firmos teikiama paslauga?	100,0%	100,0%	100,0%	
	% of Total	14,0%	86,0%	100,0%	

Pavyzdys. Tarkime, kad Jūs esate pardavimų analitikas ir norite išsiaiškinti ar yra ryšys tarp prekės A ir prekės B pirkimo? Atsitiktinai atrinktų 286 pirkėjų apklausos analizės rezultatai pateikti lentelėse (5.6 pav.).

Ar perkate prekę A ? * Ar perkate prekę B ? Crosstabulation

			Ar perkate prekę B ?		Total
			Ne	Taip	
Ar perkate prekę A ?	Ne	Count	84	32	116
		Expected Count	53,5	62,5	116,0
	Taip	Count	48	122	170
		Expected Count	78,5	91,5	170,0
Total	Count		132	154	286
	Expected Count		132,0	154,0	286,0

Chi-Square Tests

	Value	df	Asymp. Sig. (2-sided)	Exact Sig. (2-sided)	Exact Sig. (1-sided)
Pearson Chi-Square	54,150 ^b	1	,000		
Continuity Correction ^a	52,387	1	,000		
Likelihood Ratio	55,783	1	,000		
Fisher's Exact Test				,000	,000
Linear-by-Linear Association	53,961	1	,000		
N of Valid Cases	286				

a. Computed only for a 2x2 table

b. 0 cells (,0%) have expected count less than 5. The minimum expected count is 53,54.

Symmetric Measures

		Value	Approx. Sig.
Nominal by Nominal	Phi	,435	,000
	Cramer's V	,435	,000
N of Valid Cases		286	

a. Not assuming the null hypothesis.

b. Using the asymptotic standard error assuming the null hypothesis.

5.6. pav. Požymių priklausomumo analizės rezultatai.

Nulinė hipotezė atmesta, nes stebėta χ^2 statistikos su Jeito pataisa (χ^2 test with Yates' Continuity

Correction) reikšmė yra $\chi^2_{\text{Yates}} = \sum_{i=1}^I \sum_{j=1}^J \frac{(|O_{ij} - E_{ij}| - 0,5)^2}{E_{ij}} = 52,39$, o p-reikšmė = 0,000 < 0,05.

Išvada. Tarp kintamųjų X ir Y yra statistiškai reikšmingas vidutinio stiprumo ryšys (Kramerio C = 0,435). Ryšio stiprumo matavimą nagrinėsime vėliau.

5.5. Ryšio stiprumo matai tvarkos skalėje

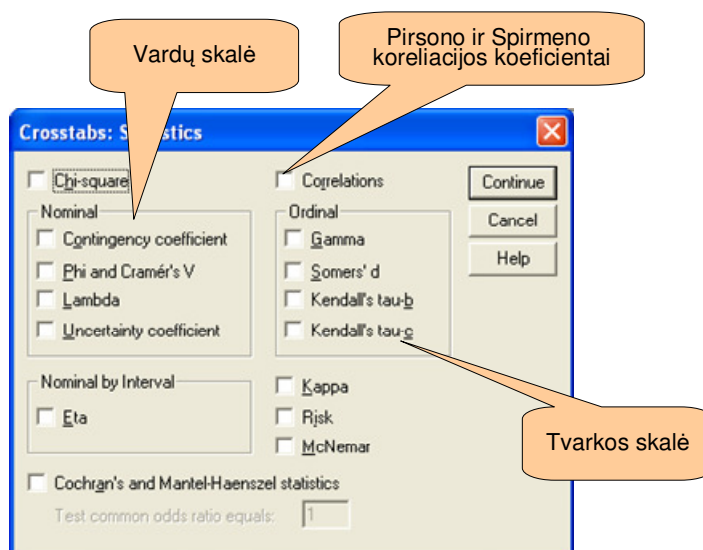
Vien konstatavimas, kad požymiai yra priklausomi, nėra pakankamas. Papildomai reikia įvertinti kintamųjų ryšio stiprumą. Šiame skyrelyje pateiksime keletą ryšio matų, kai stebimi kintamieji matuojami pagal tvarkos skalę. Visiems čia aprašytiems ryšio matams galioja ta pati taisyklė – kuo jie absoliučiau didesni, tuo požymių priklausomybė didesnė; kuo arčiau nulio – tuo priklausomybė silpnesnė. Jų reikšmės gali būti tarp -1 ir 1 arba tarp 0 ir 1.

Kendalo ranginės koreliacijos koeficientai naudojami ryšio stiprumo suderinamumo prasme įvertinimui, kai kintamieji išmatuoti santykių, intervalų ir tvarkos skalėse. Tarpusavyje lyginamos visos galimos kintamųjų (X,Y) stebėjimų reikšmių rangų poros ir tikrinama ar jos yra **suderintos** (jų skaičių žymėsime P), ar **nesuderintos** (jų skaičių žymėsime Q), ar **surištos** (jų skaičius žymėsime T). Kai imtyje nėra surištų porų naudojamas Kendalo ranginės koreliacijos koeficientas:

$$\tau_a = \frac{P - Q}{\frac{1}{2}n(n-1)}.$$

Kai imtyje yra surištų porų naudojamas Kendalo ranginės koreliacijos koeficientas τ_b :

$$\tau_b = \frac{P - Q}{\sqrt{P + Q + T_x} \sqrt{P + Q + T_y}}, \quad -1 \leq \tau_b \leq 1.$$



5.7 pav. Požymių priklausomumo lentelių koreliacijos koeficientai

Pavyzdžiui, du ekspertai 10 balų skalėje įvertino 10 projektų. Vertinimo rezultatai pateikti 2.8 pav. lentelėje (**a**-variante surištų porų nėra (kintamieji **eksp1a** ir **eksp2a**); **b**-variante yra surištų porų (kintamieji **eksp1b** ir **eksp2b**):

	eksp1a	eksp2a	eksp1b	eksp2b
1	1	2	1,0	1,0
2	2	3	2,5	2,0
3	3	1	2,5	4,5
4	4	4	4,5	4,5
5	5	6	4,5	4,5
6	6	5	6,5	4,5
7	7	9	6,5	8,0
8	8	7	8,0	8,0
9	9	8	9,5	8,0
10	10	10	9,5	10,0

5.8 pav. Vertinimo rezultatai

Vertinimo a-variante yra 40 (P=40) suderintų porų ir 5 (Q=5). nesuderintos poros. Pateiksime nesuderintų porų sąrašą: (1;2) ir (3;1); (2;3) ir (3;1); (5;6) ir (6;5); (7;9) ir (8;7); (7;9) ir (9;8). Apskaičiuosime Kendalo ranginės koreliacijos koeficientus su SPSS.

SPSS meniu: Analyze→Correlate→Bivariate...

Correlations

			Pirmas ekspertas (a)	Antras ekspertas (a)
Kendall's tau_b	Pirmas ekspertas (a)	Correlation Coefficient	1,000	,778**
		Sig. (2-tailed)	,	,002
		N	10	10
	Antras ekspertas (a)	Correlation Coefficient	,778**	1,000
		Sig. (2-tailed)	,002	,
		N	10	10
Spearman's rho	Pirmas ekspertas (a)	Correlation Coefficient	1,000	,915**
		Sig. (2-tailed)	,	,000
		N	10	10
	Antras ekspertas (a)	Correlation Coefficient	,915**	1,000
		Sig. (2-tailed)	,000	,
		N	10	10

** . Correlation is significant at the .01 level (2-tailed).

5.9. pav. Ranginės koreliacijos koeficientai

Išvada. Tarp ekspertų vertinimų yra stiprus ryšys suderinamumo prasme (Kendalo $\tau_b = 0.778$) Nulinė hipotezė apie koreliacijos koeficiento lygybę nuliui populiacijoje atmesta (p-reikšmė=0,000 <0,05). Panašūs rezultatai gaunami analizuojant ekspertų vertinimo **b**-variantą.

Jeigu **IxJ** požymių priklausomumo lentelėje I=J tai ryšio stiprumui įvertinti naudojamas Kendalo koeficientas τ_b , jei $I \neq J$, tai rekomenduojama naudoti Kendalo τ_c :

$$\tau_c = \frac{m \cdot (P - Q)}{n^2 \cdot (m - 1)}, \text{ čia } m = \min(I, J), -1 \leq \tau_c \leq 1.$$

Gudmeno ir Kraskelo koeficientas γ parodo kokių porų imtyje yra daugiau suderintų ar nesuderintų ($|\gamma|$ lygus įverčiui tikimybių skirtumo, kuris parodo kiek tikimybė, kad atsitiktinai parinkta pora bus suderinta, didesnė už tikimybę, kad ji bus nesuderinta, kai $\gamma > 0$ ir atvirkščiai, kai $\gamma < 0$):

$$\gamma = \frac{P - Q}{P + Q}.$$

Kai $\gamma = 0$, tai imtyje yra vienodas skaičius suderintų ir nesuderintų porų, kai $\gamma = 1$, tai visos poros suderintos, kai $\gamma = -1$, tai visos poros nesuderintos. Pastaba: koeficientas γ nevertina surištų porų.

5.6. Ryšio stiprimo matai vardų skalėje

Pateiksime keletą ryšio matų, kai stebimi kintamieji matuojami pagal vardų skalę. Visiems čia aprašytiems ryšio matams galioja ta pati taisyklė – kuo jie didesni, tuo požymių priklausomybė didesnė; kuo arčiau nulio – tuo priklausomybė silpnesnė. Daugelio iš jų maksimali reikšmė lygi vienetui.

Fi (Phi) koeficientas, tai ryšio matas 2x2 lentelėms dar vadinamas tarpusavio sutapimo koeficientu. Jis apibrėžiamas taip:

$$\phi = \sqrt{\chi^2 / n}.$$

Koeficiento ϕ kitimo sritis 2x2 lentelei yra [0; 1].

Kontingencijos koeficientas C - ryšio matai $I \times J$ dažnių lentelėms, dar vadinamas Pirsono kontingencijos matu. Jis apibrėžiamas taip:

$$C = \sqrt{\frac{\chi^2}{n + \chi^2}}, \quad 0 \leq C \leq \sqrt{\frac{\min(I, J) - 1}{\min(I, J)}}.$$

Matome, kad C neviršija 1. Naudojant šį koeficientą, reikia atsižvelgti į tai, kad didžiausia C reikšmė priklauso nuo eilučių ir stulpelių skaičiaus lentelėje. Pavyzdžiui, lentelei 4x4, $C_{\max}=0,866$.

Kramerio V koeficientas

$$V = \frac{\phi}{\sqrt{\min(I-1, J-1)}}, \quad 0 \leq V \leq 1.$$

Lentelei 2×2 Kramero V koeficientas sutampa su koeficientu ϕ .

Pavyzdys. Įvertinsime ryšio stiprumą tarp atsakymų į du anketos klausimus: Ar Jūs gaunate informacijos apie paslaugą? ir Ar Jūs asmeniškai naudojate paslaugą? Gauti rezultatai pateikti 5.10 pav.

Symmetric Measures

		Value	Approx. Sig.
Nominal by Nominal	Phi	,122	,000
	Cramer's V	,122	,000
	Contingency Coefficient	,121	,000
N of Valid Cases		1055	

a. Not assuming the null hypothesis.

b. Using the asymptotic standard error assuming the null hypothesis.

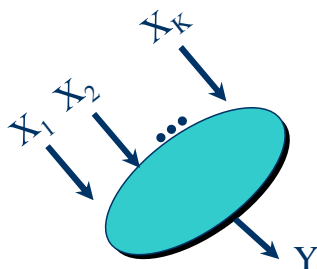
5.10 pav. Ryšio stiprumo matai vardų skalėje

Išvada. Į abu klausimus atsakė 1055 respondentai. Kramerio V koeficientas lygus 0,122. Tarp respondentų atsakymų yra labai silpnas ryšys. Nulinė hipotezė: "Populiacijos Kramerio V koeficientas lygus nuliui", atmesta (p -reikšmė $=0,000 < 0,05$).

6. REGRESINĖ ANALIZĖ

6.1. Regresinės analizės uždaviniai

Tiriant įvairius reiškinius labai dažnai sutinkami tarpusavyje susieti kintamieji (požymiai, faktoriai, rodikliai ir t.t), kai vieno kintamojo kitimas priklauso nuo to, kaip keičiasi kitų kintamųjų reikšmės. Koreliacinė analizė rodo ryšio tarp kintamųjų stiprumą, o regresinė analizė pateikia šio ryšio analizinę išraišką, kuria remiantis, galima su tam tikru patikimumu prognozuoti priklausomojo kintamojo Y reikšmę, žinant nepriklausomų kintamųjų $X_1, X_2, \dots, X_K$ reikšmes (6.1 pav.). Terminas regresija pirmą kartą buvo panaudotas 1887 metais Francio Galtono. Jis nustatė, kad vaikų, kurių tėvai yra aukšti, ūgis turi tendenciją mažėti, palyginus su vidutiniu populiacijos ūgiu. Jis įvedė regresijos terminą vieno kintamojo (vaikų ūgio) dydžiui nustatyti pagal kitą kintamąjį (tėvų ūgį).



6.1 pav. Priklausomi ir nepriklausomi kintamieji

Regresinėje analizėje gaunama regresijos lygtis susiejanti vieną kintamąjį, vadinamą priklausomu kintamuoju, su kitais – nepriklausomais kintamaisiais. Skiriamos dvi priklausomybių rūšys: funkcinė ir statistinė (atsitiktinė).

Funkcinė priklausomybė – tai neatsitiktinių dydžių priklausomybė. Esant funkciniai priklausomybei, nepriklausomų kintamųjų kitimas tiksliai apibūdina Y kitimą. Funkcinę priklausomybę žymėsime

$$Y = f(X_1, X_2, \dots, \beta_0, \beta_1, \beta_2, \dots)$$

ir suprasime kaip taisyklę, pagal kurią nepriklausomų kintamųjų reikšmėms yra priskiriamos priklausomo kintamojo Y reikšmės, čia $\beta_0, \beta_1, \beta_2, \dots$ - funkcijos parametrai. Pavyzdžiui, tiesinė funkcija

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_K X_K.$$

Statistinė priklausomybė – tai priklausomybė tarp atsitiktinių dydžių. Pavyzdžiui, žmogaus ūgis ir svoris, išlaidos reklamai ir pardavimų apimtys, akcijų ir aukso kainos rinkoje ir t.t. Esant statistinei priklausomybei, nepriklausomų kintamųjų kitimas apibūdina Y kitimą su tam tikra atsitiktine paklaida ε

$$Y = f(X_1, X_2, \dots, \beta_0, \beta_1, \beta_2, \dots) + \varepsilon.$$

Šiame kurse nagrinėsime tik tiesinę statistinę priklausomybę (klasikinę tiesinę regresinę analizę)

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_K X_K + \varepsilon.$$

Klasikiniame regresijos modelyje kintamieji $Y, X_1, X_2, \dots, X_K$ yra kiekybiniai, išmatuoti intervalų arba santykių skalėje ir turi tenkinti tam tikras sąlygas, kurios pateiktos 6.5 skyrelyje. Jeigu kintamieji netinka tiesinei regresijai, juos bandoma transformuoti. Dažnai kintamųjų negalima transformuoti taip, kad jie tiktų tiesinei regresinei analizei, tuomet galima taikyti netiesinę regresinę analizę, kuri realizuota daugumoje statistikos paketų (SPSS, SAS, STATA ir kt...). Šiame kurse netiesinės regresijos nenagrinėsime.

Be klasikinio regresijos modelio yra daug kitų regresijos modelių, kuriuose dalis kintamųjų yra kokybiniai ir išmatuoti silpnesnėse (vardų ir/arba tvarkos) skalėse. Šiame kurse apsiribosime tik pačiu paprasčiausiu modeliu, kai klasikinis regresijos modelis papildomas dvireikšmiais kokybiniais kintamaisiais, kurie dar vadinami psiaudokintamaisiais. Visi dvireikšmiai kintamieji gali įgyti tik dvi reikšmes: 0 ir 1. Pavyzdžiui,

$$Y = \beta_0 + \beta_1 X + \beta_2 D + \varepsilon, \text{ čia } D = \begin{cases} 1, & \text{moteris,} \\ 0, & \text{vyras.} \end{cases}$$

Įrašę į regresijos lygtį $D=1$, gausime regresijos lygtį moterims, o įrašę 0 gausime regresijos lygtį vyrams. Taigi gavome kiekvieną psiaudokintamojo reikšmę atitinkančią regresijos funkciją. Psiaudokintamuosius tikslinga taikyti kai regresijos tiesės abiejų kategorijų yra lygiagrečios. Dažnai į regresijos modelį įtraukiami kintamieji lytis, gyvenamoji vieta, politinės pažiūros, tautybė ir pan. Kai kategorinis kintamasis turi $m>2$ kategorijų, jis keičiams $(m-1)$ dvireikšmiu kintamuoju. Galima sudaryti regresijos modelį su tolydžių ir psiaudokintamųjų sąveikomis, pavyzdžiui:

$$Y = \beta_0 + \beta_1 X + \beta_2 D \cdot X + \varepsilon.$$

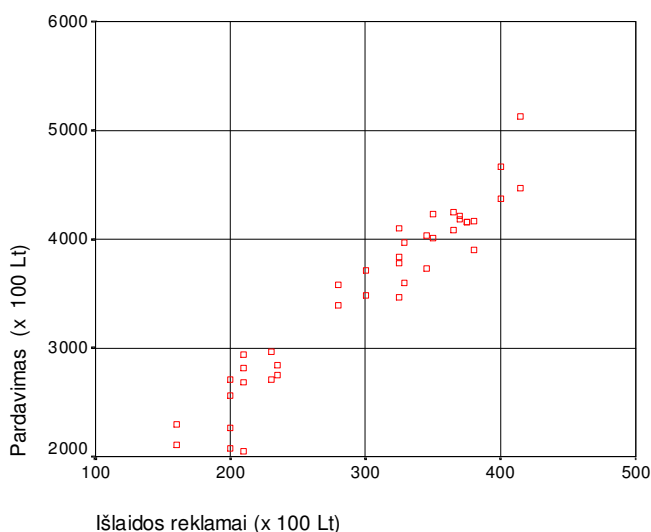
Plačiau apie kitus regresinės analizės modelius siūlome pasiskaityti [4,5,8,15,26]. Sekančiuose skyreliuose aptarsime pagrindinius daugialypės regresinės analizės uždavinius:

- Regresijos funkcijos parametrų taškinį ir intervalinių įverčių radimas.
- Hipotezių apie regresijos funkcijos parametrus tikrinimas.
- Prognozavimo paklaidų vertinimas.
- Regresijos modelio prielaidų tikrinimas.

6.2. Regresijos funkcijos koeficientai

6.2.1. Koeficientų taškiniai ir intervaliniai įverčiai

Turint duomenis, įprasta iš pradžių pavaizduoti juos grafiškai vadinamoje taškų sklaidos diagramoje (*scatter plot*). Ši diagrama (6.2 pav.) leidžia vizualiai parinkti geriausiai tinkantį regresinį modelį. Vieno nepriklausomo kintamojo atveju, kai diagramos taškai grupuojasi apie tiesę, turime pagrindą taikyti tiesinę regresiją.

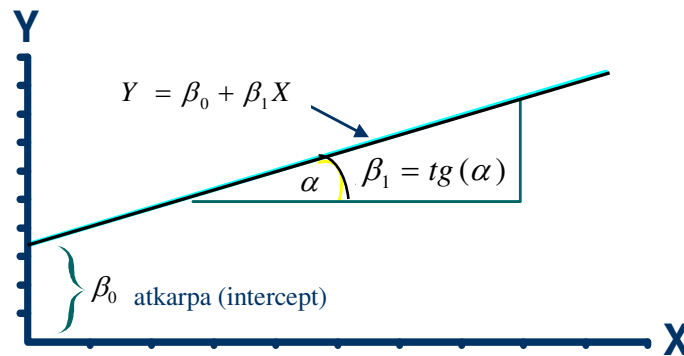


6.2 pav. Taškų sklaidos diagrama

Sekantis mūsų uždavinys įvertinti regresijos koeficientus ir užrašyti regresijos lygtį. Nagrinėsime paprasčiausią atvejį – vieno kintamojo tiesinę regresinę analizę. Tiesinė lygtis, siejanti priklausomą kintamąjį Y su nepriklausomu kintamuoju X , yra

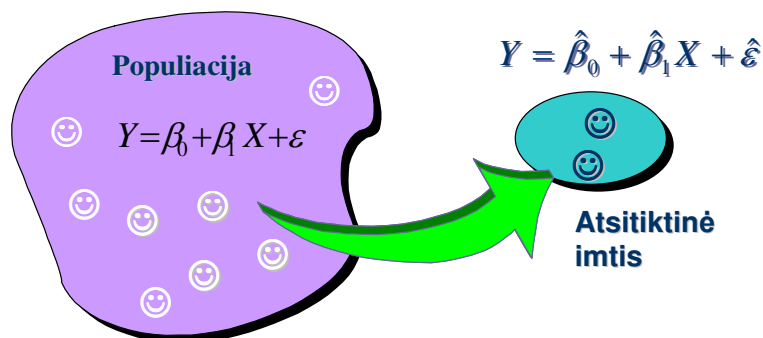
$$Y = \beta_0 + \beta_1 X,$$

čia β_1 – koeficientas, β_0 – laisvas narys. Laisvasis β_0 narys atitinka tašką, kuriame regresijos tiesė kerta Y ašį ir dažnai vadinamas konstanta (Constant) arba atkarpa (intercept). Koeficientas β_1 lygus tangentai kampo α , kuri regresijos tiesė sudaro su X ašimi (6.3 pav.).



6.3 pav. Vieno kintamojo regresijos lygties parametrų geometrinė interpretacija

Iš 6.2 pav. pateiktos diagramos matome, kad nors duomenys išsibarstę, tačiau jie gali būti aproksimuoti tiesine funkcija. Kaip gauti tokią tiesinę funkciją, kuri geriausiai apibūdintų turimus duomenis? Dažniausiai tam naudojamas mažiausiųjų kvadratų metodas. Šis metodas leidžia tarp visų galimų tiesių surasti tokią, kuri nutolusi nuo stebėtų duomenų mažiausiai. Mažiausiųjų kvadratų metodo pavadinimas atspindi faktą, kad minimizuojama atstumų tarp tiesės ir stebėtų duomenų kvadratų suma.



6.4. pav. Populiacijos ir imties regresijos lygtys

Tarkime, $Y = \beta_0 + \beta_1 X + \varepsilon$ išreiškia nežinomą tiesinės priklausomybės formą populiacijoje (populiacijos regresijos funkcija), čia ε yra populiacijos paklaida (paklaida gaunama pakeičiant tikrą priklausomybės formą tiesine), o β_0 ir β_1 – nežinomi populiacijos regresijos lygties koeficientai (6.4 pav.). Iš imties žinome tik keletą šios funkcijos realizacijų y_i , atitinkančių nepriklausomo kintamojo reikšmes x_i esant nenusakytoms kitų faktorių reikšmėms. Remiantis šiais duomenimis norime surasti tiesinės statistinės X ir Y priklausomybės nagrinėjamoje populiacijoje įvertinį, išreikštą tiesinės regresijos lygtimi $\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X = b_0 + b_1 X$, čia $\hat{\beta}_0$ ir $\hat{\beta}_1$ (arba b_0 ir b_1) yra nežinomų populiacijos koeficientų β_0 ir β_1 taškiniai įverčiai.

Pastaba: šiame kurse populiacijos regresijos lygties nežinomų koeficientų $\beta_0, \beta_1, \dots$, įverčius žymėsime $b_0, b_1, \dots$ arba $\hat{\beta}_0, \hat{\beta}_1, \dots$.

Koeficientų β_0, β_1 taškiniai įvertiniai b_0 ir b_1 yra randami iš sąlygos, kad $\sum_{i=1}^n (y_i - \hat{y}_i)^2$ turi įgyti mažiausią galimą reikšmę. Iš matematinės analizės žinome, kad dviejų kintamųjų funkcija

$$S(b_0, b_1) = \sum_{i=1}^n (y_i - b_0 - b_1 x_i)^2$$

gali įgyti ekstremumus tik taškuose, kuriuose tos funkcijos dalinės išvestinės pagal kintamuosius b_0 ir b_1 lygios nuliui. Nesileisdami į skaičiavimus, kuriuos galima rasti kiekviename matematinės statistikos vadovėlyje, pateiksime tik ieškomų įvertinių formules:

$$b_1 = \hat{\beta}_1 = \frac{\overline{xy} - \bar{x} \cdot \bar{y}}{\overline{x^2} - \bar{x}^2},$$

$$b_0 = \hat{\beta}_0 = \bar{y} - b_1 \bar{x}.$$

Čia, kaip ir anksčiau, brūkšnelis virš kintamojo žymi vidurkį. Pavyzdžiui, $\overline{x^2} = \frac{1}{n} \sum_{i=1}^n x_i^2$.

Kiekvienai iš tos pačios populiacijos sudarytai imčiai, aišku, gausime vis kitas tiesinės regresijos lygties koeficientų b_0 ir b_1 reikšmes. Vadinasi, galima kalbėti apie atsitiktinius dydžius $\mathbf{B}_0$ ir $\mathbf{B}_1$, kurių įgyjamos reikšmės konkrečiai imčiai atitinkamai lygios b_0 ir b_1 . Pažymėkime tų a.d. imties standartinius nuokrypius s_{b_0} ir s_{b_1} . Standartiniai nuokrypiai s_{b_0} ir s_{b_1} dažnai dar vadinami **standartinėmis koeficientų b_0 ir b_1 paklaidomis**. Jie apibrėžiami lygybėmis:

$$s_{b_1} = \sqrt{\frac{s_e^2}{\sum_{i=1}^n (x_i - \bar{x})^2}}, \quad s_{b_0} = \sqrt{s_{b_1}^2 \cdot \bar{x}^2 + \frac{s_e^2}{n}}.$$

Galima kalbėti ir apie koeficientų β_0 ir β_1 pasikliautinuosius intervalus. Jei patenkintos anksčiau aptartos pasikliautiniųjų intervalų sudarymo sąlygos, t.y. a.d. $\mathbf{B}_0$ ir $\mathbf{B}_1$ turi normalųjį skirstinį, tai vidurkių β_0 ir β_1 pasikliautinieji intervalai (pasiklioavimo lygmuo lygus $1 - \alpha$) randami pagal formules:

$$b_1 - t_{1-\alpha/2; n-2} \cdot s_{b_1} \leq \beta_1 \leq b_1 + t_{\alpha/2; n-2} \cdot s_{b_1},$$

$$b_0 - t_{1-\alpha/2; n-2} \cdot s_{b_0} \leq \beta_0 \leq b_0 + t_{\alpha/2; n-2} \cdot s_{b_0}.$$

Čia $t_{\alpha/2; n-2}$ pažymėtas Stjudento skirstinio su $n-2$ laisvės laipsniais $\alpha/2$ kvantilis. Analogiškai regresijos lygties koeficientų įverčiai randami, kai nepriklausomų kintamųjų yra du ir daugiau.

6.2.2. Hipotezių apie koeficientų lygybę nuliui tikrinimas

Sprendžiant regresinės analizės uždavinius dažnai domimasi klausimu, ar nepriklausomas kintamasis X_j turi įtakos Y kitimui. Paprastai X_j įtaka Y kitimui tikrinama nuline hipoteze

$H_0 : \beta_j = 0$, t.y. koeficientas prie X_j populiacijos regresijos lygtyje lygus nuliui. Alternatyvioji hipotezė $H_a : \beta_j \neq 0$ reiškia tiesinės priklausomybės tarp X_j ir Y egzistavimą.

Statistikos paketai regresinės analizės rezultatus pateikia lentelėmis. 6.1 lentelėje pateikta statistikos paketo SPSS regresijos lygties koeficientų lentelės struktūra, kurioje raidėmis b_i žymime nežinomų koeficientų β_i taškinius įverčius $b_j = \hat{\beta}_j$, $j = \overline{0, K}$.

Radę taškinius įverčius galime užrašyti imties regresijos funkciją

$$\hat{Y} = b_0 + b_1 X_1 + b_2 X_2 + \dots + b_K X_K.$$

6.1 lentelė. SPSS regresijos lygties koeficientų lentelė

Koeficientai	Regresijos lygties koeficientai b_i (Unstandardized Coefficients)		Standartizuotieji koeficientai (Standardized Coefficients)	Stebėta Stjudento statistikos reikšmė (t)	Stebėtas reikšmingumo lygmuo (p-reikšmė) (Sig.)
	Koeficientų taškiniai įverčiai (B)	Koeficientų standartinės paklaidos (Std. Error)	(Beta)		
β_0 (Constant)	b_0	s_{b_0}		$t_{imt}^0 = \frac{b_0}{s_{b_0}}$	α_{imt}^0
β_1	b_1	s_{b_1}	$BETA_1$	$t_{imt}^1 = \frac{b_1}{s_{b_1}}$	α_{imt}^1
β_2	b	s_{b_2}	$BETA_2$	$t_{imt}^2 = \frac{b_2}{s_{b_2}}$	α_{imt}^2
...	...	...	...	...	...
β_K	b_K	s_{b_K}	$BETA_K$	$t_{imt}^K = \frac{b_K}{s_{b_K}}$	α_{imt}^K

a. Priklausomas kintamasis (Dependent Variable): Y

Hipotezių apie regresijos lygties koeficientus tikrinimui

$$H_0 : \beta_j = 0,$$

$$H_a : \beta_j \neq 0,$$

naudojamas Stjudento t kriterijus $T_j = \frac{b_j}{s_{b_j}} \sim St(n - K - 1), j = 0, 1, \dots, K$. Jeigu nulinė hipotezė

atmetama, tai koeficientas β_j statistiškai reikšmingai skiriasi nuo nulio, t.y. Y reikšmės priklauso nuo X_j . Tačiau, reikia neužmiršti, kad nedidelius, bet statistiškai reikšmingus skirtumus gali lemti didelė imtis (esant dideliai imčiai fiksuojami ir mažiausi koeficientų skirtumai nuo nulio).

Lygindami koeficientus β_j tarpusavyje mes negalime nustatyti kintamųjų X_j santykinę svarbą prognozuojant Y, nes β_j didumas labai priklauso nuo X_j matavimo vienetų ir nuo duomenų sklaidos. Todėl, dažnai ieškoma ir standartizuotosios tiesinės regresijos funkcijos. Priklausomas kintamasis Y ir nepriklausomi kintamieji $X_1, X_2, \dots, X_K$ keičiami z-reikšmėmis ir mažiausiųjų kvadratų metodu randami standartizuotieji koeficientai $BETA_j, j = \overline{1, K}$. Pagal standartizuotus koeficientus $BETA_j$ galima nuspręsti, koks kintamasis X_j daro didesnę įtaką Y prognozei. Absoliučiuoju didumu didesnis $BETA_j$ rodo didesnę Y priklausomybę nuo X_j .

6.3. Standartinė regresijos paklaida ir apibrėžtumo koeficientas

Turint imties regresijos funkciją kyla natūralus klausimas, ar gerai regresijos funkcija atitinka duomenis? Vieni iš svarbiausių tinkamumo matų yra standartinė regresijos paklaida ir apibrėžtumo koeficientas. Standartinės regresijos paklaidos ir apibrėžtumo koeficientų prasmę paaiškinsime

vieno kintamojo tiesinės regresijos atveju ir po to apibendrinsime kelių kintamųjų atveju. Tarkime, kad pagal stebėjimo rezultatus (x_i, y_i) , $i = \overline{1, n}$ gavome imties regresijos lygtį

$$\hat{y} = b_0 + b_1 x.$$

Pagal imties regresijos lygtį prognozuojam a.d. Y reikšmes

$$\hat{y}_i = b_0 + b_1 x_i.$$

Nagrinėsime, ar gauta regresijos lygtis yra patikima. Regresijos tiesė tuo geriau aprašo stebėjimo duomenis, kuo arčiau šios tiesės išsidėstę taškai. Regresijos tiesės patikimumo matas yra standartinė regresijos paklaida. Ji žymima S_e ir jos interpretacija panaši į vidutinio standartinio nuokrypio interpretaciją, t.y. S_e yra sklaidos matas. Prisiminkime, kad standartinis nuokrypis nusako stebėjimų išsisklaidymą apie vidurkį, o standartinė regresijos paklaida – stebėjimų išsisklaidymą apie regresijos tiesę. Formulė, aprašanti standartinę regresijos paklaidą, taip pat primena standartinio nuokrypio formulę

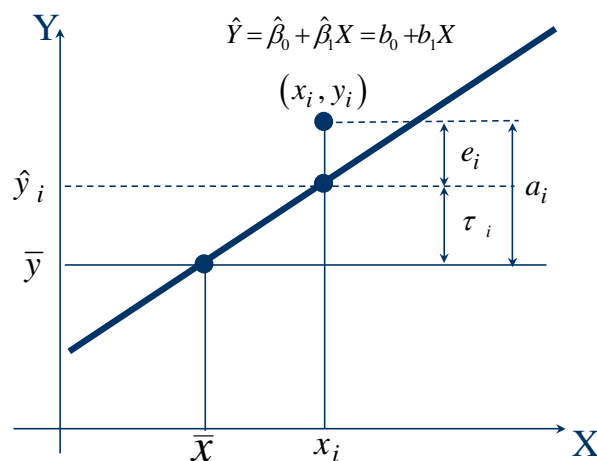
$$\hat{e} = e_i = y_i - \hat{y}_i = y_i - b_0 - b_1 x_i,$$

čia

e_i – prognozės paklaidos įvertis (residual),

y_i – stebėta reikšmė,

$\hat{y}_i$ – pagal regresijos lygtį apskaičiuota reikšmė (prognozuojama reikšmė (predicted value)).



6.5 pav.

Prognozės paklaidos standartinis nuokrypis vadinamas **standartine regresijos paklaida**

$$S_e = \sqrt{\frac{\sum_{i=1}^n (e_i - E(e_i))^2}{n-2}} = \sqrt{\frac{\sum_{i=1}^n e_i^2}{n-2}}.$$

Standartinė regresijos paklaida S_e apibūdina eksperimento taškų sklaidą apie regresijos tiesę. Aišku, kad kuo mažesnė paklaida, tuo arčiau regresijos tiesės yra išsidėstę stebėjimai. Jeigu $S_e = 0$, tai visi stebėjimai yra išsidėstę tiesėje. Daugialypės regresijos atveju $e_i = y_i - \hat{y}_i$,

$$S_e = \sqrt{\frac{\sum_{i=1}^n e_i^2}{n-K-1}}, \text{ čia } n - \text{imties didumas, } K - \text{nepriklausomų kintamųjų skaičius regresijos lygtyje.}$$

Regresinėje analizėje naudojamos trys nuokrypių kvadratų sumos:

① **Bendroji nuokrypių kvadratų suma** (total sum of squares)

$$SS_p = \sum_{i=1}^n (y_i - \bar{y})^2 = \sum_{i=1}^n a_i^2 .$$

Ji apibūdina a.d. Y reikšmių sklaidą apie vidurkį.

② **Regresinė nuokrypių kvadratų suma** (regression (model) sum of squares)

$$SS_R = \sum_{i=1}^n (\hat{y}_i - \bar{y})^2 = \sum_{i=1}^n \tau_i^2 .$$

Ji apibūdina eksperimento taškų sklaidos apie vidurkį $\bar{y}$ dalį, kuri paaiškinama Y tiesine regresija X atžvilgiu (t.y. jų tiesine priklausomybe) – „paaiškinama nuokrypių dalis“.

③ **Liekanų kvadratų suma** (residual (error) sum of squares)

$$SS_e = \sum_{i=1}^n (y_i - \hat{y}_i)^2 = \sum_{i=1}^n e_i^2$$

apibūdina eksperimento taškų sklaidą apie regresijos tiesę, t.y. taškų sklaidos apie vidurkį $\bar{y}$ dalį, kuri nepaaiškinama tiesine regresija – „nepaaiškinama nuokrypių dalis“.

Galima įrodyti, kad

$$SS_p = SS_R + SS_e .$$

Apibrėžtumo koeficientas

$$r^2 = \frac{SS_R}{SS_p}$$

parodo, kuri atsitiktinio dydžio Y sklaidos dalis apie vidurkį $\bar{Y}$ paaiškinama tiesine regresija. Kuo r^2 artimesnis vienetui, tuo didesnė dispersijos dalis paaiškinama tiesine regresija, t.y. tuo geriau regresijos lygtis aprašo kintamąjį Y, tačiau reikia neužmiršti, kad kai kintamųjų skaičius regresijos lygtyje K nedaug skiriasi nuo imties didumo n, tai apibrėžtumo koeficientas taip pat yra arti vieneto. Vieno kintamojo regresinėje analizėje apibrėžtumo koeficientas r^2 panašiai kaip ir koreliacijos koeficientas yra dviejų atsitiktinių dydžių Y ir X tiesinio ryšio matas. Daugialypės koreliacijos koeficientas yra Y tiesinės priklausomybės nuo visų kintamųjų $X_1, X_2, \dots, X_K$ matas. Daugialypėje regresinėje analizėje naudojamas koreguotas apibrėžtumo koeficientas (Adjusted R Square), kuris atsižvelgia į imties didumą n ir nepriklausomų kintamųjų skaičių regresijos lygtyje K

$$r_{adj}^2 = 1 - \frac{\frac{SS_e}{n-K-1}}{\frac{SS_p}{n-1}} .$$

Koreguotas apibrėžtumo koeficientas parodo kurią atsitiktinio dydžio Y sklaidos dalį galima paaiškinti Y tiesine regresija kintamųjų $X_1, X_2, \dots, X_K$ atžvilgiu. Statistikos paketai regresinės analizės rezultatus pateikia lentelėmis. 6.2 pateikta statistikos paketo SPSS regresinės analizės rodiklių suvestinė.

6.2 lentelė. Regresinės analizės pagrindinių rodiklių suvestinė

Modelis (Model)	Daugialypės koreliacijos koeficientas (R)	Daugialypės regresijos apibrėžtumo koeficientas (R Square)	Koreguotas daugialypės regresijos apibrėžtumo koeficientas (Adjusted R Square)	Standartinė regresijos paklaida (Std. Error of the Estimate)
--------------------	--	--	---	---

1	r	r^2	r_{adj}^2	S_e^2
---	-----	-------	-------------	---------

a. Nepriklausomi kintamieji (Predictors): konstanta (Constant), $X_1, X_2, \dots, X_K$.

b. Priklausomas kintamasis (Dependent Variable): Y

6.4. Prognozavimo pasikliautiniai intervalai

Standartinė regresijos paklaida naudojama intervalams, į kuriuos su tam tikru patikimumu patenka individualios Y reikšmės arba Y reikšmių vidurkis, apskaičiuoti. Paaiškinsime priklausomo kintamojo Y vidurkio prognozavimą vieno kintamojo regresinės analizės atveju. Esant fiksuotam X , galima prognozuoti Y vidurkio reikšmių intervalą. Jeigu fiksuojame nepriklausomo kintamojo reikšmę $X=X_p$, tuomet Y sąlyginio vidurkio $E(Y / X = X_p)$ pasikliautinąjį intervalą randame pagal formulę

$$\hat{Y} - t_{1-\alpha/2, n-2} \cdot S_{\hat{Y}} \leq E(Y / X = X_p) \leq \hat{Y} + t_{\alpha/2, n-2} \cdot S_{\hat{Y}} \quad \text{čia}$$

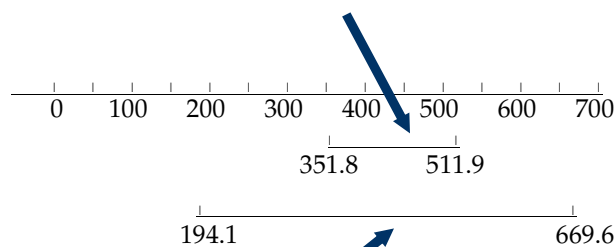
$$S_{\hat{Y}} = S_e \sqrt{\frac{1}{n} + \frac{(X_p - \bar{X})^2}{\sum_{i=1}^n (X_i - \bar{X})^2}}$$

Pasikliautinojo intervalo plotį įtakoja:

- pasikliovivimo lygmuo $1 - \alpha$,
- duomenų sklaida,
- imties didumas n ,
- taško X_p atstumas nuo X vidurkio.

Panašiai randamas ir prognozuojamų individualių Y reikšmių pasikliautinis intervalas, tik jis visada yra platesnis už sąlyginio Y vidurkio prognozės intervalą (6.5 pav.)

Y sąlyginio vidurkio pasikliautinis intervalas



Individualios Y reikšmės pasikliautinis intervalas

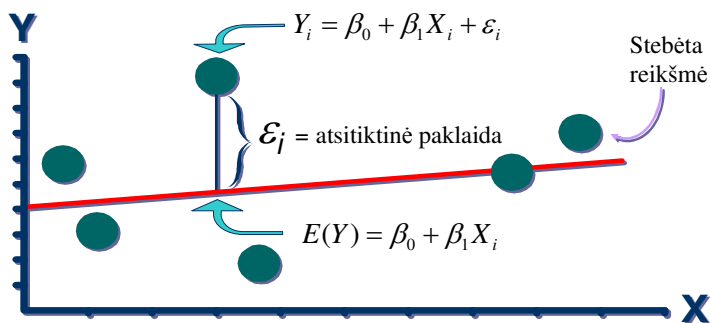
6.6 pav. Y sąlyginio vidurkio ir individualių Y reikšmių pasikliautiniai intervalai.

6.5. Regresijos prielaidų tikrinimas

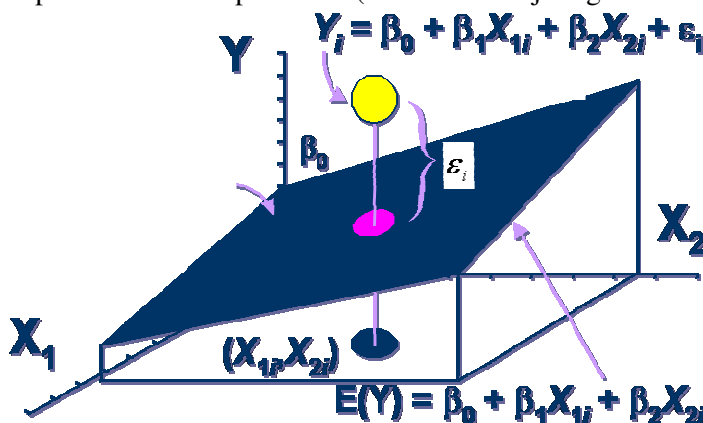
Šiame skyrelyje aptarsime kokias sąlygas turi tenkinti kintamieji ir kaip nustatyti ar tiesinės regresijos modelis tinka analizuojamiems duomenims. Užrašysime regresijos funkciją fiksuotoms (neatsitiktinėms) X_j reikšmėms

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \dots + \beta_K X_{Ki} + \varepsilon_i .$$

Didžioji dalis regresinės analizės prielaidų – tai reikalavimai, kuriuos turi tenkinti atsitiktinės paklaidos ε_i , kurios parodo, kiek stebėtoji Y reikšmė skiriasi nuo reikšmės, kurią gautume prognozuodami pagal regresijos lygtį. 6.7-6.8 pav. pavaizduotos atsitiktinės paklaidos vieno ir dviejų kintamųjų regresijos atveju.



6.7 pav. Atsitiktinės paklaidos (vieno kintamojo regresinė analizė)

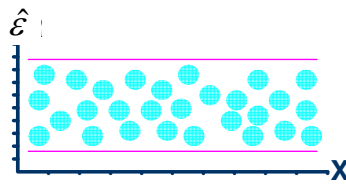


6.8 pav. Atsitiktinės paklaidos (dviejų kintamųjų regresinė analizė)

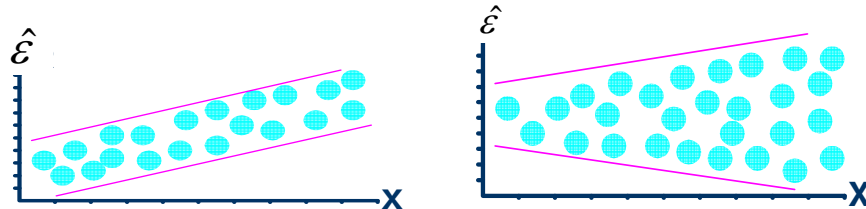
Suformuluosime pagrindines tiesinės regresinės analizės prielaidas:

1. Atsitiktinės paklaidos ε_i yra normaliai pasiskirstę atsitiktiniai dydžiai.
2. Visų ε_i vidurkiai lygūs nuliui, $E\varepsilon_i = 0$.
3. Visų ε_i dispersijos lygios (homoskedastiškumo prielaida), $D\varepsilon_i = \sigma^2$.
4. Visi ε_i nepriklausomi.
5. Duomenyse nėra išskirčių.
6. Nėra stiprios koreliacijos tarp nepriklausomų kintamųjų $X_1, X_2, \dots, X_K$ (multikolinearumo problema).

Trumpai pirmas keturias prielaidas galima užrašyti: $\varepsilon_i \sim N(0, \sigma^2)$. **Homoskedastiškumo** prielaida arba dispersijų lygybės prielaida yra reikalavimas, kad su kiekviena fiksuota X_i reikšme paklaidų sklaida būtų vienoda (6.9 pav.). Regresijos modelis yra jautrus šios prielaidos pažeidimams. Jeigu prielaida netankinama, tai prognozė gali būti iš esmės iškreipta. Būdingi homoskedastiškumo pažeidimo atvejai pavaizduoti 6.10 pav. Tokie duomenys vadinami heteroskedastiškais.



6.9 pav. Homoskedastiški duomenys



6.10 pav. Heteroskedastiški duomenys

Kai imties didumas nedidelis, tai net ir vienas, labai nuo kitų besiskiriantis stebėjimas gali statistiškai reikšmingai pakeisti regresijos lygties koeficientų įverčius. Todėl reikia mokėti rasti duomenų išskirtis ir po to išsiaiškinti ką su tomis išskirtimis daryti. Yra nemažai išskirčių nustatymo metodų, kurių šiame kurse nenagrinėsime, o apsiribosime tik išskirčių nustatymu „iš akies“ pagal paklaidų grafikus.

Multikolinearumas. Daugialypės regresinės modelis geriausiai tinka prognozuoti, kai visi nepriklausomi kintamieji $X_1, X_2, \dots, X_K$ stipriai koreliuoja su Y , o tarpusavyje nekoreliuoja arba koreliuoja silpnai. Kai tarp nepriklausomų kintamųjų $X_1, X_2, \dots, X_K$ yra stipriai koreliuojančių susiduriame su multikolinearumo problema. Dėl kintamųjų multikolinearumo negalima gerai atskirti koreliuojančių kintamųjų įtakos Y prognozei, atsiranda „ne tas“ regresijos funkcijos koeficiento ženklas, regresijos lygties koeficientai tampa labai nestabilūs – keli papildomi arba pašalinti stebėjimai gali ženkliai juos pakeisti. Kintamųjų multikolinearumas nustatomas skaičiuojant:

1. **Dispersijos mažėjimo daugiklį VIF** (Variance Inflation Factor). Jeigu $4 < \text{VIF} < 10$ - galima įtarti, kad kintamasis yra multikolinearus (nuo vidutinio iki stipraus), jeigu $\text{VIF} \geq 10$ - kintamasis „per daug multikolinearus“.
2. **Toleranciją** (Tolerance). $\text{Tolerancija} = 1/\text{VIF}$.
3. **Sąlygojimo indeksas** (Condition index). Jeigu $10 < \text{Sąlygojimo indeksas} \leq 30$ tai galima įtarti, kad kintamasis yra multikolinearus (nuo vidutinio iki stipraus). $\text{Sąlygojimo indeksas} > 30$ rodo stiprų multikolinearumą.

Multikolinearumo mažinimo metodai yra pakankamai sudėtingi. Paskelbtuose darbuose nėra vieningos nuomonės šiuo klausimu. Dažnai siūloma padidinti imtį, atsisakyti dalies multikolinearių kintamųjų, keisti kintamuosius pagrindinėmis komponentėmis (žr. 8 skyrių) ir kiti metodai. Šiame kurse multikolinearumo mažinimo metodai neapptariami.

Ieškant daugialypės regresijos lygties tikrinama hipotezė apie regresijos tiesiškumą. Jeigu visi regresijos lygties koeficientai prie nepriklausomų kintamųjų lygūs nuliui, tai regresijos modelis prognozėms visiškai netinka.

$$H_0: \beta_1 = \beta_2 = \dots = \beta_K = 0,$$

$$H_a: \text{bent vienas } \beta_j \neq 0.$$

Hipotezė tikrinama naudojant Fišerio F kriterijų su dešinine kritine sritimi

$$F = \frac{\overline{SS}_R}{\overline{SS}_e} \sim F(K, n - K - 1).$$

Hipotezės priėmimo sritis $\mathbf{F}_{H_0} = [0; F_{1-\alpha; 1; n-K-1})$, kritinė sritis $\mathbf{F}_K = [F_{1-\alpha; 1; n-K-1}; \infty)$.

Jei $F_{imt} \in \mathbf{F}_K$, tai H_0 atmetama ir priimam H_a , t.y. regresija yra tiesinė, bent vienas koeficientas nelygus nuliui ir regresijos modelis bent jau iš dalies prognozėms tinka. Hipotezės tikrinimo rezultatai pateikiami dispersinės analizės lentelėje (6.3 lentelė)..

6.3 lentelė. Dispersinės analizės lentelė

Nuokrypių šaltinis	Nuokrypių kvadratų suma (Sum of Squares)	Laisvės laipsnių skaičius (df)	Nuokrypių kvadratų vidurkis (Mean Square)	Stebėta Fišerio statistikos reikšmė (F)	Reikšmingumo lygmuo (Sig.)
Regresija (Regression)	SS_R	K	$\overline{SS}_R = \frac{SS_R}{K}$	$F_{imt} = \frac{\overline{SS}_R}{\overline{SS}_e}$	α_{imt}
Atsitiktiniai faktoriai (paklaidos) (Residual)	SS_e	n-K-1	$\overline{SS}_e = \frac{SS_e}{n - K - 1}$		
Total	SS_p	n-1			

a Nepriklausomi kintamieji (Predictors): konstanta (Constant), $X_1, X_2, \dots, X_K$

b Priklausomas kintamasis (Dependent Variable): Y

6.6. Regresinės analizės taikymo klaidos

Regresinė analizė yra galingas statistikos įrankis, deja dažnai ši analizė yra neteisingai taikoma arba traktuojama, dėl to daromos neteisingos prognozės arba priimami neteisingi sprendimai. Paminėsime dažniausiai daromas klaidas:

- Regresinė analizė taikoma, kai netenkinamos jos taikymo prielaidos (dažnai tyrėjai nepatikrina regresijos liekanų normalumo, išskirčių įtakos regresijos koeficientams, homoskedastiškumo, multikolinearumo ir kitų prielaidų).
- Taikant regresinę analizę, negalima išeiti iš intervalo (srities), kuriame galioja regresijos lygtis, ribų, nes kitame intervale (sirtyje) regresijos lygties išraiška gali būti kitokia.
- Negalima regresijos ir koreliacijos tapatinti su priežastingumu. Taikant regresinę analizę, negalima teigti, kad vieno kintamojo pakitimai sąlygoja kito kintamojo pakitimus (yra to kintamojo pasikeitimo priežastis). Kartais, taikant regresinę analizę, galima rasti ryšį tarp kintamųjų, neturinčių nieko tarpusavyje bendro. Pavyzdžiui, nustačius stiprią koreliaciją tarp išsimokslinimo ir jo gyvenamo būsto ploto nėra pagrindo teigti, kad vienas iš šių veiksmų sąlygoja kitą, todėl darant statistines išvadas būtina vadovautis sveiku protu.
- Neteisinga koreliacijos bei apibrėžtumo koeficientų interpretacija. Jei koreliacijos koeficientas, sakykime, lygus 0,7, tai negalima teigti, kad regresijos lygtis paaiškina 70% viso kintamojo Y

dispersijos. Jei $r = 0,7$, tai $r^2 = 0,49$. Taigi tik 49% visos Y dispersijos paaiškinama regresijos lygtimi.

6.7. Regresinės analizės pavyzdžiai

6.7.1. Išlaidų reklamai analizė

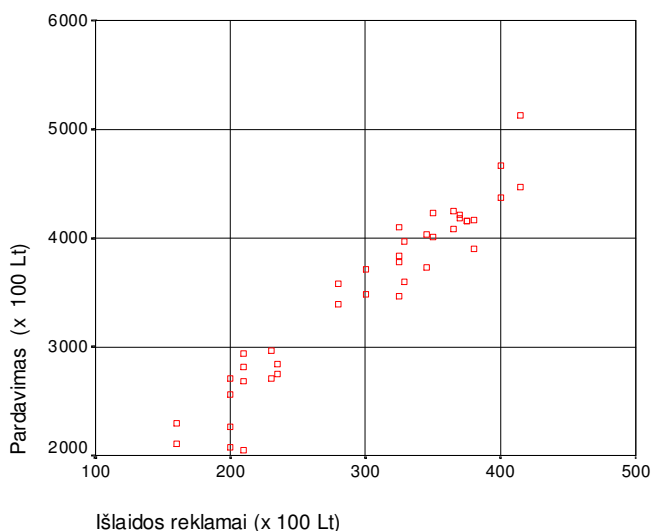
Duoti du kintamieji, kurių reikšmės išmatuotos santykių skalėje:

- **parduota** – prekių pardavimas (x100 Lt).
- **reklama** – išlaidos reklamai (x100 Lt).

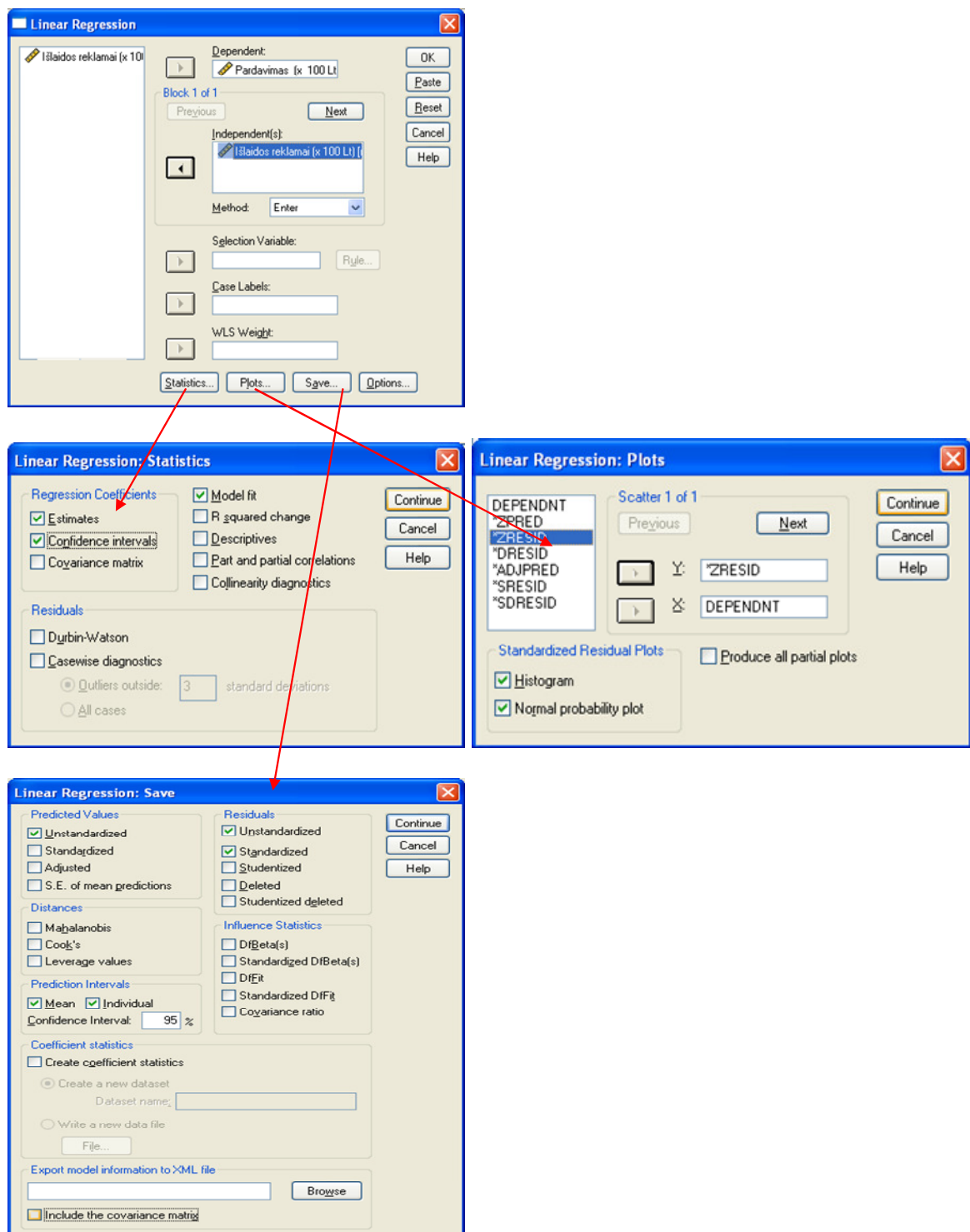
Sudarysime modelį, kurio pagalba būtų galima prognozuoti prekių pardavimus (kintamasis **parduota**) žinant išlaidas reklamai (kintamasis **reklama**). Uždavinio sprendimui taikysime vieno kintamojo tiesinės regresinės analizės modelį. Statistikos pakete SPSS tiesinei regresinei analizei skirtas meniu *Analyze → Regression → Linear*. Nubraižysime duomenų sklaidos diagramą, įvertinsime tiesinio ryšio stiprumą, apskaičiuosime Pirsono koreliacijos koeficientą, apibrėžtumo koeficientą, rasime regresijos lygties koeficientų taškinius ir intervalinius įverčius, patikrinsime hipotezes apie regresijos lygties koeficientų reikšmingumą. Vizualiai patikrinsime regresijos modelio prielaidas: ar standartizuotųjų liekanų skirstinys yra normalusis, ar standartizuotųjų liekanų sąlyginė dispersija pastovi? Patikrinsime standartizuotųjų liekanų skirstinio suderinamumo su standartiniu normaliuoju skirstiniu hipotezę. Apskaičiuosime priklausomo kintamojo individualių reikšmių ir vidurkių prognozės pasikliautinuosius intervalus fiksuotoms nepriklausomo kintamojo reikšmėms. Šio uždavinio sprendimui SPSS menių laukų užpildymas pateiktas 6.12 pav.

Nubraižome išlaidų reklamai ir pardavimų duomenų sklaidos diagramą (6.11 pav.).

Išvada. Duomenų sklaidos diagrama rodo, kad tarp išlaidų reklamai ir pardavimų apimčių yra tiesinė statistinė priklausomybė. Pasirenkame tiesinės regresijos modelį $Y = \beta_0 + \beta_1 X + \varepsilon$, čia Y- prekių pardavimas, X- išlaidos reklamai.



6.11 pav. Išlaidų reklamai ir pardavimų duomenų sklaidos diagrama



6.12 pav. SPSS tiesinės regresinės analizės meniu

6.4 lentelė. Regresijos modelio rodiklių suvestinė

Model Summary^b

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	Durbin-Watson
1	,962 ^a	,925	,923	221,52	2,162

a. Predictors: (Constant), Išlaidos reklamai (x 100 Lt)

b. Dependent Variable: Pardavimas (x 100 Lt)

Apskaičiuojame imties Pirsono koreliacijos ir apibrėžtumo koeficientus (6.4 lentelė).

Išvada. Imties Pirsono koreliacijos koeficientas $r=0,96$. Tarp stebėtų pardavimų ir išlaidų reklamai reikšmių yra labai stiprus tiesinis ryšys. Apibrėžtumo koeficiento (R Square) reikšmė $r^2=0,925$. Galime sakyti, kad tiesinės regresijos modelis paaiškina 92,5% pardavimų (Y) reikšmių sklaidos apie pardavimų (Y) vidurkį, o likusi dalis yra nepaaiškinama sklaidos dalis.

Randame regresijos lygties koeficientų taškinius įverčius, pasikliautinuosius intervalus ir patikriname hipotezes apie koeficientų lygybę nuliui (6.5 lentelė).

6.5 lentelė. Regresijos lygties koeficientai

Coefficients^a

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	95% Confidence Interval for B	
		B	Std. Error	Beta			Lower Bound	Upper Bound
1	(Constant)	520,609	142,259		3,660	,001	232,621	808,596
	Išlaidos reklamai (x 100 Lt)	9,958	,459	,962	21,681	,000	9,028	10,888

a. Dependent Variable: Pardavimas (x 100 Lt)

Gavome imties regresijos lygtį

$$\text{parduota} = 520,61 + 9,96 \cdot \text{reklama}.$$

Išvada. Padidinus išlaidas reklamai vienu vienetu, pardavimai padidėja vidutiniškai 9,96 vienetais. Abu regresijos lygties koeficientai yra reikšmingi, nes atmestos abi nulinės hipotezės apie koeficientų lygybę nuliui:

$$H_0: \beta_0 = 0, H_a: \beta_0 \neq 0, (p=0,001 < 0,05),$$

$$H_0: \beta_1 = 0, H_a: \beta_1 \neq 0, (p=0,000 < 0,05).$$

Apskaičiuojame populiacijos regresijos lygties koeficientų pasikliautinuosius intervalus:

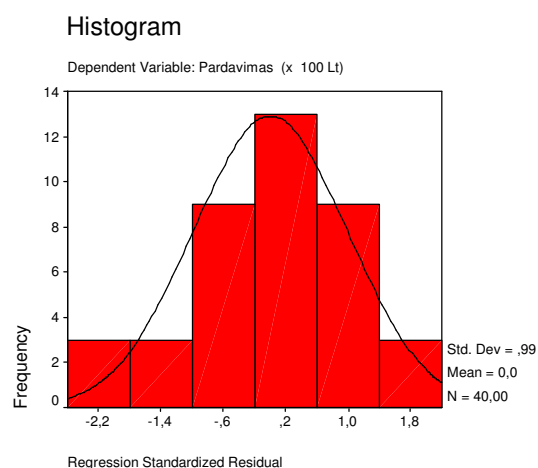
$$PI_{0,95}(\beta_0) = (232,62; 808,60),$$

$$PI_{0,95}(\beta_1) = (9,03; 10,89).$$

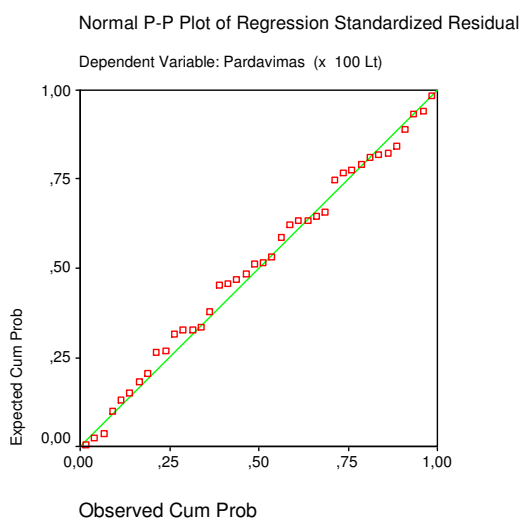
Išvada. Su 95% garantija galime prognozuoti, kad padidinus išlaidas reklamai vienu vienetu, **pardavimų** vidutinis padidėjimas bus intervale nuo 9,03 iki 10,89 vienetų.

Nubraižome standartizuotųjų liekanų histogramą (6.13 pav.).

Išvada. Vizualiai palyginę, galime teigti, kad standartizuotųjų liekanų histograma yra suderinta su standartinio normaliojo skirstinio kreive (t.y. standartizuotųjų liekanų skirstinys yra suderintas su standartiniu normaliuoju skirstiniu).



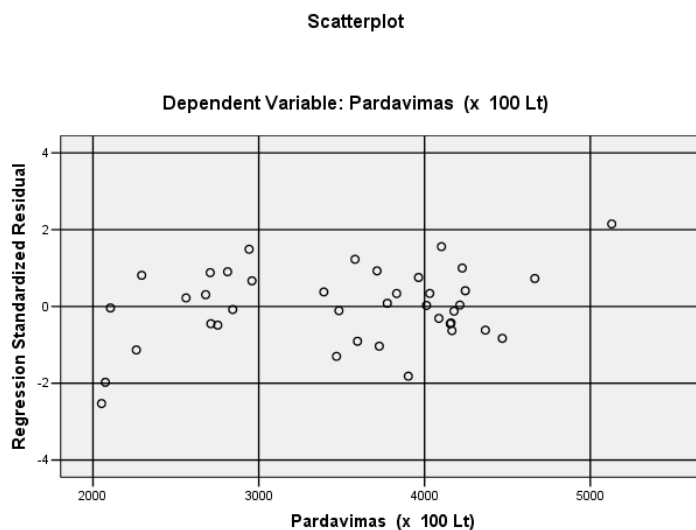
6.13 pav. Standartizuotųjų liekanų histograma.



6.14 pav. Standartizuotųjų liekanų P-P grafikas.

Nubraižome standartizuotųjų liekanų P-P grafiką (ištiesintų standartinio normaliojo skirstinio ir empirinio skirstinio funkcijų palyginimo grafiką) (6.14 pav.).

Išvada. Vizualiai palyginę, galime teigti, kad standartizuotųjų liekanų skirstinys yra suderintas su standartiniu normaliuoju skirstiniu.



6.15 pav. Pardavimų ir standartizuotųjų liekanų reikšmių sklaidos diagrama.

Išvada. Pardavimų ir standartizuotųjų liekanų reikšmių sklaidos diagrama (6.15 pav.) rodo, kad sąlyginė standartizuotųjų liekanų dispersija yra pastovi.

Patikrinsime suderinamumo hipotezę: „Standartizuotųjų liekanų skirstinys yra standartinis normalusis“. Hipotezės tikrinimo rezultatai pateikti 6.6 lentelėje.

Išvada. Suderinamumo hipotezė neatmesta ($p=0,977>0,05$), standartizuotųjų liekanų skirstinys yra suderintas su standartiniu normaliuoju skirstiniu. **Teorinės prielaidos tenkinamos, tiesinės regresijos modelis tinka analizuojamiems duomenims.**

Apskaičiuojame **pardavimų vidutinių** reikšmių prognozės pasikliautinosius intervalus fiksuotoms **išlaidų reklamai** reikšmėms (6.7 lentelėje stulpeliai LMCI_1 ir UMCI_1) ir nubraižome pasikliautiniųjų intervalų grafiką (6.16 pav.).

Apskaičiuojame **pardavimų individualių** reikšmių prognozės pasikliautinosius intervalus fiksuotoms **išlaidų reklamai** reikšmėms (6.7 lentelėje stulpeliai LICI_1 ir UICI_1) ir nubraižome pasikliautiniųjų intervalų grafiką (6.17 pav.).

6.6 lentelė. Suderinamumo hipotezės tikrinimo rezultatai
(Kolmogorovo-Smirnovo statistika)

One-Sample Kolmogorov-Smirnov Test		ZRE_1 Standardized Residual
N		40
Normal Parameters(a,b)	Mean	,0000
	Std. Deviation	,98710
Most Extreme Differences	Absolute	,075
	Positive	,055
	Negative	-,075
Kolmogorov-Smirnov Z		,477
Asymp. Sig. (2-tailed)		,977

a Test distribution is Normal.

b Calculated from data.

6.7 lentelė

pardavimai konspekto pvz.sav - SPSS Data Editor

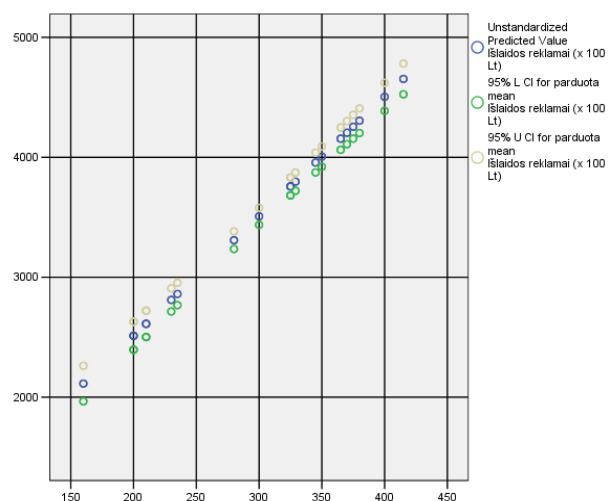
File Edit View Data Transform Analyze Graphs Utilities Window Help

7 :

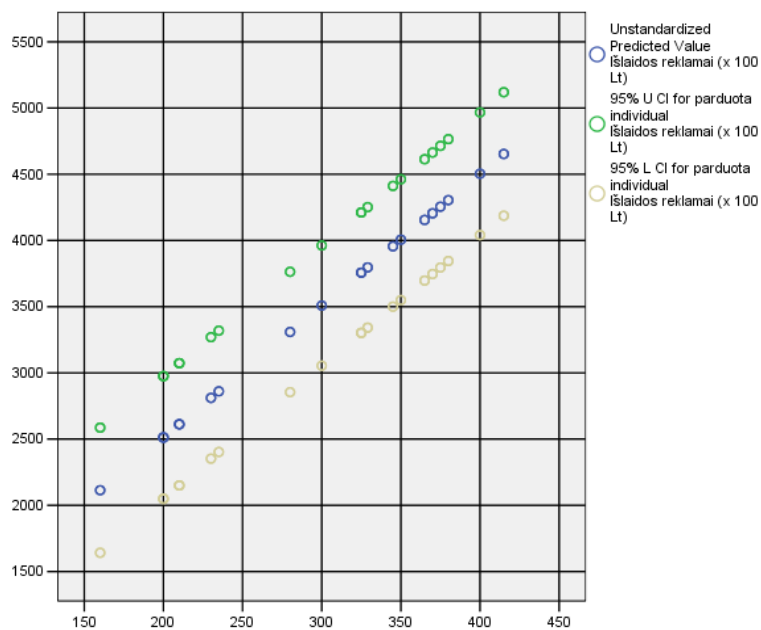
	parduota	reklama	PRE_1	RES_1	ZPR_1	ZRE_1	LMCI_1	UMCI_1	LICI_1	UICI_1
1	2843	235	2860,75	-17,5	,84	-,08	2767,46	2954,04	2402,70	3318,80
2	2680	210	2611,80	68,16	-1,17	,31	2501,98	2721,62	2150,10	3073,50
3	2105	160	2113,90	-8,62	-1,82	-,04	1965,50	2262,29	1641,53	2586,26
4	4031	345	3956,14	75,18	,58	,34	3873,90	4038,37	3500,21	4412,07
5	3775	325	3756,98	18,46	,32	,08	3682,41	3831,54	3302,37	4211,58
6	4086	365	4155,30	-68,9	,84	-,31	4062,25	4248,35	3697,29	4613,30
7	4665	400	4503,83	161,5	1,29	,73	4387,05	4620,61	4040,42	4967,24
8	4213	370	4205,09	8,12	,90	,04	4108,97	4301,21	3746,45	4663,73
9	4012	350	4005,93	5,97	,64	,03	3921,24	4090,61	3549,55	4462,31
10	2562	200	2512,22	49,33	-1,30	,22	2395,14	2629,30	2048,74	2975,70
11	2261	200	2512,22	-251	-1,30	-1,13	2395,14	2629,30	2048,74	2975,70
12	2052	210	2611,80	-560	-1,17	-2,53	2501,98	2721,62	2150,10	3073,50
13	2959	230	2810,96	147,9	-,91	,67	2714,59	2907,34	2351,27	3269,65
14	3581	280	3308,86	272,0	-,26	1,23	3235,51	3382,22	2854,45	3763,28

Data View Variable View /

SPSS Processor is ready



6.16 pav. Pardavimų vidutinių reikšmių prognozės pasikliautiniai intervalai



6.17 pav. Pardavimų individualių reikšmių prognozės pasikliautinieji intervalai

Išvada. Pagal 6.7 lentelės arba 6.16 ir 6.17 grafikų duomenis galime prognozuoti **pardavimų vidurkių** ir **pardavimų** reikšmes fiksuotoms išlaidų reklamai reikšmėms. Pavyzdžiui, su 95% garantija galime prognozuoti, kad išleidus reklamai 20000 Lt, **pardavimų vidurkis** bus intervale nuo 239514 Lt iki 262930 Lt. (x.4 lentelė ir x.5 pav.), o **pardavimai** bus intervale nuo 204874 Lt iki 297570 Lt.

6.7.2. Kauno miesto butų kainų analizė

Pateiksime daugialypės regresinės analizės pavyzdį su tolydaus kintamojo ir psiaudokintamųjų sąveikomis. Duomenys paimti 2008 m. birželio mėn. 1 d. iš didžiausio Lietuvoje skelbimų portalo **www.aruodas.lt**. Duomenų pjūvio parametrai:

- Miestas: Kaunas.
- Mikrorajonas: Žaliakalnis, Šilainiai, Dainava.
- Namų tipai: blokinis, mūrinis.
- Būklė: pilnai įrengti ir tvarkingi.
- Skelbimai: ne senesni kaip 2 mėnesių.

Sudarysime regresinį modelį Kauno miesto butų kainų prognozavimui pagal jų plotą, namo tipą (mūrinis arba blokinis) ir miesto rajoną (Žaliakalnis arba kitas - Dainava arba Šilainiai). Analizuojamų kintamųjų vardai ir jų žymės, kintamųjų reikšmių žymės ir matavimo skalės pateiktos 6.4 lentelėje.

6.8 lentelė. Analizuojami kintamieji

Kintamojo vardas	Kintamojo žymė	Kintamojo reikšmių žymės	Matavimo skalė
Adresas	Adresas		Vardų
Rajonas	Kauno m. rajonas	1-Žaliakalnis 2-Šilainiai 3-Dainava	Vardų
Kambariu_sk	Kambarių skaičius		Santykių
Plotas	Buto plotas, kv.m		Santykių
Aukstas	Buto aukštas		Santykių
Aukstu_sk	Namo aukštų sk.		Santykių
Kaina	Buto kaina, Lt		Santykių
Murinis	Mūrinis namas	0-Blokinis 1-Mūrinis	Vardų
Zaliakalnis	Žaliakalnis	0-Dainava arba Šilainiai 1-Žaliakalnis	Vardų
LnKaina	Ln(Buto kaina)		Santykių

Sudarysime regresinį modelį su tolydžiojo kintamojo ir psiaudokintamųjų sąveikomis

$$Y = \beta_0 + \beta_1 X + \beta_2 D_1 \cdot X + \beta_3 D_2 \cdot X,$$

čia

Y- buto kainos (Lt) logaritmas (kintamasis LnKaina),

X- buto plotas kv. m. (kintamasis Plotas),

D₁- psiaudokintamasis Murinis,

D₂- psiaudokintamasis Zaliakalnis.

Kintamasis LnKaina (buto kainos logaritmas) pasirinktas todėl, kad jis geriau tinka tiesinei regresinei analizei. Prieš atliekant regresinę analizę su kintamųjų sąveikomis, reikia sukurti sąveikų kintamuosius. Mūsų atveju buvo sukurti du kintamieji:

- Murinis_plotas, kuris lygus sandaugai kintamųjų Murinis ir Plotas (sąveika D_1X),
- Zaliakalnis_plotas, kuris lygus sandaugai kintamųjų Zaliakalnis ir Plotas (sąveika D_2X).

Modelio koreguotas apibrėžtumo koeficientas pateiktas 6.9 lentelėje.

6.9 lentelė. Regresijos modelio rodiklių suvestinė

Model Summary^a

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	Change Statistics				
					R Square Change	F Change	df1	df2	Sig. F Change
1	,866 ^a	,749	,747	,205480	,749	267,203	3	268	,000

a. Predictors: (Constant), Zaliakalnis_Plotas, Buto plotas, kv.m, Murinis_Plotas

b. Dependent Variable: Ln (Buto kaina, Lt)

Išvada. Koreguoto apibrėžtumo koeficiento (Adjusted R Square) reikšmė lygi **0,747**. Galime sakyti, kad 74,7% Y sklaidos apie vidurkį galime paaiškinti Y tiesine regresija nepriklausomų nepriklausomų kintamųjų atžvilgiu, o likusi dalis yra nepaaiškinama sklaidos dalis.

6.10 lentelė. Dispersinės analizės lentelė

ANOVA^a

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	33,846	3	11,282	267,203	,000 ^a
	Residual	11,315	268	,042		
	Total	45,161	271			

a. Predictors: (Constant), Zaliakalnis_Plotas, Buto plotas, kv.m, Murinis_Plotas

b. Dependent Variable: Ln (Buto kaina, Lt)

Patikriname hipotezę apie regresijos tiesiškumą.

$$H_0: \beta_1 = \beta_2 = \beta_3 = 0,$$

$$H_a: \text{ bent vienas } \beta_j \neq 0, j=1, \dots, 3.$$

Hipotezė tikrinama naudojant Fišerio F kriterijų su dešinine kritine sritimi. Hipotezės tikrinimo rezultatai pateikti 6.10 lentelėje. H_0 atmesta ($p=0,000<0,05$) ir priimam H_a , t.y. regresija yra tiesinė, bent vienas koeficientas nelygus nuliui ir regresijos modelis bent jau iš dalies prognozėms tinka.

Randame regresijos lygties koeficientų taškinius įverčius, pasikliautinuosius intervalus ir patikriname hipotezes apie koeficientų lygybę nuliui (6.11 lentelė).

6.11 lentelė. Regresijos lygties koeficientai ir multikolinearumo statistikos.

Coefficients ^a										
		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	95% Confidence Interval for B		Collinearity Statistics	
		B	Std. Error	Beta			Lower Bound	Upper Bound	Tolerance	VIF
1	(Constant)	11,3692	,041		275,236	,000	11,288	11,450		
	Buto plotas, kv.m	,0166	,001	,713	21,426	,000	,0151	,0181	,844	1,185
	Murinis_Plotas	,0024	,001	,183	4,494	,000	,0013	,0035	,564	1,773
	Zaliakalnis_Plotas	,0021	,001	,142	3,672	,000	,0010	,0032	,623	1,605

a. Dependent Variable: Ln (Buto kaina, Lt)

Gavome imties regresijos lygtį

$$\text{LnKaina} = 11,37 + 0,0116 \cdot \text{Plotas} + 0,0024 \cdot \text{Murinis} \cdot \text{Plotas} + 0,0021 \cdot \text{Zaliakalnis} \cdot \text{Plotas}.$$

Išvada. Visi regresijos lygties koeficientai yra reikšmingi, nes atmestos nulinės hipotezės apie koeficientų lygybę nuliui:

$$H_0 : \beta_j = 0, H_a : \beta_j \neq 0, j=0, \dots, 3, (p=0,000 < 0,05),$$

Padidinus buto plotą vienu kv. m., buto kainos logaritmas $\text{Ln}(\text{Kaina})$ blokiname name (Murinis=0) Žaliakalnyje (Zaliakalnis=1) padidėja vidutiniškai $0,0116 + 0,0021 = 0,0137$ vienetais, t.y. buto kaina vidutiniškai padidėja 1019 Lt ($\text{Ln}(\text{Kaina}) = 0,0137$, $\text{Kaina} = e^{0,0137} = 1,0139$ tūkst. Lt).

Apskaičiuojame populiacijos regresijos lygties koeficientų pasikliautinuosius intervalus:

$$PI_{0,95}(\beta_0) = (11,288; 11,450),$$

$$PI_{0,95}(\beta_1) = (0,0151; 0,0181),$$

$$PI_{0,95}(\beta_2) = (0,0013; 0,0035),$$

$$PI_{0,95}(\beta_3) = (0,0010; 0,0032).$$

Išvada. Su 95% garantija galime prognozuoti, kad buto plotą padidinus vienu kv. m., blokiname name (Murinis=0), esančiame Dainavos arba Šilainių mikrorajone (Zaliakalnis=0), buto kainos logaritmo $\text{Ln}(\text{Kaina})$ vidutinis padidėjimas bus intervale nuo 0,0151 iki 0,0181, t.y. kainos vidutinis padidėjimas bus intervale nuo 1015 Lt iki 1018 Lt. (nes $\text{Ln}(\text{Kaina}) = 0,0151 \rightarrow \text{Kaina} = e^{0,0151} = 1,0152$ tūkst. Lt, $\text{Ln}(\text{Kaina}) = 0,0181 \rightarrow \text{Kaina} = e^{0,0181} = 1,0183$ tūkst. Lt). Jeigu rinksimės butą mūriniame name, tai papildomas 1 kv. m. kainos logaritmo vidutinis padidėjimas bus intervale nuo 0,0013 iki 0,0035.

Visi trys nepriklausomi kintamieji nėra multikolinearūs, nes dispersijos mažėjimo daugikliai $VIF < 4$, o sąlygojimo indeksai (Condition index) < 10 (6.12 lentelė).

6.12 lentelė. Multikolinearumo statistika (Sąlygojimo indeksas)

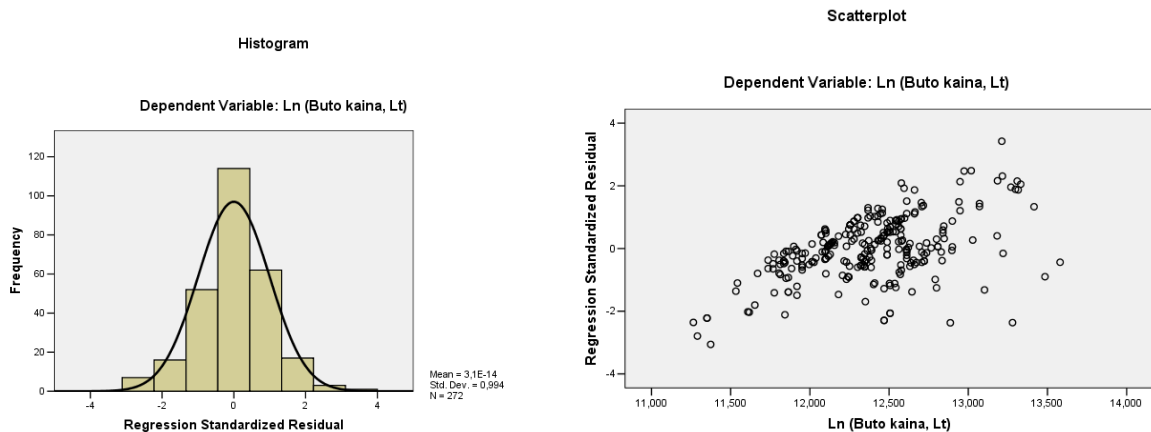
Collinearity Diagnostics ^a							
Model		Dimension	Eigenvalue	Condition Index	Variance Proportions		
					(Constant)	Buto plotas, kv.m	Murinis_Plotas
1	1		3,079	1,000	,01	,01	,03
	2		,638	2,197	,04	,02	,06
	3		,239	3,587	,01	,00	,84
	4		,043	8,416	,94	,97	,07

a. Dependent Variable: Ln (Buto kaina, Lt)

Patikriname pagrindines regresinės analizės prielaidas (6.18 pav.).

Išvada. Vizualiai palyginę, galime teigti, kad standartizuotųjų liekanų histograma yra suderinta su standartinio normaliojo skirstinio kreive (t.y. standartizuotųjų liekanų skirstinys yra suderintas su standartiniu normaliuoju skirstiniu). Standartizuotųjų liekanų reikšmių sklaidos diagrama rodo, kad sąlyginė standartizuotųjų liekanų dispersija yra pastovi. Suderinamumo hipotezė „Standartizuotųjų

liekanų skirstinys yra standartinis normalusis“ neatmesta ($p=0,102>0,05$), standartizuotųjų liekanų skirstinys yra suderintas su standartiniu normaliuoju skirstiniu. Prielaidos tenkinamos, tiesinės regresijos modelis tinka analizuojamiems duomenims.



One-Sample Kolmogorov-Smirnov Test

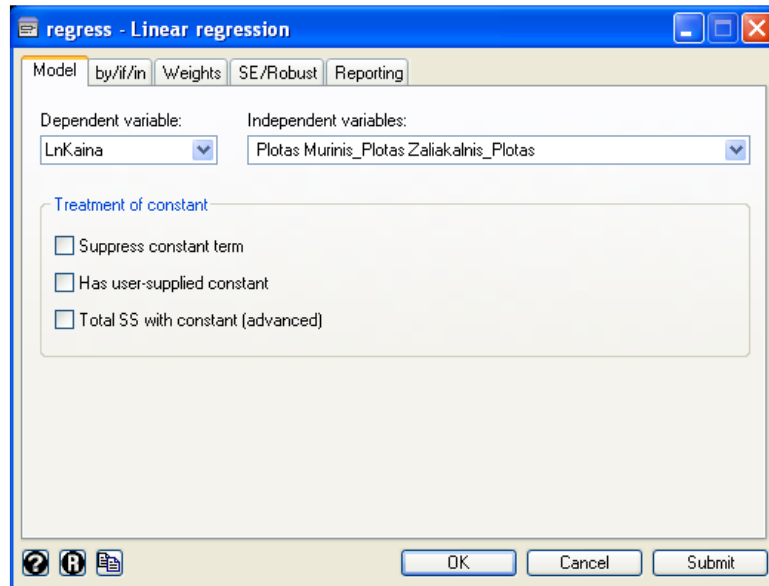
		Standardized Residual
N		272
Normal Parameters ^{a,b}	Mean	,0000000
	Std. Deviation	,99444954
Most Extreme Differences	Absolute	,074
	Positive	,040
	Negative	-,074
Kolmogorov-Smirnov Z		1,219
Asymp. Sig. (2-tailed)		,102

a. Test distribution is Normal.

b. Calculated from data.

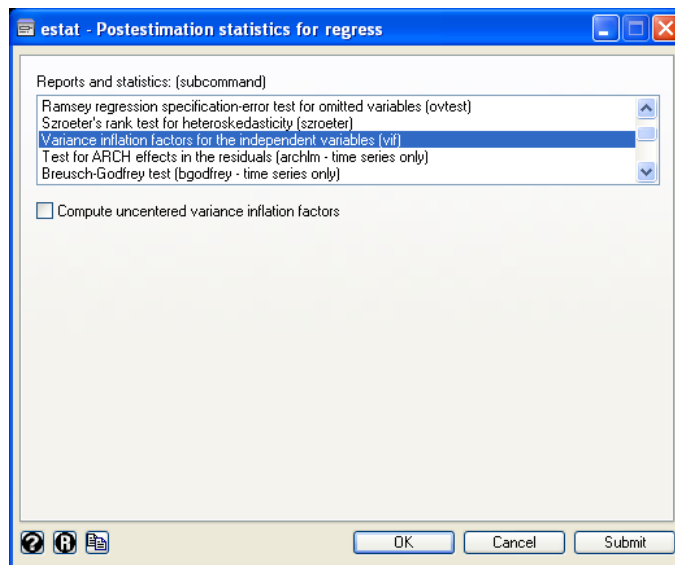
6.18 pav. Regresinės analizės prielaidų tikrinimas

Sprendžiant tą patį uždavinį pilnosios jungties metodu su STATA pasirenkame meniu: **Statistics → Linear models and related → Linear regression**. Atsidariusiame dialogo langelyje įkeliamė regresinės analizės priklausomus ir nepriklausoms kintamuosius – LnKaina, Plotas, Murinis_plotas, Zaliakalnis_plotas → OK



6.19. pav. STATA tiesinės regresinės analizės langas

Tikrinant regresinės analizės prielaidas su STATA pasirenkame meniu: **Statistics → Linear regression diagnostics → Specification tests, etc.** Atsidariusiame lange nurodome kokias prielaidas tikriname, pavyzdžiui multikolinearumą (6.20 pav.) → OK



6.20. pav. STATA regresinės analizės prielaidų tikrinimo langas

Su STATA gaui tie patys rezultatai, kaip ir su SPSS (6.21-6.22 pav.). Skiriasi tik rezultatų pateikimo tikslumas (skaitmenų skaičius po kablelio).

Source	SS	df	MS	Number of obs = 272	
Model	33.8455273	3	11.2818424	F(3, 268) =	267.20
Residual	11.3154808	268	.042221943	Prob > F =	0.0000
				R-squared =	0.7494
Total	45.1610081	271	.166645786	Adj R-squared =	0.7466
				Root MSE =	.20548

LnKaina	Coef.	Std. Err.	t	P> t	Beta
Plotas	.0166122	.0007753	21.43	0.000	.7130483
Murinis_Pl~s	.0024024	.0005345	4.49	0.000	.1829805
Zaliakaln~as	.0020841	.0005676	3.67	0.000	.1422398
_cons	11.36915	.0413069	275.24	0.000	.

6.21 pav. STATA regresinės analizės rezultatai

Variable	VIF	1/VIF
Murinis_Pl~s	1.77	0.563986
Zaliakaln~as	1.60	0.623079
Plotas	1.18	0.844138
Mean VIF	1.52	

6.22 pav. STATA multikolinearumo statistikos.

7. KLASTERINĖ ANALIZĖ

7.1. Klasifikavimo uždaviniai

Klasifikavimas - tiriamų objektų arba jų savybių jungimas klases, grupes ar tipus, pagal giminingumą, panašumą, skirtumus, atstumą, koreliacinio ryšio stiprumą ar kitus **artumo matus**. Daugelis dabartiniu metu naudojamų klasifikavimo metodų atsirado psichologų, sociologų, biologų ir matematikų bendro darbo dėka. Sprendžiamų problemų įvairovė nulėmė ir klasifikavimo metodų gausą. Klasifikavimo metodai sėkmingai taikomi rinkodaroje, analizuojant socialines ekonomines problemas, vertinant teritorinius skirtumus. Kai objektų arba savybių (požymių) skaičius didelis, derinami klasterinės ir faktorinės analizės metodai. Pavyzdžiui, marketingo specialistai klasifikavimo metodus taiko išskirdami vartotojų grupes, kuriamos taikomos atskiros marketingo programos; draudikai - skirstydami draudimo įmonių klientus į rizikos grupes pagal pareikštą pretenzijų dydžius ir t.t. Įvairūs klasifikavimo, metodai leidžia nustatyti stebimų objektų arba jų savybių (požymių) panašumą, sudaryti taisykles, pagal kurias jie rūšiuojami, vertinti atskirų parametrų įtaką daromam sprendimui. Klasifikavimo metodų sukurta daug. Dažnai pasitaiko, kad taikant tiems patiems duomenims skirtingus metodus, gaunami skirtingi rezultatai. Yra daug euristinių, teorinio pagrindimo neturinčių, metodų. Todėl, pagrindinė šių metodų taikymo problema – kaip konkrečiu atveju parinkti tinkamiausią metodą [4,8,11,15,33].

Klasifikuojant objektus arba jų savybes (požymius) taikomi parametriniai ir neparametriniai metodai. Parametriniai metodai taikomi kai populiacija yra kelių žinomų skirstinių mišinys. Mišinį sudarančių skirstinių parametrų vertinimui naudojami parametriniai metodai. Neparametriniai metodai taikomi, kai turime tik bendro pobūdžio žinias apie populiaciją (pvz. skirstinys yra tolydusis ir t.t.).

Klasių egzistavimas priklauso nuo klasifikuojamų objektų arba savybių (požymių) pasiskirstymo daugiamatėje erdvėje. Pavyzdžiui:

1. Objektų savybių (požymių) skirstinys yra daugiamatis normalusis (objektai pasiskirstę taip, kad sudaro vieną sutankėjimą - monomodalinis skirstinys). Tokiu atveju jokių relių klasių nėra, tačiau ir čia objektų sutankėjimo sritį galime išskirti į atskirą klasę.
2. Objektų savybių (požymių) skirstinys yra daugiamodalinis, t.y. egzistuoja kelios sutankėjimo sritys. Klasėmis laikomos objektų sutankėjimo sritys ir jas supančios retesnio objektų išsidėstymo sritys.

Objektyviai egzistuojančios klasės nagrinėjamoje savybių (požymių) erdvėje gali pakankamai gerai neatsiskirti. Tyrinėjant gamtos reiškinius objektų klasės atsiskiria geriau lyginant su socialiniais reiškiniiais. Gamtos moksluose klasifikacija dažniausiai pagrįsta vienu ar keliais esminiais požymiais (pvz. chemijoje - periodinė elementų sistema; biologijoje - gyvūnų ir augalų klasifikacija ir t.t.)

Problema yra optimaliaus klasių skaičiaus parinkimas. Kiekviena klasifikacija siekiama apibrėžtų tikslų, todėl objektų požymių skaičius yra **baigtinis**. Skirtingų požymių besilaikantys tyrinėtojai, **skirtingai klasifikuos** tuos pačius objektus. Bet tai neišvengiamas mokslinio pažinimo etapas.

Objektų klasifikavimas yra grupės panašių, bet neidentiškių objektų pakeitimas vienu „apibendrintu“ „tipišku“ tai klasei objektu. Todėl klasifikavimas (grupavimas) visada susijęs su tam tikros dalies **informacijos praradimu**. Išryškėja tam tikras **prieštaravimas - kuo daugiau klasių (grupių, tipų) tuo mažiau prarandame informacijos**. Todėl optimalia laikoma tokia klasifikacija, kurią

taikant gaunamas minimalus klasių skaičius su sąlyga, kad informacijos praradimas neviršija iš anksto nustatyto dydžio.

Yra mažiausiai trys priežastys, kurios leidžia sumažinti požymių erdvės matavimą minimaliai prarandant informaciją:

- informacijos dubliavimas (tarp požymių yra stiprus koreliacinis ryšys);
- mažas požymių informatyvumas (maža požymio reikšmių sklaida populiacijoje);
- galima kelis požymius apjungti (pvz., vieną požymį išreikšti kitų požymių tiesiniu dariniu).

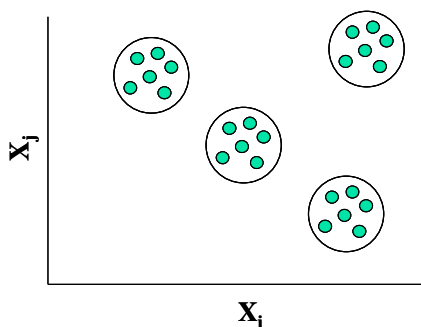
Pavyzdžiui, žmogaus rūbai turi du arba tris požymius (dydis, ūgis, talijos apimtis). Pagal tuos tris požymius gauname pakankamai gerą žmonių klasifikaciją. Gera siuvėja matuoja 11 požymių.

Formalus klasifikavimo teisingumo patikrinimas pagrįstas statistinėmis imties savybėmis. Pavyzdžiui, dvi nepriklausomos atsitiktinės imtys imamos iš tos pačios populiacijos ir klasifikuojamos nepriklausomai viena nuo kitos, o paskui palyginami gauti rezultatai.

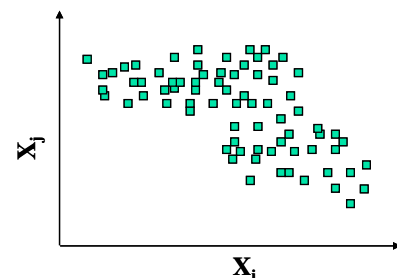
Klasifikavimo metodai skirstomi į dvi pagrindines klases - **klasifikavimą su mokytoju ir be mokytojo**. Klasifikavimo su mokytoju metodai (tiesinė, kvadratinė, branduolinė, artimiausių kaimynų diskriminantinė analizė, logistinė regresinė analizė, neuroniniai tinklai, klasifikavimo medžiai) taikomi tada, kai egzistuoja mokomoji imtis, t.y. informacija apie dalies tiriamų objektų (savybių) priklausomybę konkrečiai klasei.

Klasterinė analizė - tai metodas skirtas identifikuoti homogenines objektų arba savybių grupes (klasterius), kai nėra informacijos apie dalies tiriamų objektų (savybių) priklausomybę konkrečiai klasei. Objektai (savybės) suskirstomi į klasterius taip, kad skirtumai klasterių viduje būtų kuo mažesni, o tarp klasterių - kuo didesni. Panašūs objektai priklauso tam pačiam klasteriui, nutolę objektai - skirtingiems klasteriams. Pavyzdžiui, valstybės gali būti suskirstytos į homogenines grupes pagal biudžeto ir ekonomikos rodiklius, turizmo rinka - pagal įvairius paklausos rodiklius ir t. t.

Pats paprasčiausias būdas suvokti dviejų savybių, kurių duomenys sudaro keletą homogeninių grupių, skirstymo į klasterius reiškinį, yra nubraižyti šių dviejų savybių sklaidos (Scatterplot) diagramą. 7.1 pav. pavaizduota situacija, kada duomenys gerai tinka klasifikacijai. Tačiau realiai tokia galimybė pasitaiko labai retai - paprastai klasterių struktūros nėra aiškiai atskirtos ir persidengia (7.2 pav.), o kintamųjų skaičius žymiai viršija du, t. y. analizė atliekama p-matėje erdvėje.



7.1 pav. Duomenys gerai tinka objektų klasifikacijai pagal savybes X_i ir X_j



7.2 pav. Praktikoje dažniausiai pasitaikanti situacija klasifikuojant objektus pagal savybes X_i ir X_j

Skiriamos dvi pagrindinės klasterinės analizės metodų grupės- **hierarchiniai (hierarchical) ir skaidymo (partitional)**. Toliau nagrinėsime tik *hierarchinius klasterizavimo* metodus. Hierarchiniais metodais nustatoma bendra visų klasterių tarpusavio priklausomybių struktūra ir tik tada sprendžiama, koks klasterių skaičius optimalus. Savo ruožtu hierarchiniai metodai skirstomi į jungimo ir skaidymo metodus. *Skaidymo (partitional) klasterizavimo* metodai paprastai taikomi tada, kai iš anksto žinomas klasterių skaičius ir norima tiriamus objektus (savybes) klasterizuoti. Mes nagrinėsime tik nepersidengiančių klasterių metodus, t.y., kai objektas (savybė) priskiriama tik vienam klasteriui. Be jų yra persidengiančių klasterių metodai, kurie gali priskirti objektą (savybę) keliems klasteriams [33].

7.2. Klasterinės analizės modelis

Tarkime, kad reikia klasifikuoti objektus (arba jų savybes), kai nėra mokomųjų imčių. **Duota:** n objektų $O_1, O_2, \dots, O_n$, kurių savybes nusako matrica “**objektai – savybės**”

$$X = \begin{pmatrix} x_1^{(1)} & x_2^{(2)} & \dots & x_n^{(p)} \\ x_2^{(1)} & x_2^{(2)} & \dots & x_n^{(p)} \\ \vdots & \vdots & & \vdots \\ x_n^{(1)} & x_n^{(2)} & \dots & x_n^{(p)} \end{pmatrix}$$

arba **artumų** matrica

$$\rho = \begin{pmatrix} \rho_{11} & \rho_{12} & \dots & \rho_{1m} \\ \rho_{21} & \rho_{22} & \dots & \rho_{2m} \\ \dots & \dots & \dots & \dots \\ \rho_{m1} & \rho_{m2} & \dots & \rho_{mm} \end{pmatrix},$$

čia $x_i^{(k)}$ – objekto O_i k – toji savybė, ρ_{ij} – artumas tarp objektų (savybių). Matrica “**objektai – savybės**” transformuojama į **artumų matricą** nurodžius artumo matą, pagal kurį apskaičiuojamas artumas tarp objektų (savybių). Klasifikavimo procese naudojamos tokios sąvokos, kaip panašumas, nepanašumas, skirtumas, atstumas, giminingumas, homogeniškumas. Taigi klasifikavimo metodai apima tam tikrą matų sistemą objektų (savybių) tarpusavio santykiams įvertinti. Priklausomai nuo to, pagal kokią skalę išmatuoti analizuojami kintamieji, naudojami skirtingi artumo matai. Kai $m=n$ poromis lyginami objektai, kai $m=p$ poromis lyginamos savybės.

Artumų matricos elementais gali būti:

- objektų O_i ir O_j panašumų arba nepanašumų matai;
- intensyvumai kokiame nors procese (pavyzdžiui i -tos pramonės šakos produkcijos srautas į j -ąją šaką);
- geometriniai atstumai tarp objektų O_i ir O_j ;
- pirmumo sąryšis (pavyzdžiui, $\rho_{ij} = 1$,jeigu objektas O_i yra ne blogesnis už O_j , $\rho_{ij} = 0$ priešingu atveju);
- savybių X_i ir X_j koreliacijos koeficientai.

Klasterinės analizės uždavinys - analizuojamą objektų aibę $O = \{O_i\}, i = 1, \dots, n$ užduotą matricas X arba ρ suskaidyti į nedidelį skaičių (iš anksto žinomą arba nežinomą) homogeninių klasių (klasterių). Klasterinėje analizėje pirmiausia reikia apibrėžti objektų homogeniškumo sąvoką. Tai

sunkus uždavinys, nes jam spręsti nėra formalių metodų. Paprastai artumų matricoje elementai ρ_{ij} apibūdina arba atstumą tarp objektų $d(O_i, O_j)$ arba įvertina jų panašumo (nepanašumo) laipsnį $r(O_i, O_j)$ daugiamatėje erdvėje.

Apibrėžiant panašumo laipsnį reikia neužmiršti akivaizdžių reikalavimų:

- $r(O_i, O_j) = r(O_j, O_i)$ (simetriškumas, t.y. panašumas neturi priklausyti nuo to, kuris iš lyginamų objektų imamas pirmuoju, kuris -antruoju),
- $r(O_i, O_i) = \max r(O_i, O_j)$ (objektas labiausiai panašus į save).

Artumo tarp objektų mato parinkimas yra vienas iš pagrindinių klasterinės analizės uždavinių. Nuo jos pasirinkimo priklauso ir galutinis objektų grupavimo į klasterius variantas. Artumo mato parinkimo uždavinys sprendžiamas kiekvienu atveju atskirai ir priklauso nuo:

- klasifikavimo tikslo;
- fizinės ir statistinės stebimų objektų (savybių) prigimties;
- pradinių duomenų apie objektus (savybes) pilnumo;
- X tikimybinio skirstinio.

Kartais tikslinga matricos X reikšmes transformuoti. Pavyzdžiui, jeigu kintamųjų kitimo skalės yra labai skirtingos, jie dažniausiai standartizuojami taip, kad vidurkiai būtų lygūs nuliui, o dispersijos –vienetui. Tiesa, tokia procedūra gali sumažinti klasterių (klasių) skirtumus. Kitas kintamųjų transformavimo būdas – svorių suteikimas. Taigi kintamųjų transformavimas, leidžia jiems vaidinti didesnę arba mažesnę vaidmenį, matuojant artumą tarp objektų (savybių). Svoriai dažniausiai parenkami aprioriškai, tačiau tai reikia gerai teoriškai pagrįsti. Dauguma statistikos paketų turi kintamųjų transformavimo galimybę. Pavyzdžiui SPSS tai galima atlikti meniu: Analyze→ Classify → Hierarchical Cluster → Method → Transform values.



7.3 Klasterinės analizės etapai

7.3. pav. pateikti pagrindiniai klasterinės analizės etapai, kuriuos detaliau aptarsime sekančiuose skyreliuose.

7.3. Artumo matai

7.3.1. Artumo matai intervalų ir santykių skalėse

Kai objektus apibūdinantys požymiai matuojami pagal intervalų arba santykių skalę dažnai naudojami metriniai artumo matai (pvz. SPSS skaičiuoja 6 matus naudoja. Šiuos matus galima būtų vadinti skirtingumo matais — kuo didesnė reikšmė, tuo objektai mažiau panašūs. Nevienodai skirtingai matuojamų požymių įtakai išvengti naudojamos šių kintamųjų standartizuotosios reikšmės (z-reikšmės). Intervalinių kintamųjų panašumo matas gali būti ir koreliacijos koeficientas. Metriniai atstumo matai pasižymi šiomis savybėmis:

$$\begin{aligned}d(i, j) &\geq 0, \\d(i, i) &= 0, \\d(i, j) &= d(j, i), \\d(i, j) &\leq d(i, k) + d(k, j).\end{aligned}$$

Pateiksime dažniausiai praktikoje naudojamus metrinis atstumo matus:

1. **Euklido atstumas** (Euclidean distance). Euklido atstumas tarp taško X_i ir X_j — tai trumpiausias atstumas tarp šių objektų (taškus p -matėje erdvėje). Plokštumoje arba trimatėje erdvėje — tai objektus (taškus) jungianti tiesė. Bendru p -matės erdvės atveju šis atstumas apskaičiuojamas pagal formulę

$$d_E(i, j) = \sqrt{\sum_{s=1}^p (x_i^{(s)} - x_j^{(s)})^2}.$$

2. **Euklido atstumas su svoriais**

$$d_{ES}(i, j) = \sqrt{\sum_{s=1}^p \omega_s (x_i^{(s)} - x_j^{(s)})^2},$$

čia $0 \leq \omega_s \leq 1$, $s = \overline{1, p}$, ω_s - svorio koeficientas, kurio dydis proporcingas vektoriaus X komponentės $x^{(s)}$ reikšmingumui, skirstant objektus į klases.

3. **Euklido atstumo kvadratas** (Squared Euclidean distance)

$$d_{E2}(i, j) = \sum_{s=1}^p (x_i^{(s)} - x_j^{(s)})^2.$$

Šiuo matu tiksliau įvertinami dideli skirtumai. Euklido atstumo kvadratas turi būti naudojamas pasirinkus centrų arba Vordo (Ward's) klasterių sudarymo metodus.

4. **Čebyševio atstumas** (Chebychev distance)

$$d_{CH}(i, j) = \max_{1 \leq s \leq p} |x_i^{(s)} - x_j^{(s)}|.$$

Atstumu tarp dviejų objektų yra laikoma didžiausia absoliuti skirtumo tarp visų šių objektų porų reikšmė.

5. **Manhateno atstumas** (City – block or Manhattan distance)

$$d_{MH}(i, j) = \sum_{s=1}^p |x_i^{(s)} - x_j^{(s)}|.$$

7.3.2. Dvireikšmių vektorių artumo matai

Tarkime, kad objektas O_i gali turėti p savybių. Žymėsime $x_i^{(s)} = 1$, jeigu jis s-ją savybę turi ir $x_i^{(s)} = 0$, jeigu jos neturi. Tuomet objektą O_i apibūdina vektorius

$$X_i = (x_i^{(1)}, x_i^{(2)}, \dots, x_i^{(p)}), \text{ čia } x_i^{(s)} \in \{0,1\}.$$

Sudarant dviejų objektų panašumo matus yra naudojami porų $(x_i^{(s)}, x_j^{(s)})$ reikšmių dažniai, kuriuos patogiu pateikti lentelėje

$O_i \setminus O_j$	0	1
0	$v_{ij}^{(00)}$	$v_{ij}^{(01)}$
1	$v_{ij}^{(10)}$	$v_{ij}^{(11)}$

čia $v_{ij}^{(00)}$ – porų (0;0) dažnis, t.y. savybių, kurių neturi abu objektai dažnis (skaičius);

$v_{ij}^{(11)}$ – porų (1;1) dažnis, t.y. savybių, kurias turi abu objektai dažnis (skaičius) ir t.t.

Aišku, kad $v_{ij}^{(00)} + v_{ij}^{(01)} + v_{ij}^{(10)} + v_{ij}^{(11)} = p$. Jeigu visos p objekto savybės yra vienodai svarbios ir vertinant objektų panašumą nulių sutapimo (savybės neturėjimo) reikšmė yra tokia pati kaip ir vienetukų sutapimo, tai objektams, apibūdinamiems dvireikšmiaisiais vektoriais, galime įvesti paprastą paprastą suderinamumo matą (Simple matching)

$$r(i, j) = \frac{v_{ij}^{(00)} + v_{ij}^{(11)}}{v_{ij}^{(00)} + v_{ij}^{(11)} + v_{ij}^{(01)} + v_{ij}^{(10)}},$$

kuris parodo kokia dalis objektų O_i ir O_j savybių sutampa. Objektai yra tuo panašesni, kuo daugiau sutampančių koordinatų turi vektoriai ir atvirkščiai.

Pavyzdys. Duota: $p = 10$, $O_1 = (0,0,1,1,0,0,1,1,0,1)$, $O_2 = (0,1,1,1,0,0,0,0,0,1)$. Sudarome dažnių lentelę ir apskaičiuojame parastą suderinamumo matą.

$X_1 \setminus X_2$	0 (NE)	1 (TAIP)
0 (NE)	4	1
1 (TAIP)	2	3

Abu objektai O_1 ir O_2 panašūs pagal 7 savybes (abu neturi 4 savybių $v_{ij}^{(00)} = 4$, ir turi 3 tas pačias savybes $v_{ij}^{(11)} = 3$). Pagal likusias 3 savybes jie yra nepanašūs (vienas objektas turi savybę, o kitas ne). Apskaičiuojame parasto suderinamumo mato reikšmę

$$r(1,2) = \frac{4+3}{10} = 0,7.$$

Dvireikšmių vektorių panašumo/nepanašumo matų yra labai daug. Pavyzdžiui SPSS skaičiuoja 27 panašumo/nepanašumo matus. Čia pateiksime kelis iš jų:

1. **Euklido atstumas** (Euclidian distance)

$$r_E(i, j) = \sqrt{v_{ij}^{(01)} + v_{ij}^{(10)}} \quad , \quad r_E(i, j) \geq 0$$

yra intervalų ir santykių skalės Euklido atstumo analogas dvireikšmiams kintamiesiems.

2. **Euklido atstumo kvadratas** (Squared Euclidian distance)

$$r_{E2}(i, j) = v_{ij}^{(01)} + v_{ij}^{(10)} \quad , \quad r_{E2}(i, j) \geq 0.$$

3. **Didumų skirtumo** (Size difference)

$$r_{SD}(i, j) = \frac{(v_{ij}^{(01)} - v_{ij}^{(10)})^2}{(v_{ij}^{(00)} + v_{ij}^{(11)} + v_{ij}^{(01)} + v_{ij}^{(10)})^2} \quad ,$$

matuoja savybių nesutapimų asimetriją, $r_{SD}(i, j) \geq 0$.

4. **Struktūrų skirtumo** (Pattern difference)

$$r_{PD}(i, j) = \frac{v_{ij}^{(01)}v_{ij}^{(10)}}{(v_{ij}^{(00)} + v_{ij}^{(11)} + v_{ij}^{(01)} + v_{ij}^{(10)})^2} \quad , \quad 0 \leq r_{PD}(i, j) \leq 1.$$

Dvireikšmių vektorių panašumo matai (Similarity measures for Binary Data):

1. **Paprasčiausias suderinamumo matas** (Simple matching)

$$r(i, j) = \frac{v_{ij}^{(00)} + v_{ij}^{(11)}}{v_{ij}^{(00)} + v_{ij}^{(11)} + v_{ij}^{(01)} + v_{ij}^{(10)}} \quad ,$$

parodo kokia dalis objektų O_i ir O_j savybių sutampa.

2. **Raselo ir Rao panašumo matas** (Russel and Rao)

$$r_{RR}(i, j) = \frac{v_{ij}^{(00)}}{v_{ij}^{(00)} + v_{ij}^{(11)} + v_{ij}^{(01)} + v_{ij}^{(10)}} \quad ,$$

parodo kokią dalį sudaro savybės, kurių neturi abu objektai O_i ir O_j .

3. **Julo Q panašumo matas** (Yule Q similarity measure). Julo Q dar vadinamas keturlaukės koreliacijos koeficientu.

$$r_Q(i, j) = \frac{v_{ij}^{(11)}v_{ij}^{(00)} - v_{ij}^{(01)}v_{ij}^{(10)}}{v_{ij}^{(11)}v_{ij}^{(00)} + v_{ij}^{(01)}v_{ij}^{(10)}} \quad , \quad -1 \leq r_Q(i, j) \leq 1.$$

Kuo $r_Q(i, j)$ absoliutiniu didumu arčiau vieneto tuo objektai yra panašesni. Pastaba. $r_Q(i, j)$ nenaudojamas kai bent vienas dažnis yra lygus 0, nes tuomet $|r_Q(i, j)| = 1$, tačiau tai nereiškia visiško objektų panašumo.

7.4. Klasterių sudarymo metodai

Pažymėkime S_i i -tą objektų klasterį (objektų, kurie priklauso klasteriui S_i , numerių aibė), n_i – objektų skaičių klasteryje S_i ; $\bar{S}(i)$ – i -to klasterio savybių reikšmių vidurkių vektorius (i -to klasterio „svorio centrą“), $\rho(S_l, S_m)$ – atstumą tarp klasterių S_l ir S_m .

Dažniausiai naudojamos šie klasterių sudarymo metodai:

1. **Vienetinės jungties arba „Artimiausio kaimyno“** (Single linkage or „Nearest neighbor“) Atstumas tarp klasterių yra apibrėžiamas kaip atstumas tarp priklausančių skirtingiems klasteriams dviejų artimiausių objektų

$$\rho_{\min}(S_l, S_m) = \min_{i \in S_l, j \in S_m} d(i, j).$$

2. **Pilnosios jungties arba „Tolimiausio kaimyno“** (Complete linkage or „Furthest neighbor“). Atstumas tarp klasterių yra apibrėžiamas kaip atstumas tarp priklausančių skirtingiems klasteriams dviejų artimiausių objektų arba savybių

$$\rho_{\max}(S_l, S_m) = \max_{i \in S_l, j \in S_m} d(i, j).$$

3. **Vidutinės jungties** (Between-groups linkage, Unweighted pair-group method using averages) Vidutinis atstumas tarp visų galimų dviejų klasių objektų porų. Atstumas tarp klasterių yra lygus atstumo tarp visų galimų stebėjimų porų, kai vienas stebėjimas yra imamas iš vieno klasterio, o kitas - iš antro, vidurkiui. Metoda pranašesnis už artimiausių ir tolimiausio kaimyno metodus, nes naudoja informaciją apie visas dviejų klasterių objektų poras, o ne tik apie artimiausius arba tolimiausius kaimynus

$$\rho_{\text{vid}}(S_l, S_m) = \frac{1}{n_l n_m} \sum_{i \in S_l} \sum_{j \in S_m} d(i, j).$$

4. **Centrų metodas (Centroid)**. Matuojamas atstumas tarp klasterių „sunkio centrų“

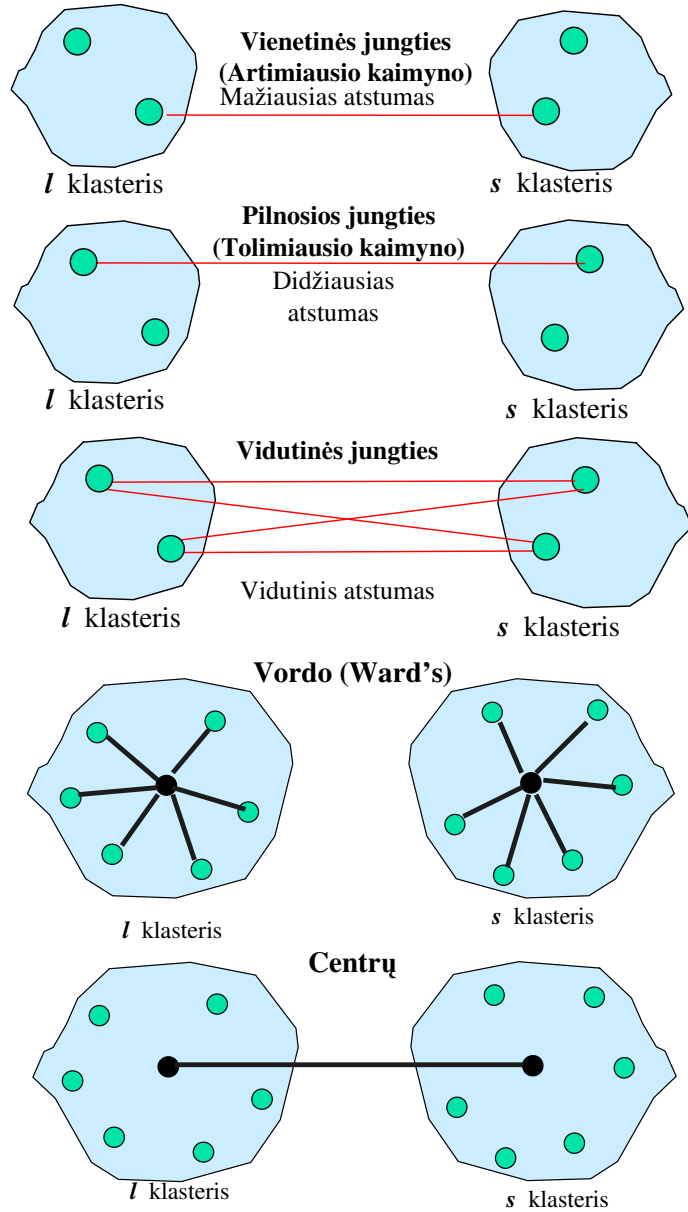
$$\rho_c(S_l, S_m) = d(\bar{S}(l), \bar{S}(m)).$$

5. **Vordo** (Ward's method). Dar vienas populiarius hierarchinės klasterizacijos metodas yra Vordo metodas. Jis remiasi klasterių vidinės dispersijos minimizavimo principu. Čia naudojama tikslo funkcija – klasterio vidinių kvadratinių nuokrypių suma. Kiekvienu žingsniu jungiami tie objektai arba klasteriai, kuriems nuokrypių suma įgyja mažiausią reikšmę. Taikant Vordo (Ward's) metodą, imama klasterių pora ir apskaičiuojama savybių vektorių nuokrypių nuo jungtinio klasterio centro kvadratų suma. Analogiškai daroma su kita klasterių pora ir t. t. Jungiami tie klasteriai, kuriuos sujungus nuokrypių suma yra mažiausia.

Iš visų aprašytųjų metodų dažniausiai yra naudojamas - vidutinės jungties metodas (Between-groups linkage). Dažnai naudojamas ir Vordo metodas (Ward's method), tačiau taikant šį metodą gaunama daugiau (mažesnių) klasterių, metodo rezultatams didelę įtaką daro objektai su savybių išskirtimis. Apskritai išskirtys iškraipo visų klasterizavimo metodų rezultatus, todėl prieš klasterizuojant pageidautina jas pašalinti. Objektams klasterizuoti rekomenduojama taikyti keletą klasterizavimo metodų - tada galima tikėtis patikimesnių rezultatų. Geometrinė klasterių sudarymo metodų interpretacija pateikta 7.4 pav.

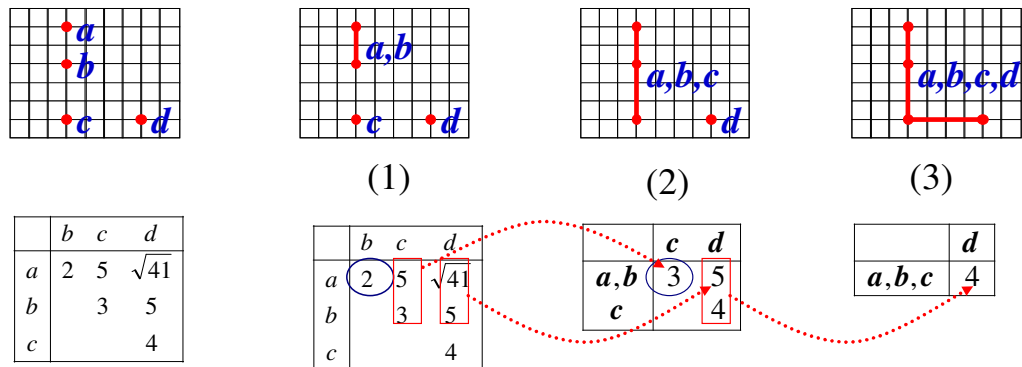
Pateiksime keturių objektų **a**, **b**, **c**, ir **d** jungimo į klasterius pavyzdį panaudojant vienetinės jungties ir pilnosios jungties metodus. Analizuosime dvimatį atvejį. Objektų **a**, **b**, **c**, ir **d** išsidėstymas plokštumoje ir Euklido atstumų tarp jų matrica pateikta 7.5. pav. ir 7.6. pav. kairėje pusėje. Šiuose paveiksluose taip pat pavaizduoti visi trys hierarchinio jungimo algoritmo etapai ir atstumų matricos transformavimo rezultatai. Naudojant vienetinės jungties metodą pirmiausia sujungiami

objektai **a** ir **b**, po to prie jų prijungiamas objektas **c** ir paskutiniame žingsnyje prijungiamas objektas **d**. Jungiant pilnosios jungties metodu pirmiausia sujungiami objektai **a** ir **b**, po to **c** ir **d** ir paskutiniame žingsnyje abu anksčiau gauti klasteriai apjungiami į vieną klasterį. Gautų dendogramų palyginimas pateiktas 7.7 pav.



7.4 pav. Atstumai tarp klasterių naudojant skirtingus klasterių sudarymo metodus

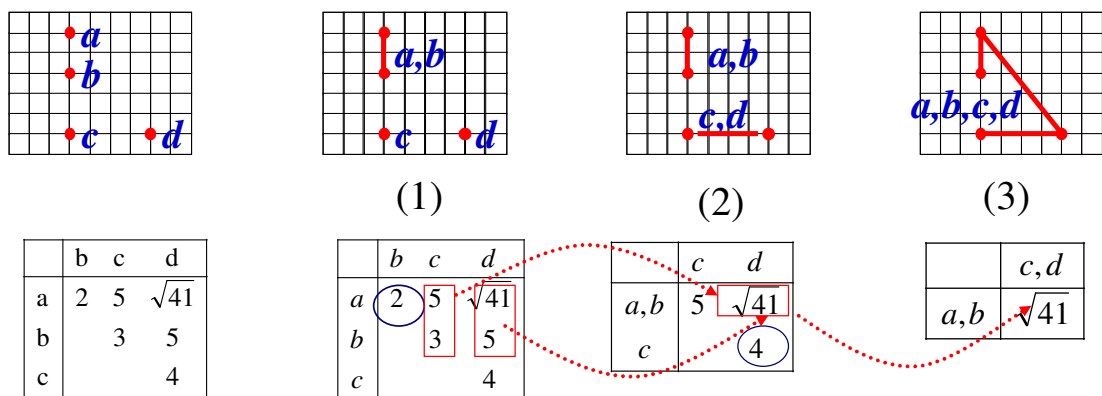
Euklido atstumas



Atstumų matrica

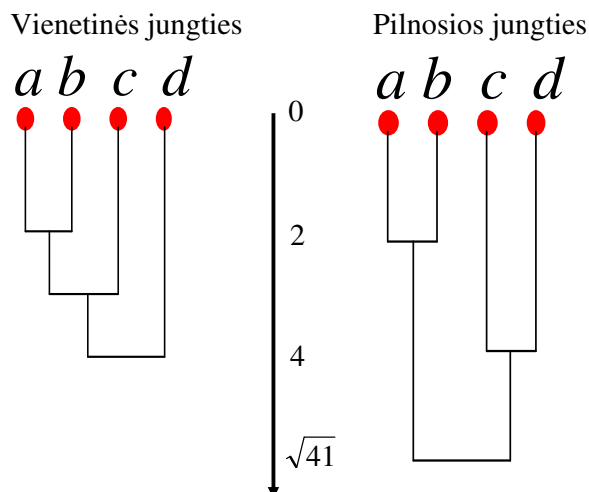
7.5 pav. Klasterių sudarymo vieninės jungties metodu pavyzdys

Euklido atstumas



Atstumų matrica

7.6 pav. Klasterių sudarymo pilnosios jungties metodu pavyzdys



7.7 pav. Vieninės ir pilnosios jungties metodais gautų dendogramų palyginimas

7.5. Klasifikavimo kokybės kriterijai

Klasifikavimo kokybės kriterijai naudojami lyginant skirtingus klasifikavimo algoritmus. Kaip palyginti du skirtingus tų pačių objektų (savybių) sugrupavimus į klasterius? Norint atsakyti į šį klausimą reikia apibrėžti klasifikavimo kokybės kriterijus. Kai klasterių skaičius žinomas iš anksto ir jis lygus k , dažniausiai naudojami šie kriterijai:

a) klasterių vidinių dispersijų suma

$$Q_1(S) = \sum_{l=1}^k \sum_{i \in S_l} d^2(i, \bar{S}(l))$$

b) atstumų tarp klasių elementų kvadratų suma

$$Q_2(S) = \sum_{l=1}^k \sum_{\substack{i \in S_l \\ j \in S_l}} d^2(i, j)$$

arba atstumų tarp klasterių elementų kvadratų vidurkių suma

$$Q_2^*(S) = \sum_{l=1}^k \frac{1}{n_l} \sum_{\substack{i \in S_l \\ j \in S_l}} d^2(i, j),$$

c) apibendrinta klasterių vidinė dispersija

$$Q_3(S) = \det \left(\sum_{l=1}^k n_l W_l \right),$$

čia W_l – klasės S_l empirinė kovariacinė matrica. Jos elementai apskaičiuojami pagal formulę

$$\omega_{qm}(l) = \frac{1}{n_l} \sum_{i \in S_l} (x_i^{(q)} - \bar{x}^{(q)}(l))(x_i^{(m)} - \bar{x}^{(m)}(l)),$$

čia $q = \overline{1, p}$, $m = \overline{1, p}$, $\bar{x}^{(m)}(l)$ – klasės S_l m -tos savybės vidurkis.

Klasifikavimo kokybės kriterijai, kai klasių skaičius nežinomas dažniausiai sudaromi iš dviejų dalių

$$Q(S) = I_1(S) + I_2(S),$$

čia $I_1(S)$ yra nedidėjanti funkcija, priklausanti nuo klasterių skaičiaus $k(S)$ ir dažniausiai charakterizuoja vidinę klasterių sklaidą. Antra funkcija $I_2(S)$ yra nemažėjanti klasių skaičiaus $k(S)$ funkcija. Jis gali būti įvairiai interpretuojamas. Pavyzdžiui, baudų dydis už nedidelius klasterius:

$$I_1(S) = \sum_{l=1}^{k(S)} \sum_{x \in S_l} d(X_i, \bar{X}(l)) \text{ ir}$$

$$I_2(S) = ck(S),$$

$$I(S) = I_1(S) + I_2(S),$$

čia $k(S)$ – klasių skaičius, esant suskaidymui S , c – konstanta, $c > 0$.

7.6. Hierarchiniai klasifikavimo algoritmai

Hierarchiniai metodai skirstomi į jungimo ir skaidymo metodus. Taikant jungimo metodus, iš pradžių visi objektai (savybės) traktuojami kaip atskiri klasteriai. Pirmuoju žingsniu du objektai (savybės) yra sujungiami į klasterį, kiekvienu kitu žingsniu naujas objektas (savybė) yra jungiamas prie esamo klasterio arba sujungiami du anksčiau sudaryti klasteriai. Suformuotas klasteris vėliau jau negali būti skaidomas — jis gali būti tik jungiamas su kitais klasteriais. Veiksmai kartojami tol, kol lieka vienas klasteris. Tyrėjas pats sprendžia, kuriuo etapu objektų arba savybių paskirstymas į klasterius yra optimalus. Skaidymo metodai yra loginė jungimo metodų priešingybė — vienintelis pradinis klasteris, kuriame yra visi objektai (savybės) nuosekliai skaidomas į dalis.

Skaidymo algoritmai. Objektų aibė skaidoma į dvi klases, po to, tas klases vėl į dvi ir t.t.

$$(S_1^*, S_2^*) = \arg \max_{S_1 \cup S_2 = S} \rho(S_1, S_2)$$

čia S_1 ir S_2 aibės S poaibiai.

Jungimo algoritmai. Pradinis skaidinys

$$S^{(0)} = (S_1^{(0)}, \dots, S_n^{(0)}), \text{ čia } S_i^0 = \{O_i\}$$

k – lygio skaidinys

$$S^{(k)} = (S_1^{(k)}, \dots, S_{n-k}^{(k)}),$$

gaunamas iš $S^{(k-1)}$, $k \geq 1$ skaidinio apjungus klasių porą (S_1^*, S_2^*)

$$(S_1^*, S_2^*) = \arg \min_{\substack{S_1 \neq S_2 \\ S_1, S_2 \in S^{(k-1)}}} \rho(S_1, S_2).$$

Galutinę hierarchiją sudaro įdėtų skaidinių sistema $S^{(0)} \subset S^{(1)} \subset \dots \subset S^{(n-1)} \equiv O$, kurią galima atvaizduoti grafiškai. Šio proceso schema vaizduojama grafiku vadinamu **dendograma**. Vienoje dendogramos ašyje atidedami objektų numeriai, kitoje atstumai tarp jų. Objektai į klasterius jungiami laužtėmis. Tyrėjas analizuodamas dendogramą vizualiai, nusprendžia, kuriame žingsnyje objektų paskirstymas į klasterius yra optimalus. Tiksliau optimalus klasterių skaičius parenkamas panaudojus klasifikavimo kokybės kriterijus (žiūr. 7.5 skyrelį).

7.7. Klasterinės analizės pavyzdys

Tarkime, kad automobilių gamintojai nori išsiaiškinti jų gaminamų automobilių konkurencines savybes. Šio uždavinio sprendimui galima panaudoti klasterinę analizę. SPSS faile car_sales.sav pateiktos parduodamų automobilių charakteristikos (7.1 lentelė). Iš SPSS failo car_sales.sav suformuotas duomenų pjūvis (failas automobiliu pardavimai.sav, 7.8 pav.), kuriame įrašyti duomenys apie gerai parduodamus automobilius (pardavimai viršija 100 tūkst. vienetų). Panaudoję hierarchinius klasterizavimo metodus suklausifikuosime geriausiai parduodamus automobilius pagal kainą ir technines charakteristikas.

Automobiliu pardavimai.sav - SPSS Data Editor

File Edit View Data Transform Analyze Graphs Utilities Window Help

11 : mpg 27

	manufact	model	sales	price	engine_s	horsepow	wheelbas	width	length	curb_wgt	fuel_cap	mpg
1	Chevrolet	Cavalier	145,52	13,260	2,2	115	104,1	67,9	180,9	2,676	14,3	27
2	Chevrolet	Malibu	135,13	16,535	3,1	170	107,0	69,4	190,4	3,051	15,0	25
3	Chevrolet	Impala	108,00	18,890	3,4	180	110,5	73,0	200,0	3,389	17,0	27
4	Ford	Mustang	113,37	21,560	3,8	190	101,3	73,1	183,2	3,203	15,7	24
5	Ford	Taurus	245,82	17,885	3,0	155	108,5	73,0	197,6	3,368	16,0	24
6	Ford	Focus	175,67	12,315	2,0	107	103,0	66,9	174,8	2,564	13,2	30
7	Honda	Civic	199,69	12,885	1,6	106	103,2	67,1	175,1	2,339	11,9	32
8	Honda	Accord	230,90	15,350	2,3	135	106,9	70,3	188,8	2,932	17,1	27
9	Pontiac	Grand A	131,10	19,720	3,4	175	107,0	70,4	186,3	3,091	15,2	25
10	Toyota	Corolla	142,54	13,108	1,8	120	97,0	66,7	174,0	2,420	13,2	33
11	Toyota	Camry	247,99	17,518	2,2	133	105,2	70,1	188,5	2,998	18,5	27

Data View Variable View /

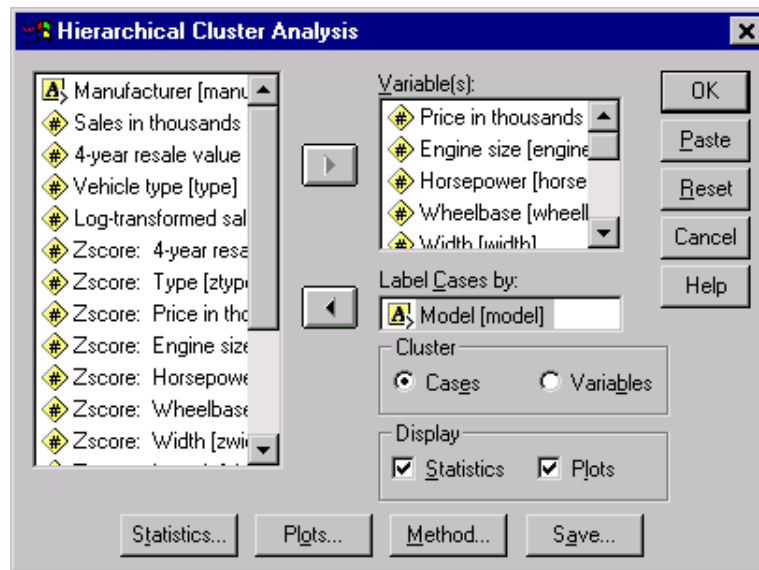
SPSS Processor is ready

7.8 pav. Failo automobilių pardavimai.sav duomenys

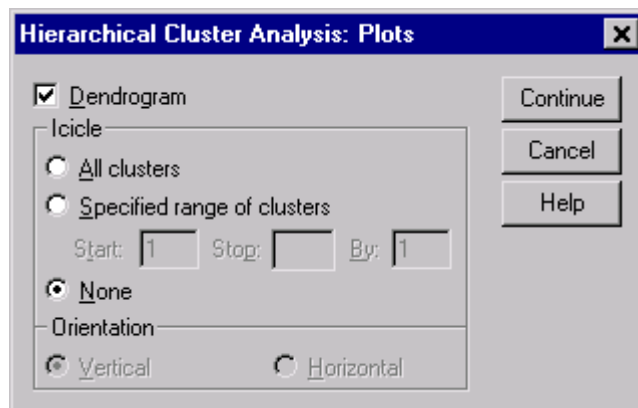
7.1 lentelė. Analizuojami kintamieji

Kintamųjų vardai	Kintamųjų žymės	Matavimo skalė
manufact	Gamintojas	Vardų
model	Modelis	Vardų
sales	Pardavimai (tūkst. vnt.)	Santykių
price	Kaina (tūkst. dolerių)	Santykių
engine_s	Variklio tūris	Santykių
horsepow	Galingumas (AJ)	Santykių
wheelbas	Važiuklės bazė	Santykių
width	Plotis	Santykių
length	Ilgis	Santykių
curb_wgt	Automobilio masė (be keleivių)	Santykių
fuel_cap	Degalų bako talpa (galonais)	Santykių
mpg	Degalų sąnaudos (mylios/galonas)	Santykių

Išspręsimе uždavinį vienietinės jungties (artimiausio kaimyno) metodu. Klasterinei analizei skirtas SPSS meniu: **Analyze → Classify → Hierarchical Cluster...** Atsidariusiame lange **Hierarchical Cluster Analysis** įkeliame analizuojamus kintamuosius į laukelį **Variable(s)**. Klasifikuojant objektus pažymime laukelį **Cases**. Spragtelėję mygtuką **Plots**, naujame lange **Hierarchical Cluster Analysis: Plots** pasirenkame **Dendogramas** braižymą (7.10 pav.).



7.9 pav. SPSS hierarchinės klasterinės analizės pagrindinis dialogo langas



7.10 pav. SPSS hierarchinės klasterinės analizės rezultatų grafinio vaizdavimo dialogo langas

Spragtelėję mygtuką **Method...**, naujame lange **Hierarchical Cluster Analysis: Method** sąrašė **Cluster Method** pasirenkame) klasterių sudarymo metodą - *vienetinės jungties (artimiausių kaimynų)* (7.11 pav.).

Artumo matų pasirinkimas priklauso nuo to, kokiai skalei priklauso analizuojami kintamieji: intervalų, išreikšti dažniais ar dvejetainiais. Pažymėję atitinkamą **Measure** grupės laukelį — **Interval**, **Counts** ar **Binary** pasirenkame vieną iš artumo matų –Euklido atstumo kvadratą (Squared Euclidian distance). **Transform Values** sąrašė **Standardize** nurodome, kad prieš klasifikaciją stebėtas reikšmes standartizuoti ir naudoti z-reikšmes (**Z scores**) (7.11 pav.)

Hierarchical Cluster Analysis: Method

Cluster Method:

Measure:

☒ Interval:

Power: Root:

☐ Counts:

☐ Binary:

Present: Absent:

Transform Values:

Standardize:

☒ By variable

☐ By case

Transform Measures:

☐ Absolute values

☐ Change sign

☐ Rescale to 0-1 range

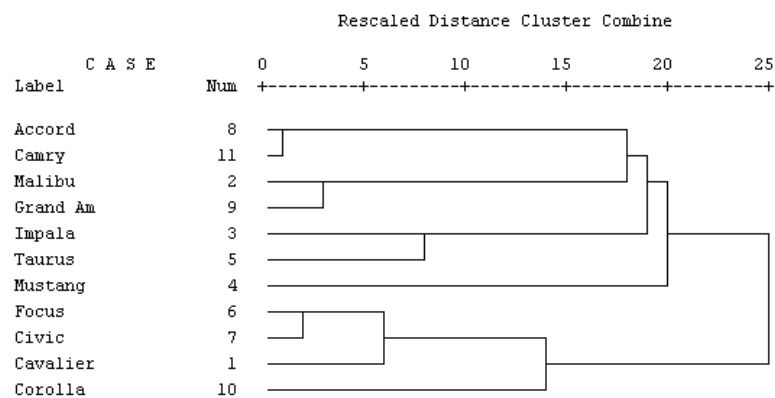
Continue Cancel Help

7.11 pav. SPSS hierarchinės klasterinės analizės klasterių sudarymo metodo, artumo mato ir duomenų standartizavimo metodo pasirinkimo langas

7.2 lentelė. Klasterizavimo eigos aprašas (vienetinės jungties metodas)

Stage	Cluster Combined		Coefficients	Stage Cluster First Appears		Next Stage
	Cluster 1	Cluster 2		Cluster 1	Cluster 2	
1	8	11	1.260	0	0	7
2	6	7	1.579	0	0	4
3	2	9	1.625	0	0	7
4	1	6	2.318	0	2	6
5	3	5	2.619	0	0	8
6	1	10	3.670	4	0	10
7	2	8	4.420	3	1	8
8	2	3	4.505	7	5	9
9	2	4	4.774	8	0	10
10	1	2	5.718	6	9	0

Dendrogram using Single Linkage



7.12 pav. Vienetinės jungties metodu gauta dendrograma

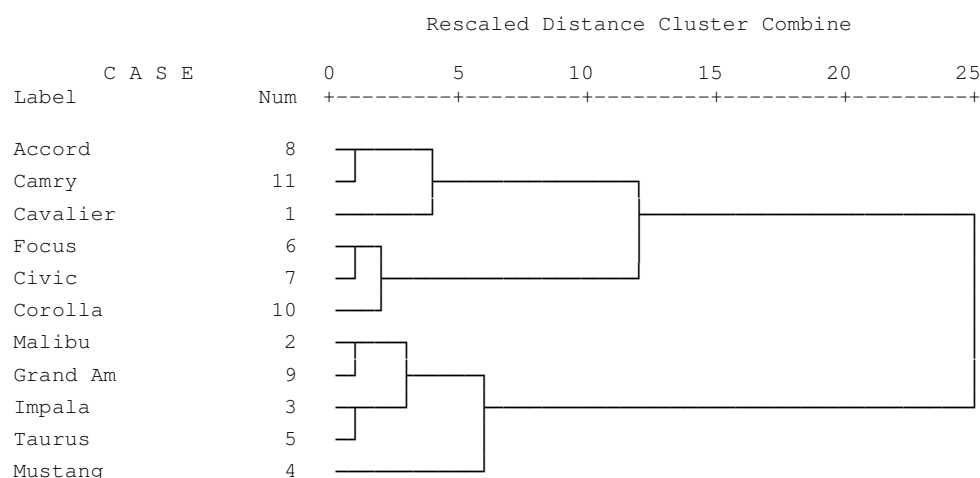
Klasterizavimo eigos aprašas pateiktas 7.2 lentelėje, o dendograma 7.12 pav. 7.2 lentelės stulpelyje **Stage** nurodomi klasterizavimo žingsnių numeriai. Mūsų pavyzdyje klasifikuojami 11 automobilių, todėl klasterizavimo žingsnių yra 10. Stulpelyje **Cluster Combined** nurodoma kokie klasteriai buvo jungiami atitinkamame žingsnyje, o stulpelyje **Coefficients** pateikiama atstumo tarp jungiamų klasterių reikšmė. Pavyzdžiui, 1-u žingsniu jungiami 8 ir 11 klasteriai (atstumas tarp jų lygus 1,26) ir jam suteikiamas numeris 8 (pirmojo jungiamo klasterio numeris). Prie gauto klasterio 7-u žingsniu prijungiamas 2 klasteris. Kas sudaro 2 klasterį? Tai galima išsiaiškinti analizuojant stulpelius **Stage Cluster First Appears**. Čia yra skaičiai 3 ir 1. Tai reiškia, kad 7 žingsniu jungiami klasteriai buvo gauti 3-u ir 1-u žingsniais. 3-u žingsniu buvo sujungti 2 ir 9 klasteriai, t.y. automobiliai Malibu ir Grand Am. Išvada, po 7-o žingsnio į vieną klasterį pateko 2, 8, 9 ir 11 klasteriai (automobiliai Accord, Camry, Malibu ir Grand Am). Klasterių hierarchija pavaizduota dendogramoje (7.12 pav.), bet klasterizavimo eigos aprašas pateikia detalesnę informaciją lyginant su dendograma.

Jeigu 7.2 lentelėje nagrinėsime atstumų intervalą nuo 2 iki 3,5 tai gausime 6 klasterius, jeigu nagrinėsime atstumų intervalą virš 5 gausime du klasterius. Pastaba: SPSS dendogramoje pateikia ne originalias atstumo reikšmes, bet padaugintas iš mastelio daugiklio (Rescaled distance). Matome, kad 6 klasterių sprendinyje yra didesni atstumai tarp klasterių. Dideli atstumai tarp klasterių rodo, kad klasteriai geriau atsiskiria. Šio sprendinio trūkumas, kad tuos klasterius sunku interpretuoti. Dviejų klasterių sprendinys yra nepatenkinamas, nes klasteriai gerai neatsiskiria, tarp jų yra mažas atstumas.

7.3 lentelė . Klasterizavimo eigos aprašas (pilnosios jungties metodas)

Agglomeration Schedule						
Stage	Cluster Combined		Coefficients	Stage Cluster First Appears		Next Stage
	Cluster 1	Cluster 2		Cluster 1	Cluster 2	
1	8	11	1,260	0	0	7
2	6	7	1,579	0	0	5
3	2	9	1,625	0	0	6
4	3	5	2,619	0	0	6
5	6	10	4,012	2	0	9
6	2	3	7,333	3	4	8
7	1	8	9,183	0	1	9
8	2	4	12,440	6	0	10
9	1	6	25,486	7	5	10
10	1	2	54,607	9	8	0

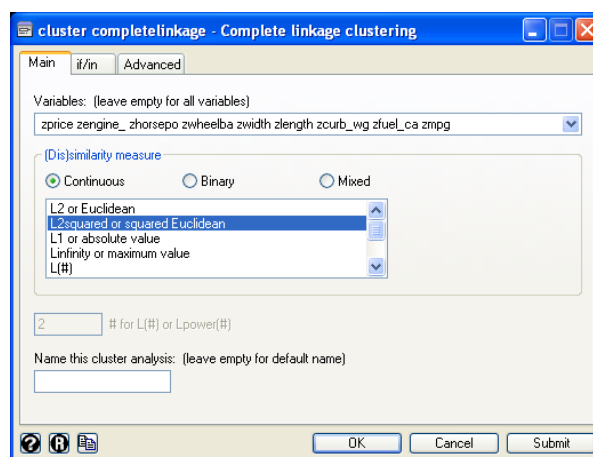
Dendrogram using Complete Linkage



7.13 pav. Pilnosios jungties metodu gauta dendrograma

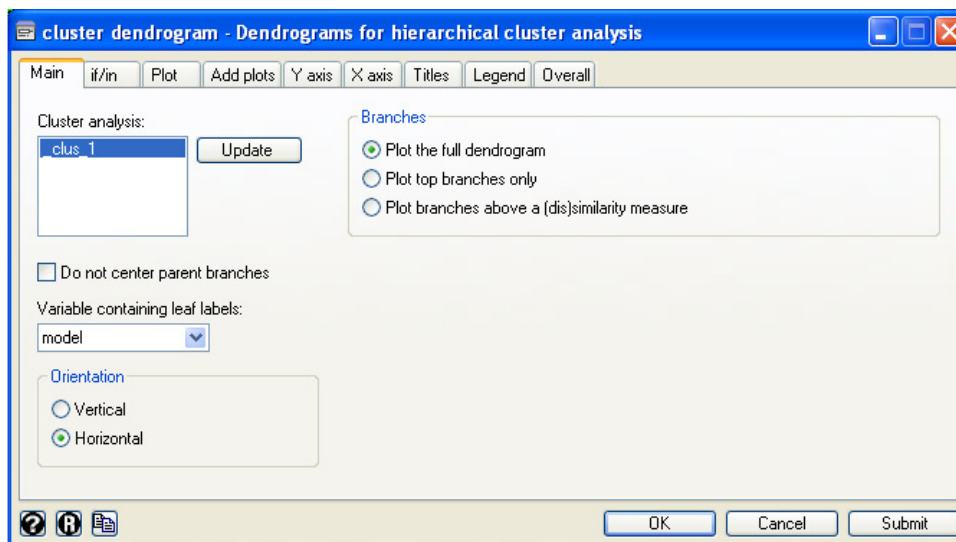
Išspręsimė tą patį uždavinį pilnosios jungties (tolimiausio kaimyno) metodu. Gauti rezultatai pateikti 7.3 lentelėje ir 7.13 pav. Pilnosios jungties metodu gerai išsiskiria du arba trys klasteriai. 2 klasterių sprendinyje išsiskiria mažesni (Accord, Camry, Cavalier, Focus, Civic, Carolla) ir didesni automobiliai. 3 klasterių sprendinyje mažesni automobiliai išskaidomi į dvi klases: maži automobiliai (Focus, Civic, Carolla) ir ekonominės klasės automobiliai (Accord, Camry, Cavalier). Be to Civic (Honda) ir Carolla (Toyota) yra mažesni ir pigesni už tų pačių gamintojų automobilius Accord (Honda) ir Camry (Toyota). Taigi, pilnosios jungties metodu geriau atsiskiria klasteriai (atstumai tarp jų yra didesni) ir geriau galima interpretuoti gautus rezultatus. Kurdami naują automobilį gamintojai gali įrašyti kainą ir techninius parametrus į duomenų failą ir patikrinti ar jų kuriamas automobilis patenka į perkamiausių automobilių klasterius.

Sprendžiant tą patį uždavinį pilnosios jungties metodu su STATA, pasirenkame meniu: **Statistics → Multivariate analysis → Cluster analysis → Cluster data → Complete linkage**. Atsidariusiame lange įkeliamė kintamuosius – automobilių kainos ir techninių parametrų z-reikšmės (zprice zengine_ zhorsepo zwheelba zwidth zlength zcurb_wg zfuel_ca zmpg) pagal kurias bus klasifikuojami automobiliai, nurodome artumo matą („L2squared or squared Euclidean“ - Euklido atstumo kvadratas) → OK



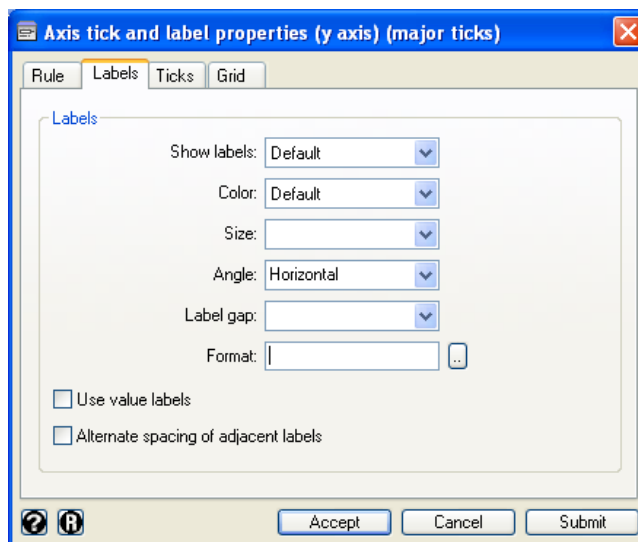
7.14 pav. STATA pilnosios jungties metodo dialogo langas

STATA sukuria tris naujus kintamuosius `_clus_1_id`, `_clus_1_ord`, `_clus_1_hgt`, kuriais remiantis sprendžiami klasterinės analizės uždaviniai. Šie kintamieji tai tarpiniai rezultatai ir interpretavimui nenaudojami. Remiantis šiais kintamaisiais, galime spręsti kitus klasterinės analizės uždavinius. Pavyzdžiui, nubraižyti dendrogramą. Braižant dendrogramą pasirenkame STATA meniu: **Statistics** → **Multivariate analysis** → **Cluster analysis** → **Postclustering** → **Dendrograms**.

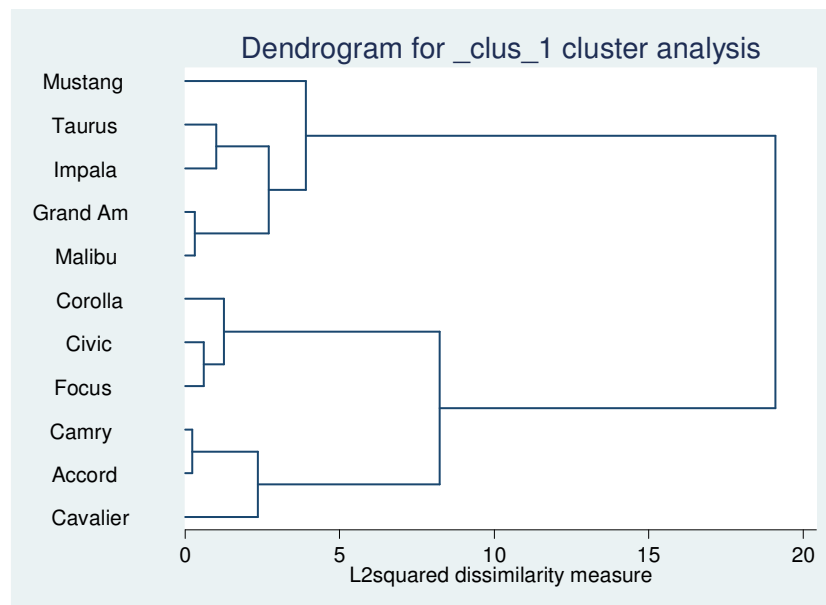


7.15 pav. STATA dendrogramos braižymo dialogo langas

Atsidariusiame lange pasirenkame kintamąjį **clus_1**, nurodome, kad dendrogramoje klasifikuojamų objektų –automobilių pavadinimai būtų paimti iš kintamojo **model** (pasirenkame šį kintamąjį iš „**variable containing leaf labels**“ meniu), taip pat pasirenkame horizontalią dendrogramos orientaciją ir horizontalų klasifikuojamų objektų pavadinimų pateikimą Y ašyje → **OK** arba **Submit**. Gauta dendrograma pateikta 7.17 pav.



7.16. pav. STATA Y ašies žymių parametrų nustatymo langas



7.17 pav. STATA pilnosios jungties metodu gauta dendrograma

8. FAKTORINĖ ANALIZĖ

8.1. Įvadas

Faktorinė analizė (*factor analysis*) –taikoma dideliame kintamųjų skaičiui sumažinti (*data reduction*), juos pakeičiant tiesiogiai nestebimais (latentiniais) faktoriais (*latent factors*). Šis metodas plačiai taikomas šiuolaikinėje psichologijoje, socialiniuose ir humanitariniuose moksluose. Faktorinės analizės pradininku yra Čarlzas Spirmenas (*Charles Spearman*), kuris prieš beveik 100 metų iškėlė hipotezę, kad dauguma žmogaus protinių gebėjimų - matematinių, kalbos, meno, loginio mąstymo ir t.t. - gali būti paaiškinti vienu bendruoju intelekto „faktoriumi“. Spirmenas tyrė mokinių įvairių mokymosi dalykų įvertinimus ir nustatė, kad visiškai skirtingų dalykų įvertinimai yra stipriai teigiamai koreliuoti. Tai leido jam teigti, kad egzistuoja vienas bendrasis intelekto faktorius, charakterizuojantis žmogaus intelektualius gebėjimus, kurį jis pavadino faktoriumi g. 1931 metais L. L. Terstounas (L. L. Thurstone) savo moksliniame darbe pirmasis pavartojo terminą „Faktorinė analizė“ [15], o 1960 metais buvo paskelbtas amerikiečių mokslininko G. Čarmano (G. Charman) darbas „Šiuolaikinė faktorinė analizė“. Pradžioje, faktorinė analizė daugiausiai buvo taikoma tik psichologijos duomenų analizei, tačiau vėliau ji pradėta plačiai taikyti sociologiniuose tyrimuose, medicinoje, karinėje pramonėje, ekonomikoje ir t.t. . Šiandien sunku rasti mokslinių tyrimų sritį, kurioje faktorinės analizės metodai netaikomi.

Kokie uždaviniai sprendžiami naudojant faktorinę analizę?

1 pavyzdys. Tarkime, mus domina, kokios respondentų nuostatos lėmė apklausos rezultatus. Atlikę apklausos rezultatų koreliacinę analizę pastebėjome, kad tarp atsakymų į klausimus egzistuoja tam tikra struktūra: atsakymus apie mokesčius sieja koreliacinis ryšys, atsakymus apie karines problemas sieja koreliacinis ryšys ir t.t. Faktorinės analizės pagalba, galime nustatyti, kokie ir kiek tiesiogiai nestebimų faktorių charakterizuoja stebimą reiškinį, šiuo atveju – respondentų nuostatas. Jeigu apklausos anketą sudaro 100 klausimų, o faktorinės analizės rezultatas – 12 faktorių, tuomet pradinę 100-o tarpusavyje koreliuotų kintamųjų aibę galime apibendrinti 12-os nekoreliuotų faktorių rinkiniu. Be to, kiekvienam respondentui galime apskaičiuoti išskirtų faktorių reikšmių įverčius ir juos panaudoti kituose statistikos metoduose. Pavyzdžiui, faktorių įverčiai gali būti panaudoti regresijos modeliui balsavimo rezultatų ateinančiuose rinkimuose prognozavimui. Faktorinės analizės taikymo sritis yra labai plati. Vietoj požiūrio į politiką galėtume tirti jų požiūrį į maisto produktus, automobilius ar gyvenamosios vietos pasirinkimą. Taigi , faktorinė analizė gali būti pritaikyta pradinių duomenų transformavimui tokiu būdu, kad jie atitiktų kitų statistikos metodų keliamus reikalavimus. Pavyzdžiui, taikant daugialypės regresinės analizės metodus, reikalaujama, kad nepriklausomi kintamieji tarpusavyje būtų nekoreliuoti. Kai ši sąlyga netenkinama, pradžioje galima taikyti faktorinės analizės metodą, kurio pagalba nepriklausomų kintamųjų skaičius sumažinamas ir kuriuos galima panaudoti regresinei analizei, neprarandant duomenyse esančios informacijos.

2 pavyzdys. Pavyzdžiui, automobilių prekybos agentas nori prognozuoti automobilių pardavimus. Jo turimi duomenys – automobilių modeliai, jų charakteristikos, pardavimo kainos ir pan. Agentas baiminasi, kad tarp gausybės kintamųjų yra nemažai tarpusavyje koreliuojančių, todėl prognozės rezultatai gali būti netikslūs. Pritaikę faktorinę analizę nustatytume nekoreliuotų faktorių rinkinį, charakterizuojantį agento turimus duomenis. Todėl prognozėms agentas galėtų naudoti faktorių reikšmių įverčius vietoj pradinių kintamųjų. Pabrėšime, kad faktorių skaičius būtų žymiai mažesnis už pradinių kintamųjų skaičių ir faktoriai būtų nekoreliuoti tarpusavyje.

3. pavyzdys. Analizuojant socialinių grupių ar atskirų asmenų vartojimo įpročius, nuomones ar elgesį pirkimo metu, dažnai formuluojamos pradinės hipotezės apie šių savybių tarpusavio ryšius. Faktoriinės analizės metodai leidžia įvertinti vienai ar kitai grupei būdingos savybės ir jų ryšius. Šie metodai gali būti taikomi tikrinant ir kito tipo hipotezes, pavyzdžiui, sparčią vartojimo plėtrą ir socialinį stabilumą, kartu įvertinant ir kitas savybes, kurios gali daryti įtaką tiriamiems procesams.

Faktoriinės analizės metodais sprendžiami dviejų tipų uždaviniai: tiriamieji ir patvirtinamieji. Pagal šiuos uždavinius ir faktoriinę analizę vadinama tiriamąja ir patvirtinamąja.

Tiriamoji faktoriinė analizė taikoma tada, kai turima kintamųjų aibė ir neaišku, nei kiek gali būti faktorių, nei kokie kintamieji juos sudaro. Pavyzdžiui, turint kelių metų duomenis apie regionus (valstybes, miestus) sprendžiamas uždavinys, kiek ir kokius nepriklausomus faktorius matuoja šie požymiai. Toliau nagrinėsime tik tiriančiąją faktoriinę analizę.

Patvirtinančioji faktoriinė analizė taikoma tada, kai norime jau turimas žinias patvirtinti ar išplėsti.

8.1. Faktoriinės analizės modelis

Pirmiausiai apibrėžime faktoriaus ir faktoriaus svorio sąvokas. **Faktorius** F_j – tai tiesiogiai nestebimas (latentinis) kintamasis, kuris yra susijęs su daugiau negu vienu kintamuoju X_i [4].

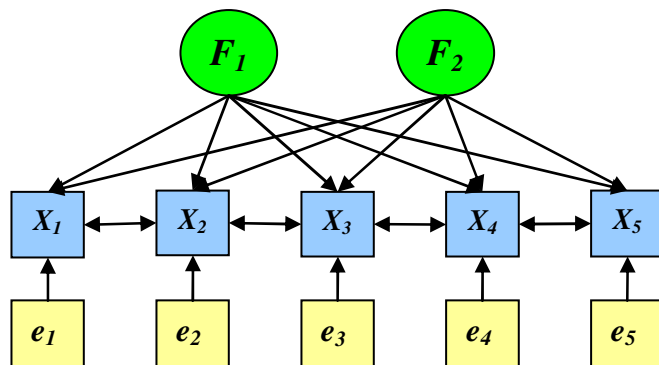
Faktoriaus svoris (*factor loading*) a_{ij} – tai koeficientas, nusakantis faktoriaus F_j ir stebimo kintamojo X_i koreliacinio ryšio stiprumą.

Matematiškai faktoriinės analizės idėją galime aprašyti taip: kiekvienas stebimas kintamasis X_i ($i = 1, \dots, n$) gali būti išreikštas tiesiogiai nestebimų (latentinių) faktorių F_j ($j = 1, \dots, k$) tiesine daugyda:

$$X_i = a_{i1}F_1 + a_{i2}F_2 + \dots + a_{ik}F_k + e_i,$$

čia $F_1, F_2, \dots, F_k$ yra bendrieji latentiniai faktoriai, $a_{i1}, a_{i2}, \dots, a_{ik}$ – faktorių svoriai (*weights*), o e_i – specifinis kintamojo X_i faktorius (*unique factor*). Tarkime, kad tiek specifiniai, tiek bendrieji faktoriai yra nekoreliuoti tarpusavyje, tuomet stebimų kintamųjų X_i tarpusavio ryšius galime charakterizuoti faktoriais $F_1, F_2, \dots, F_k$.

Faktoriinės analizės užduotis – nustatyti faktorių F_k skaičių, kuriais galėtume charakterizuoti visą kintamųjų X_i ($i = 1, \dots, n$) aibę, bei įvertinti kiek informacijos apie kintamuosius X_i prarandame, pakeisdami juos faktoriais F_j ($j = 1, \dots, k$).



8.1 pav. Faktoriinės analizės modelio pavyzdys

Tarkime, kad apklausos anketą sudaro 100 klausimų, o faktorinės analizės rezultatas – 12 faktorių, tuomet pradinę 100-o tarpusavyje koreliuotų kintamųjų aibę galime apibendrinti 12-os nekoreliuotų faktorių rinkiniu. Be to, kiekvienam respondentui galime apskaičiuoti nustatytų faktorių reikšmių įverčius ir juos panaudoti tolimesnėje analizėje vietoje pradinių kintamųjų.

Pateiksime faktorinės analizės pavyzdį. Tarkime, kad mokiniai atliko žinių ir gebėjimo testą. Testo rezultatams pritaikėme faktorinę analizę ir nustatėme, kokie faktoriai charakterizuoja stebėtus kintamuosius. Faktorių svoriai pateikti 8.1 lentelėje.

8.1 lentelė

Kintamieji	F_1	F_2	F_3
Loginis mąstymas, X_1	0,82	0,63	0,44
Gramatika, X_2	0,68	0,64	0,21
Literatūra, X_3	0,28	0,59	0,18
Algebra, X_4	0,45	0,20	0,38
Geometrija, X_5	0,50	0,17	0,69
Fizika, X_6	0,41	0,13	0,37
Lotynų, X_8	0,58	0,70	0,20
Prancūzų, X_9	0,32	0,68	0,17
Istorija, X_{10}	0,25	0,43	0,12
Technologijos, X_{11}	0,49	0,09	0,60

Visus 8.1 lentelės kintamuosius $X_1, X_2, \dots, X_{11}$ galime aprašyti faktoriais F_1, F_2 ir F_3 . Faktorių svoriai parodo, kaip stipriai kintamasis koreliuoja su faktoriumi. 8.2 lentelėje pateikta faktorių svorių interpretacija.

8.2 lentelė

Faktoriaus svoris, a_{ij}	Interpretacija
$ a_{ij} \geq 0,6$	Faktorių F_j ir kintamąjį X_i sieja stiprus ryšys
$0,3 \leq a_{ij} < 0,6$	Faktorių F_j ir kintamąjį X_i sieja ryšys
$ a_{ij} < 0,3$	Tarp faktoriaus F_j ir kintamojo X_i ryšys neegzistuoja

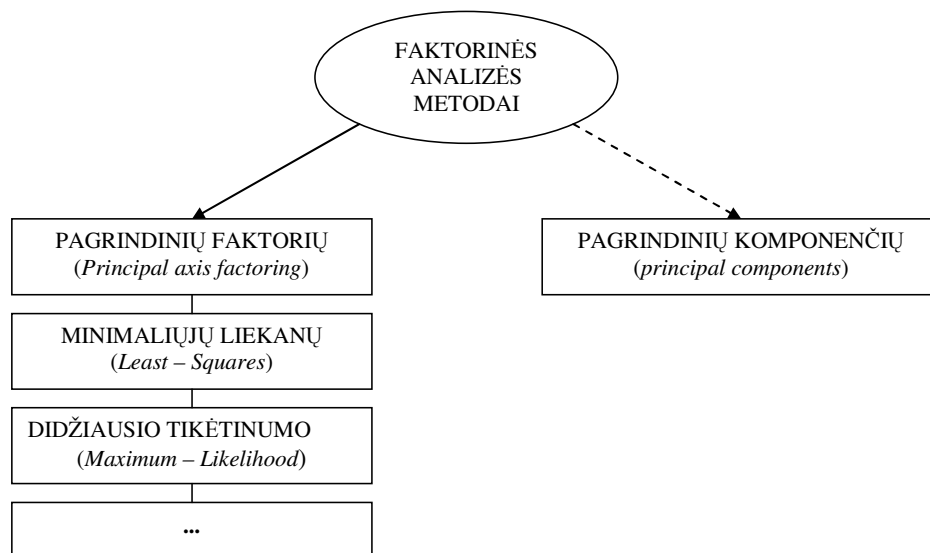
Pirmasis faktorius F_1 yra susijęs beveik su visais dalykais, bet stipriausiai su kintamuoju $X_1 =$ „Loginis mąstymas“. Išanalizavę visus faktoriaus F_1 ryšius su kintamaisiais X_i ($i = 1, \dots, 11$), galime teigti, kad faktorius F_1 nusako bendrąjį mokinio gebėjimą mąstyti, kuris yra svarbus visuose dalykuose. Taigi, faktorių F_1 galime pavadinti „Bendrieji protiniai gebėjimai“. Pastebėsime, kad faktorius F_1 skirtingai įtakoja nagrinėjamus kintamuosius X_i . Pavyzdžiui, Geometrijoje ($a_{51} = 0,50$) ir Lotynų kalboje ($a_{81} = 0,58$) mokinio protiniai gebėjimai turi didesnę svarbą negu Istorijoje ($a_{101} = 0,25$). Panašiai samprotaudami, galime nustatyti ir kitų faktorių prasmę. Antrasis faktorius F_2 yra stipriausiai susijęs su kalbomis, gramatika bei literatūra. Todėl šį faktorių galima pavadinti „Kalbiniai įgūdžiai“. Pastebėsime, kad faktoriaus F_2 svoriai kalbos kintamųjų atžvilgiu (a_{22}, a_{32}, a_{82} ir a_{92}) yra ženkliai didesni nei kitų. Trečiąjį faktorių F_3 galime pavadinti „Inžinieriniai gebėjimai“, nes jis koreliuoja su tikslųjų mokslų vertinimais (Algebra, Geometrija ir Fizika) bei su Technologijomis ir labai silpnai koreliuoja su humanitarinių dalykų vertinimais.

Išskyrę ir apibrėžę faktorius, galime atlikti tolimesnę duomenų analizę nenaudodami pradinių kintamųjų X_i ($i = 1, \dots, 11$), o naudodami faktorius F_j ($j = 1, 2, 3$). Pavyzdžiui, jeigu norėtume prognozuoti mokinių testavimo rezultatus panaudodami tiesinę regresiją, tai sudarant regresijos lygtį galėtume naudoti ne pradinių kintamųjų X_i reikšmes, o faktorių F_j reikšmių įverčius. Taigi, faktorinės analizės pagalba nustatėme, kad stebimą reiškinį aprašančių 11 kintamųjų X_i aibę galime apibendrinti 3 faktorių F_j rinkiniu.

Šiame kurse nagrinėsime tik tiriančiosios faktorinės analizės metodus, todėl toliau žodį „tiriančioji“ praleisime. Faktorinė analizė yra gana sudėtingas ir prieštaringas daugiamatės statistinės analizės uždavinys, nes jos rezultatai labai priklauso nuo paties tyrėjo, nes:

- šiam uždaviniui dažniausiai neegzistuoja vienintelis sprendinys, t.y. išskirti faktoriai yra nevienareikšmiai;
- išskirtus faktorius ne visada lengva interpretuoti;
- ne visada įmanoma išskirti faktorius, apibendrinančius stebimo reiškinio kintamuosius.

8.2. Faktorinės analizės metodai



8.2 pav. Faktorinės analizės metodai

Faktorinės analizės metodus galime suskirstyti į dvi grupes. Pirmiausia reikėtų paminėti, kad pagrindinių komponentių metodas iš tikrųjų nėra priskiriamas faktorinei analizei, nors ir turi daug su ja bendro. Pagrindinė to priežastis – paaiškinamos kintamųjų dispersijos dalies skirtumas (*total variance explained*).

Taikant faktorinės analizės metodus yra priimama, kad tik tam tikra stebėtų kintamųjų dispersijos dalis gali būti visiškai paaiškinta išskirtais faktoriais, o likusi dispersijos lieka nepaaiškinta dėl kintamųjų specifiškumo.

Priešingai negu taikant faktorinės analizės metodus, pagrindinių komponentių metodu išskirti faktoriai paaiškina visą stebimų kintamųjų dispersiją. Vadinasi, pagrindinių komponentių metodas neįvertina kiekvieno kintamojo atskirai dėl jo specifiškumo.

8.3. Faktorinės analizės etapai

8.3.1. Aprašomoji statistika ir duomenų tikamumas faktorinei analizei

Faktorinę analizę galime taikyti tik kiekybiniais kintamiesiems, išmatuotiems intervalų arba tvarkos skalėje, kai tarp kintamųjų yra koreliacinis ryšys. Nagrinėdami kovariacijų (koreliacijų) matricą galime nustatyti faktorinei analizei netinkamus (silpnai koreliuotus) kintamuosius ir juos pašalinti iš pradinių kintamųjų rinkinio. SPSS skaičiuoja *KMO* (*Kaiser-Meyer-Olkin statistics*) ir *Bartleto sferiškumo statistikas* (*Bartlett's Test of Sphericity*).

Bartleto sferiškumo kriterijus parodo, ar tarp stebėtų kintamųjų yra reikšmingai koreliuojančių. Tikrinama hipotezė:

H_0 : „Visi stebėti kintamieji yra nekoreliuoti“,

H_a : „Tarp stebėtų kintamųjų yra reikšmingai koreliuojančių“.

KMO kriterijus tikrina ar stebimus kintamuosius išvis įmanoma apibendrinti tam tikrų faktorių rinkiniu. Jeigu KMO yra mažesnis už 0,6, tuomet stebėtiems kintamiesiems faktorinė analizė yra netaikytina. Maža KMO reikšmė parodo, kad kintamųjų tarpusavio koreliacija negali būti paaiškinta kitais kintamaisiais.

Prieš taikant faktorinę analizę, rekomenduojama patikrinti kintamųjų tinkamumą faktorinei analizei apskaičiuojant kiekvieno kintamojo X_i ($i = 1, \dots, n$) MSA_i (*Measure of Sampling Adequacy*) statistikos reikšmę. Ši statistika parodo kiekvieno kintamojo X_i tinkamumą faktorinei analizei. Jeigu $MSA_i < 0,5$, tai kintamasis X_i faktorinei analizei yra netinkamas.

8.3.2. Faktorių išskyrimas

Šiame skyrelyje supažindinsime su pagrindinių komponentių ir pagrindinių faktorių metodais ir aptarsime jų panašumus bei skirtumus.

Pagrindinių komponentių metodas. *Pagrindinė komponentė* galime išreikšti stebimų kintamųjų tiesine daugdara:

$$C_i = a_{i1}X_1 + a_{i2}X_2 + \dots + a_{in}X_n,$$

kur

C_i - išskirta i -oji komponentė,

a_{ij} - i -osios komponentės svoris j -ojo kintamojo atžvilgiu (kintamojo j ir komponentės i koreliacijos koeficientas),

X_j - kintamojo j reikšmės,

Sprendžiant pagrindinių komponentių analizės uždavinį reikia rasti svorius a_{ij} , ir nustatyti, kiek pradinių kintamųjų dispersijos paaiškina kiekviena komponentė C_i .

Trumpai šio metodo idėją galima nusakyti taip: išskiriame pirmąją komponentę C_1 , kuri paaiškina didžiausią dalį pradinių (stebėtų) kintamųjų X_j ($j = 1, \dots, n$) dispersijos, ir apskaičiuojame jos svorių vektorius $b_1 = (b_{11}, b_{12}, \dots, b_{1n})^T$ atitinkančią tikrinę reikšmę λ_1 (*eigenvalue*), kuri parodo paaiškinamos dispersijos dalį ($DC_1 = \lambda_1$). Tuomet ieškome antrosios komponentės C_2 , kuri paaiškina didžiausią dalį likusios dispersijos ir yra nekoreliuota su pirmąja, $cov(C_1, C_2) = 0$ bei randame tikrinę reikšmę λ_2 . Procesą tęsiame tol, kol išskiriamos visos komponentės $C_3, \dots, C_k$ ir

visa kintamųjų X_j ($j = 1, \dots, n$) dispersija yra paaiškinama komponentėmis, t.y. $\lambda_1 + \lambda_2 + \dots + \lambda_k = DX_1 + DX_2 + \dots + DX_n$. Praktikoje dažniausiai kelios pirmosios komponentės paaiškina beveik visą pradinių kintamųjų dispersiją.

Pagrindinių faktorių metodas. Pagrindinių faktorių modelyje, kaip ir visuose faktorinės analizės modeliuose stebimus kintamuosius X_i išreiškiame faktorių F_j tiesinėmis daugdaromis:

$$X_i = a_{i1}F_1 + a_{i2}F_2 + \dots + a_{ik}F_k + e_i,$$

X_i - kintamojo reikšmės,

a_{ij} - faktoriaus j svoris (kintamojo i ir faktoriaus j koreliacijos koeficientas),

F_j - išskirtas j -asis faktorius,

e_i - specifinis faktorius, susijęs tik su kintamuoju X_i .

Pagrindinių komponentių analizės modelyje, komponentes C_j išreiškiame stebėtų kintamųjų $X_1, X_2, \dots, X_k$ tiesinėmis daugdaromis, o pagrindinių faktorių analizėje, kiekvieną kintamąjį X_i išreiškiame faktorių F_j tiesinėmis daugdaromis.

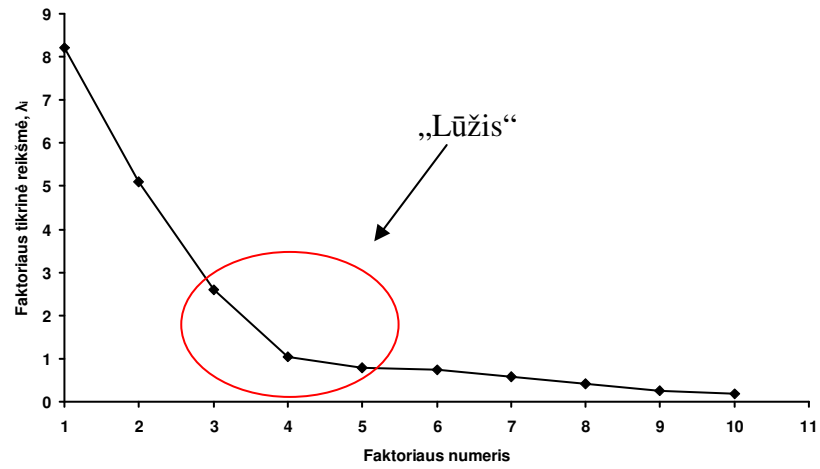
Pagrindinėmis komponentėmis galime paaiškinti visą kiekvieno kintamojo X_i dispersiją (DX_i), o pagrindinių faktorių metodu išskirti faktoriai paaiškina tik bendrąją kintamojo dispersiją, o specifinė dispersija lieka nepaaiškinta.

Dėl paaiškinamos dispersijos dalies skirtumo, pagrindinių komponentių analizę yra netikslinga taikyti, kai norime nustatyti pradinių kintamųjų $X_1, X_2, \dots, X_k$ priklausomybės struktūrą aprašančių faktorių prigimtį ir skaičių. Pagrindinių komponentių analizė neatskiria bendrosios kintamojo dispersijos nuo specifinės, todėl nustatyti faktoriai gali būti klaidingi. Tačiau abi šios procedūros yra lygiavertės, jei norime sumažinti stebėtų kintamųjų skaičių.

8.3.3. Faktorių skaičiaus nustatymas

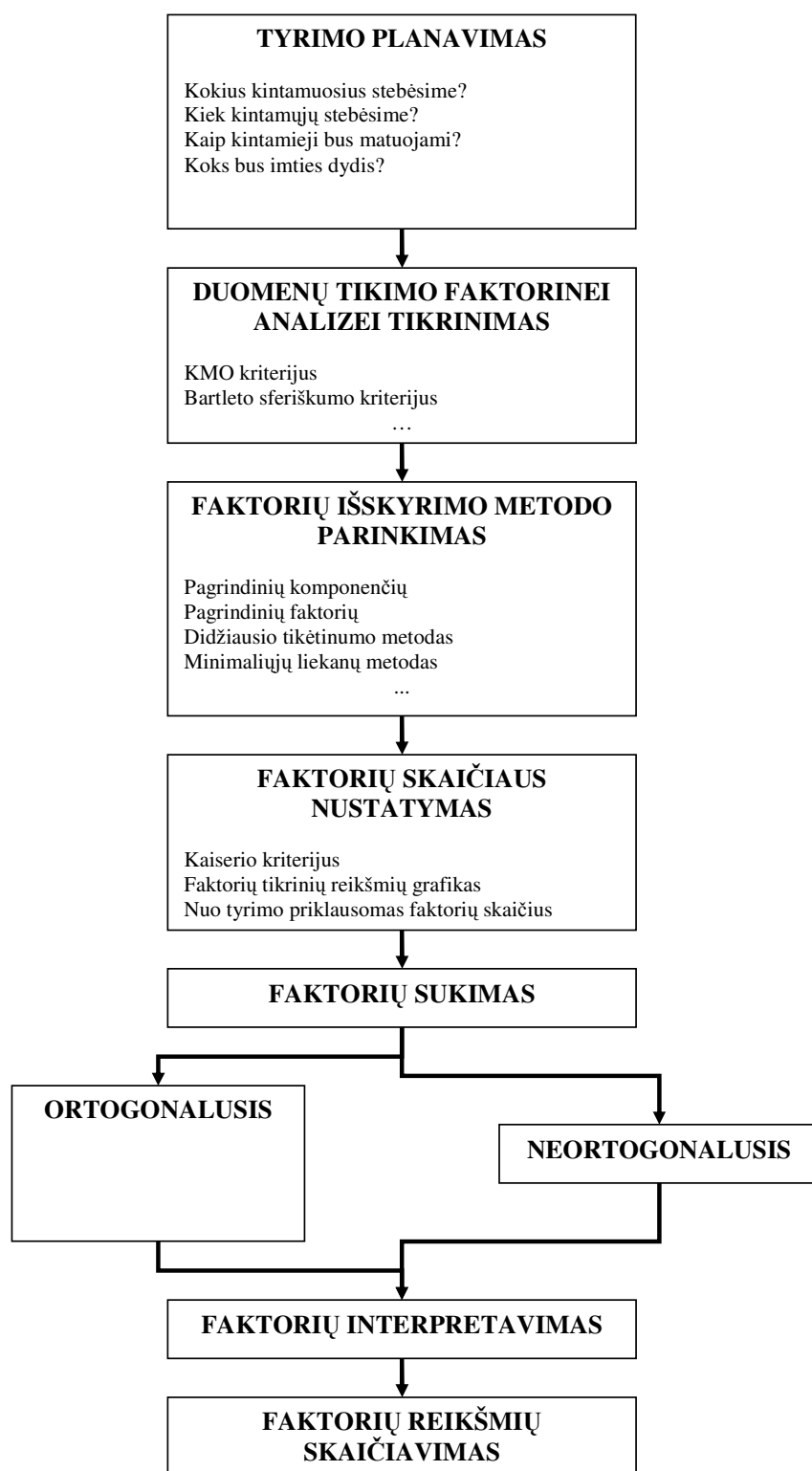
Kaiserio kriterijus (*Kaiser criterion*). Dažniausiai faktorių skaičius nustatomas naudojant *Kaiserio kriterijų* (Kaiser), kuris rekomenduoja į faktorių rinkinį įtraukti tik tuos faktorius, kurių tikrinės reikšmės yra didesnės už 1, t.y. $\lambda_j > 1$.

Faktorių tikrinių reikšmių λ_j grafikas (*Scree plot*). Tai grafinis faktorių skaičiaus nustatymas. Braižomas faktorių tikrinių reikšmių λ_j grafikas. Grafike paprastai yra matomas „lūžis“ tarp tikrinių reikšmių dydžių (8.3 pav.). Faktoriai su mažomis tikrinėmis reikšmėmis (esantys po „lūžio“) yra pašalinami iš faktorių rinkinio.



8.3 pav. Faktorių tikrinių reikšmių grafikas

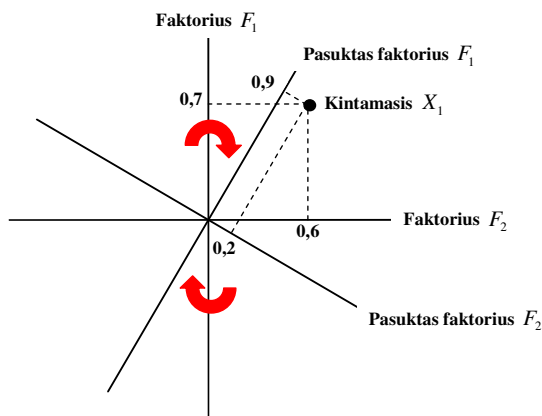
Faktorių skaičių galime parinkti ir patys, priklausomai nuo sprendžiamo uždavinio. Kartais prieš atlikdami faktorinę analizę jau žinome, kiek faktorių reikia išskirti arba, kokia pradinių kintamųjų dispersijos dalis (pvz. 80%) turi būti paaiškinta faktoriais.



8.4. pav. Faktorinės analizės etapai

8.3.4. Faktorių sukimas ir interpretavimas

Dažnai išskirtus faktorius (arba komponentes) yra sunku tinkamai interpretuoti, nes vienas kintamasis yra reikšmingai vienodai susijęs su keletu faktorių. Tarkime turime kintamąjį X_1 ir faktorius F_1 ir F_2 , kurių svoriai yra atitinkamai lygūs $a_{11} = 0,7$ ir $a_{12} = 0,6$. Šiuo atveju gana sudėtinga nuspręsti, kuris faktorius geriau apibūdina kintamąjį X_1 , nes abiejų faktorių svoriai yra didesni arba lygūs 0,6, t.y. kintamąjį ir faktorius sieja stiprus ryšys. Jeigu ir daugiau kintamųjų turi šį požymį, tuomet išskirtus faktorius prasmingai interpretuoti tampa labai sudėtinga. Norėdami išspręsti šią problemą turime atlikti *faktorių sukimą*. Faktorių sukimas – tai faktorių svorių matricos transformavimas į lengviau interpretuojamą pavidalą. Faktorių sukimą paaiškinsime grafiškai (8.5 pav). Faktorai F_1 ir F_2 vaizduojami koordinačių ašimis, o kintamasis X_1 – taškas plokštumoje F_1 ir F_2 . Akivaizdu, kad kintamasis X_1 yra panašiai nutolęs nuo abiejų faktorių ašių, todėl sunku nustatyti, kuris faktorius geriau charakterizuoja šį kintamąjį. Po faktorių F_1 ir F_2 sukimo, jų svoriai atitinkamai lygūs $a_{11} = 0,9$ ir $a_{12} = 0,2$. Taigi, dabar kintamąjį X_1 geriau apibūdina faktorius F_1 . Sukdami faktorius siekiame, kad kintamasis būtų arčiau vienos iš faktorių ašių, o idealiausia, kad jis būtų ant ašies.



8.5 pav. Ortogonalus faktorių sukimo pavyzdys

8.5 pav. pavaizduotas ortogonalus faktorių sukimas (*orthogonal rotation*). **Ortogonalus faktorių sukimas** – tai faktorių ašių sukimas išlaikant statų kampą (90 laipsnių) tarp ašių. Statas kampas tarp faktorių ašių yra būtinas, jeigu norime kad faktoriai būtų nekoreliuoti tarpusavyje. Faktorai gali būti sukami taip, kad kampas tarp ašių nebūtų status. Tai **neortogonalus faktorių sukimas** (*oblique rotation*). Toks faktorių sukimas yra retai naudojamas praktikoje, nes jis leidžia faktoriams koreliuoti tarpusavyje. Statistinės analizės paketuose įdiegta įvairių faktorių sukimo procedūrų. Populiariausi iš jų yra šie ortogonalieji sukimai:

- **Varimax** (ortogonalus VARIMAX sukimas). Minimizuojamas su kiekvienu faktoriumi reikšmingai susijusių kintamųjų skaičius.
- **Quartimax** (ortogonalus QUARTIMAX sukimas). Minimizuojamas kiekvieną kintamąjį aprašančių faktorių skaičius.
- **Equamax** (ortogonalus EQUAMAX sukimas). Kombinuotas VARIMAX ir QUARTIMAX metodas. Minimizuojamas kiekvieną kintamąjį aprašančių faktorių skaičius, bei su kiekvienu faktoriumi reikšmingai susijusių kintamųjų skaičius.

8.4. Faktorinės analizės pavyzdys

Tarkime, kad mus domina, kokie faktoriai įtakoja žmogaus santykius su dabartiniu partneriu. Turime pirminės informacijos, kad žmogaus santykiai su dabartiniu partneriu gali priklausyti nuo alternatyvos bei dabartinių santykių verčių.

Alternatyvos vertė parodo, ar žmogus linkęs nutraukti santykius su savo dabartiniu partneriu. Jeigu šio faktoriaus reikšmė yra aukšta, tai žmogus norėtų palikti savo dabartinį partnerį dėl kito(-os) arba tiesiog nutraukti santykius ir likti viena(-as).

Dabartinių santykių vertė parodo, kaip žmogus vertina santykius su savo dabartiniu partneriu. Pavyzdžiui, žmogus su savo dabartiniu partneriu turi daug bendrų draugų, laisvalaikį praleidžia kartu ir turi daug bendrų interesų. Toks žmogus nemažai prarastų, jei santykiai nutrūktų, todėl jo dabartinių santykių vertė yra aukšta.

Sudarykime trumpą anketą (8.3 lentelė) ir apklauskime žmones, kurie šiuo metu yra vedę/ištekę arba turi nuolatinį gyvenimo draugą(-ę).

8.3 lentelė. Anketa

Kiekvieną teiginį įvertinkite skalėje nuo 1 iki 7, kur 1 = „Visiškai nesutinku“, o 7 = „Visiškai sutinku“.		
	Teiginys	Įvertinimas
1.	Aš labai stengiuosi stiprinti santykius su savo dabartiniu partneriu	
2.	Jei išsiskirčiau su savo dabartiniu partneriu, prarasčiau daug mėgstamų laisvalaikio užsiėmimų	
3.	Jei išsiskirčiau su savo dabartiniu partneriu, prarasčiau daug bendrų draugų	
4.	Įdomiau būtų pradėti naują pažintį, negu tęsti santykius su dabartiniu partneriu	
5.	Labiau norėčiau būti vienas, negu su dabartiniu partneriu	
6.	Aš linkęs silpninti santykius su dabartiniu partneriu	

Tarkime, kad pagal šią anketą buvo apklausta 50 respondentų. Apklausos duomenys įrašyti faile *atsakymai.sav*. Teiginių įvertinimus žymėsime:

X_1 = „Aš labai stengiuosi stiprinti santykius su savo dabartiniu partneriu“,

X_2 = „Jei išsiskirčiau su savo dabartiniu partneriu, prarasčiau daug mėgstamų laisvalaikio užsiėmimų“,

X_3 = „Jei išsiskirčiau su savo dabartiniu partneriu, prarasčiau daug bendrų draugų“,

X_4 = „Įdomiau būtų pradėti naują pažintį, negu tęsti santykius su dabartiniu partneriu“,

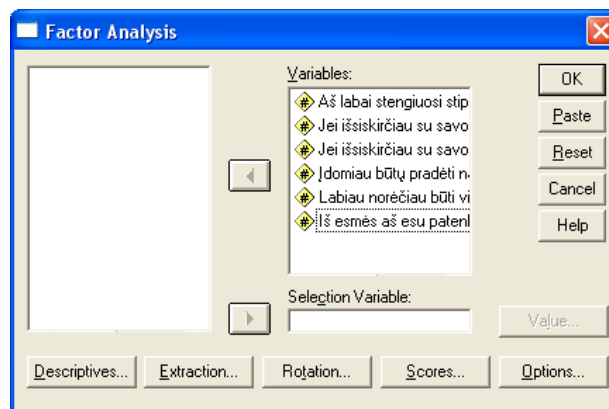
X_5 = „Labiau noriu būti vienas, negu su dabartiniu partneriu“,

X_6 = „Aš linkęs silpninti santykius su dabartiniu partneriu“.

	X_1	X_2	X_3	X_4	X_5	X_6
1	7	7	6	1	2	2
2	7	7	6	1	1	1
3	1	1	1	4	2	5
4	2	2	2	6	3	3
5	5	5	1	6	6	6
6	6	6	6	5	2	4
7	6	3	3	1	1	2
8	7	6	6	2	1	2
9	7	6	6	2	1	2
10	4	5	4	4	4	4
11	1	1	1	3	3	2
12	4	4	4	3	4	3
13	5	5	6	2	1	2
14	5	4	3	3	3	2
15	6	7	7	2	2	2

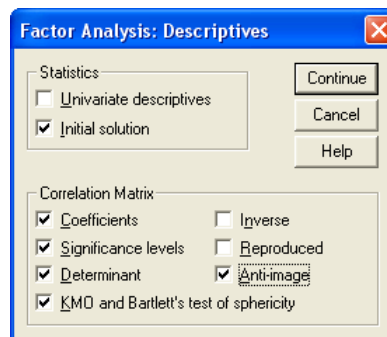
8.6 pav. Duomenų failo atsakymai.sav fragmentas

8.6 pav. pateiktas duomenų failo atsakymai.sav fragmentas. Atverkime pagrindinį SPSS faktorinės analizės langą (*Analyze ➔ Data reduction ➔ Factor...*).



8.7 pav. SPSS faktorinės analizės lango pavyzdys

Kintamuosius X_1 , X_2 , X_3 , X_4 , X_5 ir X_6 perkeliame į langelį **Variables** (faktorinės analizės kintamieji) (8.7 pav.).

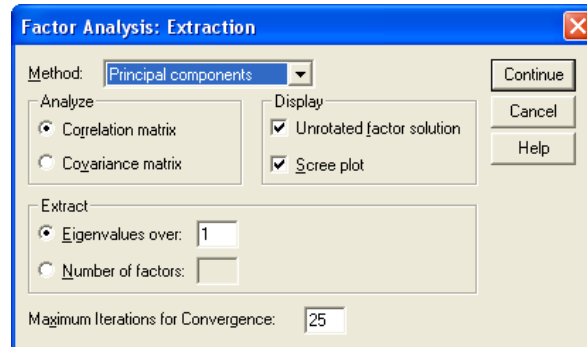


8.8 pav. Aprašomosios statistikos lango pavyzdys

Descriptives (aprašomoji statistika) meniu 8.8 pav.) pažymėkime:

- **Initial solution** (pradinis sprendinys),
- **Coefficients** (kintamųjų koreliacinė matrica),

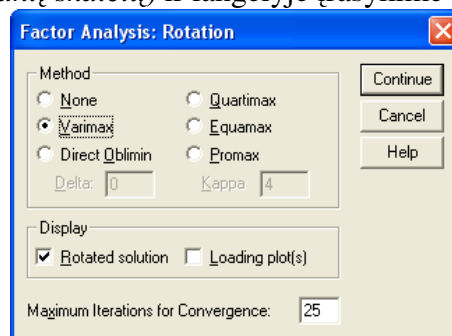
- **Significance levels** (hipotezių apie koreliacijos koeficientų lygybę nuliui tikrinimas),
- **Determinant** (koreliacinės matricos determinantas),
- **KMO and Bartlett's test of sphericity** (KMO ir Bartleto sferiškumo kriterijus).
- **Anti-image** (kintamųjų stebėjimų tinkamumo matas).



8.9 pav. Faktorių išskyrimo lango pavyzdys

Faktorių išskyrimui naudosime pagrindinių komponentų metodą, todėl **Extraction** (faktorių išskyrimas) langas (8.9 pav.) **Method** skiltyje pasirinkime **Principal Components** (pagrindinių komponentų metodas). Taip pat šiame meniu pažymėkime:

- **Correlation matrix** (faktoriai išskiriami naudojant kintamųjų koreliacinę matricą),
- **Unrotated factor solution** (nepasuktas sprendinys),
- **Scree plot** (faktorių tikrinių reikšmių grafikas),
- **Eigenvalues over** (į faktorių rinkinį įtraukiami tik tie faktoriai kurių tikrinės reikšmės didesnės už langelyje esantį skaičių) ir langelyje įrašykime **1**.



8.10 pav. Faktorių sukimo metodo parinkimo meniu

Rotation (faktorių sukimas) lango (8.10 pav.) pažymėkime:

- **Varimax** (ortogonalus Varimax faktorių sukimas),
- **Rotated solution** (rezultatų lange rodyti pasukto sprendinio svorių matricą).

Duomenų tinkamumas faktorinei analizei. Faktorinė analizė neturi prasmės nekoreliuotiems duomenims, todėl pirmiausia reikia analizuoti kintamųjų $X_1, X_2, \dots, X_6$ koreliacinę matricą. Koreliacinės matricos analizės tikslas - nustatyti visiškai nekoreliuotus arba silpnai tarpusavyje koreliuotus kintamuosius ir pašalinti juos iš pradinių kintamųjų rinkinio. Pastebėsime, kad kartais kintamieji gali ir per daug stipriai koreliuoti tarpusavyje (multikolinearumo problema).

Correlation Matrix ^a							
	X_1	X_2	X_3	X_4	X_5	X_6	
Correlation	X_1	1,000	,857	,747	-,373	-,420	-,296
	X_2	,857	1,000	,816	-,219	-,238	-,142
	X_3	,747	,816	1,000	-,255	-,348	-,219
	X_4	-,373	-,219	-,255	1,000	,621	,592
	X_5	-,420	-,238	-,348	,621	1,000	,664
	X_6	-,296	-,142	-,219	,592	,664	1,000
Sig. (1-tailed)	X_1		,000	,000	,004	,001	,019
	X_2	,000		,000	,063	,048	,162
	X_3	,000	,000		,037	,007	,063
	X_4	,004	,063	,037		,000	,000
	X_5	,001	,048	,007	,000		,000
	X_6	,019	,162	,063	,000	,000	

a. Determinant = ,019

8.11 pav. Koreliacinė matrica

Kintamieji X_1 , X_2 ir X_3 reikšmingai stipriai koreliuoja tarpusavyje, hipotezės apie koreliacijos koeficientų lygybę nuliui, atmetos ($p = 0,000 < 0,05$). Tačiau visi jie silpnai koreliuoja su kintamaisiais X_3 , X_4 ir X_5 . Kintamieji X_3 , X_4 ir X_5 reikšmingai koreliuoja tarpusavyje, tačiau silpnai koreliuoja su kintamaisiais X_1 , X_2 ir X_3 . Taigi, kintamieji X_1 , X_2 ir X_3 sudaro vieną grupę, o kintamieji X_3 , X_4 ir X_5 - kitą. Koreliacinės matricos determinantas yra lygus 0,019 ($> 0,00001$), todėl stebėti kintamieji neturi multikolinearumo problemos.

KMO and Bartlett's Test		
Kaiser-Meyer-Olkin Measure of Sampling Adequacy.		,747
Bartlett's Test of Sphericity	Approx. Chi-Square	183,601
	df	15
	Sig.	,000

8.12 pav. Duomenų tinkamumo faktorinei analizei tyrimo rezultatai

Bartleto sferiškumo kriterijus (žr. **KMO and Bartlett's Test**) parodo, ar tarp analizuojamų kintamųjų yra statistiškai reikšmingai koreliuojančių. Iš gautų rezultatų matome, kad ši hipotezė yra atmetama ($p = 0,000 < 0,05$).

KMO kriterijus parodo, ar analizuojamiems duomenims iš viso galima takyti faktorinę analizę. Jei šio kriterijaus reikšmė yra mažesnė už 0,6, tuomet kintamųjų porų koreliacijos negali būti paaiškinta kitais kintamaisiais ir faktorinė analizė yra netaikytina. Mūsų atveju ši reikšmė yra 0,747 ($> 0,6$), todėl turimiems duomenims faktorinę analizę galime taikyti.

Neigiamų dalinių koreliacijų ir kovariacijų matricos (**Anti – image Matrices**) parodo, kurių kintamųjų koreliacijų negalima paaiškinti išskirtais faktoriais. Jeigu daugumos matricos elementų, neesančių pagrindinėje įstrižainėje, reikšmės yra mažos (artimos 0) tuomet kintamojo koreliacinius ryšius galime gerai paaiškinti faktoriais. Iš gautų rezultatų matome, kad visų kintamųjų koreliaciniai ryšiai sąlyginai gerai paaiškinami faktoriais.

Pagrindinėje neigiamų dalinių koreliacijų (arba kovariacijų) matricos įstrižainėje yra **Kintamojo stebėjimų tinkamumo matas – MSA (Measure of sampling Adequacy)**. Rekomenduojama iš pradinių kintamųjų rinkinio pašalinti tuos kintamuosius, kurių $MSA < 0,5$. Mūsų atveju visų kintamųjų MSA reikšmės yra didesnės už 0,5 - visi kintamieji $X_1, X_2, \dots, X_6$ stebėjimai yra tinkami faktorinei analizei.

Faktorių išskyrimui naudojome pagrindinių komponentių metodą (**Extraction Method: Principal Components**). Faktoriaus F_1 tikrinė reikšmė (**Initial Eigenvalues**) yra lygi 3,297 ir tai sudaro net

55% bendros visų kintamųjų dispersijos. Pastebėsime, kad faktoriais F_1 ir F_2 galime paaiškinti 82% visos kintamųjų X_1 , X_2 , X_3 , X_4 , X_5 ir X_6 dispersijos. Likę faktoriai $F_3, \dots, F_6$ paaiškina tik nedidelę dispersijos dalį (18%), todėl juos atmetame.

Anti-image Matrices							
		X_1	X_2	X_3	X_4	X_5	X_6
Anti-image Covariance	X_1	,210	-,131	-,013	,045	,066	,013
	X_2	-,131	,172	-,125	-,007	-,058	-,017
	X_3	-,013	-,125	,307	-,020	,071	,004
	X_4	,045	-,007	-,020	,540	-,152	-,157
	X_5	,066	-,058	,071	-,152	,426	-,203
	X_6	,013	-,017	,004	-,157	-,203	,504
Anti-image Correlation	X_1	,756 ^a	-,690	-,053	,133	,221	,040
	X_2	-,690	,649 ^a	-,546	-,022	-,215	-,057
	X_3	-,053	-,546	,811 ^a	-,049	,195	,010
	X_4	,133	-,022	-,049	,824 ^a	-,317	-,301
	X_5	,221	-,215	,195	-,317	,735 ^a	-,438
	X_6	,040	-,057	,010	-,301	-,438	,767 ^a

a. Measures of Sampling Adequacy(MSA)

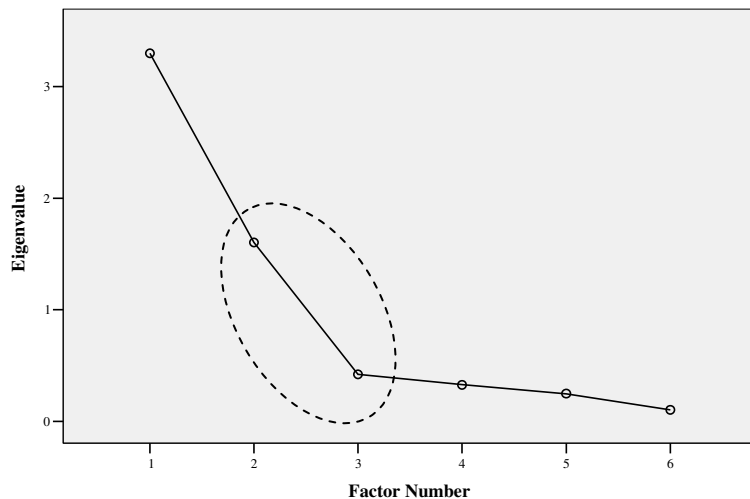
8.13 pav. Stebėjimų tinkamumo matai (MSA –Measure of sampling Adequacy)

Total Variance Explained									
Component	Initial Eigenvalues			Extraction Sums of Squared Loadings			Rotation Sums of Squared Loadings		
	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %
1	3,297	54,951	54,951	3,297	54,951	54,951	2,611	43,524	43,524
2	1,603	26,718	81,669	1,603	26,718	81,669	2,289	38,145	81,669
3	,420	7,005	88,674						
4	,328	5,465	94,139						
5	,248	4,131	98,270						
6	,104	1,730	100,000						

Extraction Method: Principal Component Analysis.

8.14 pav. Faktorių išskyrimo rezultatai

Scree Plot



8.15 pav. Faktorių tikrinių reikšmių grafikas (Scree plot)

Kaiserio kriterijus rekomenduoja išskirti tik tuos faktorius, kurių tikrinės reikšmės yra didesnės už 1. Gauti rezultatai parodo, kad tikslinga išskirti tik faktorius F_1 ir F_2 (tikrinės reikšmės atitinkamai lygios 3,297 ir 1,603).

Faktorių tikrinių reikšmių grafike (8.15 pav.) aiškiai matomas „lūžis“ tarp tikrinių reikšmių dydžių, kuris irgi rodo, kad reikėtų nagrinėti tik faktorius F_1 ir F_2 , o likusius faktorius $F_3, \dots, F_6$ atmesti.

Pastebėsime, kad atlikus faktorių sukimą, jų struktūra buvo optimizuota (8.14 pav. žr. **Rotation Sums of Squared Loadings**), t.y. dabar faktoriai F_1 ir F_2 yra beveik vienodos svarbos. Prieš sukimą faktorius F_1 paaiškino 55% bendros visų kintamųjų dispersijos, o po sukimo šis rodiklis sumažėjo iki 44%. Tačiau faktoriui F_2 šis rodiklis padidėjo nuo 27% iki 38%. Bendra faktoriais F_1 ir F_2 paaiškinama dispersijos dalis yra nepakitusi (82%).

Communalities			Component Matrix ^a		
	Initial	Extraction		Component	
				1	2
X_1	1,000	,874	X_1	,865	,356
X_2	1,000	,927	X_2	,780	,564
X_3	1,000	,837	X_3	,797	,449
X_4	1,000	,716	X_4	-,653	,538
X_5	1,000	,777	X_5	-,713	,519
X_6	1,000	,769	X_6	-,610	,630

Extraction Method: Principal Component Analysis.

Extraction Method: Principal Component Analysis.
a. 2 components extracted.

8.16 pav. Faktorių išskyrimo rezultatai

Faktorių išskyrimui naudojome pagrindinių komponentių metodą, kuris remiasi prielaida, kad visa stebėtų kintamųjų dispersija gali būti paaiškinta išskirtais faktoriais $F_1, F_2, \dots, F_6$. Dėl šios prielaidos pradiniai kintamųjų bendrumai (. **Initial Communalities**) yra lygūs 1.

Kintamųjų bendrumai po faktorių išskyrimo (**Extraction Communalities**) rodo, kokią kiekvieno kintamojo dispersijos dalį galime paaiškinti faktoriais F_1 ir F_2 . Pavyzdžiui, 87,4% kintamojo X_1 dispersijos galime paaiškinti faktoriais F_1 ir F_2 . Analogiškai interpretuojami ir likusių kintamųjų $X_2, \dots, X_6$ bendrumai (8.16 pav.).

Faktorių svorių matrica (žr. **Component Matrix**) parodo kintamųjų ir faktorių ryšius. Priminsime, kad faktorius ir kintamasis laikomi stipriai susiję, jeigu faktoriaus svoris $|a_{ij}| \geq 0.6$. Faktorių svorių matricoje matome, kad pavyzdžiui, faktorius F_1 yra reikšmingai susijęs su visais kintamaisiais X_1, X_2, X_3, X_4, X_5 ir X_6 . Tačiau minėti kintamieji aprašo skirtingus dalykus, todėl faktoriaus F_1 neįmanoma tinkamai interpretuoti. Norėdami palengvinti išskirtų faktorių interpretavimą, turime juos pasukti.

Faktorių sukimas

Rotated Component Matrix ^a		
	Component	
	1	2
X_1	,894	-,276
X_2	,961	-,061
X_3	,901	-,160
X_4	-,161	,831
X_5	-,220	,854
X_6	-,070	,874

Extraction Method: Principal Component Analysis.
Rotation Method: Varimax with Kaiser Normalization.
a. Rotation converged in 3 iterations.

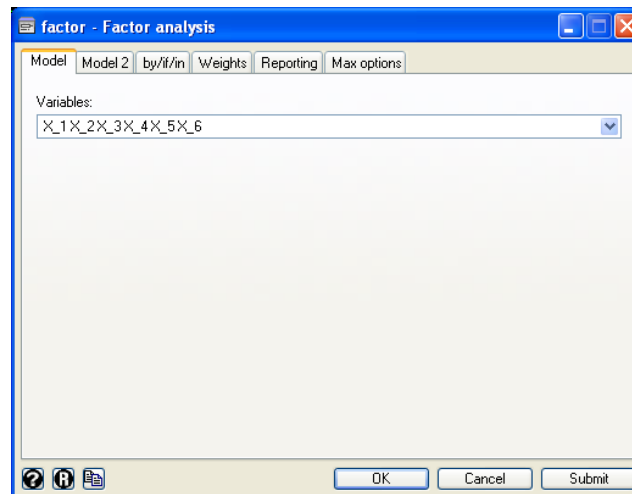
8.17 pav. Faktorių išskyrimo rezultatai po sukimo

Faktoriai buvo pasukti naudojant ortogonalųjį *Varimax* sukimą (žr. ***Rotated Component Matrix***). Po sukimo faktorius F_1 yra stipriai susijęs su kintamaisiais X_1 , X_2 ir X_3 (svoriai didesni už 0,6) ir visiškai nesusijęs su kintamaisiais X_4 , X_5 ir X_6 (svoriai moduliu mažesni už 0,3). Faktorius F_2 yra stipriai susijęs su kintamaisiais X_4 , X_5 ir X_6 (svoriai moduliu didesni už 0,6) ir visiškai nesusijęs su kintamaisiais X_1 , X_2 ir X_3 (svoriai moduliu mažesni už 0,3). Taigi, dabar faktorius interpretuoti yra žymiai lengviau.

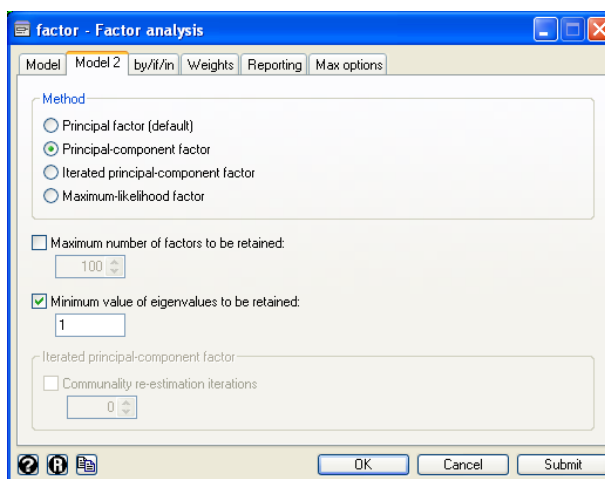
Faktorius F_1 yra stipriai susijęs su kintamaisiais X_1 , X_2 ir X_3 , kuriais įvertinama dabartinių santykių vertė, todėl faktorių F_1 galime pavadinti „Dabartinių santykių vertė“. Panašiai samprotaudami apibūdinsime ir kitą faktorių: kintamaisiais X_4 , X_5 ir X_6 įvertinama alternatyvos vertė, todėl F_2 galime įvardinti „Alternatyvos vertė“.

Respondentų atsakymai į anketos klausimus buvo analizuojami panaudojant faktorinę analizę. Faktorių išskyrimui panaudotas pagrindinių komponentų metodas. Išskirti faktoriai pasukti panaudojant ortogonalųjį *Varimax* sukimą. Tik pirmųjų dviejų faktorių F_1 ir F_2 tikrinės reikšmės yra didesnės už 1, todėl likusieji faktoriai buvo atmesti kaip nereikšmingi. Tikrinių reikšmių grafikas taip pat parodė, kad tik faktoriai F_1 ir F_2 yra reikšmingi. Faktoriais F_1 ir F_2 galime paaiškinti 82% visos kintamųjų $X_1, X_2, \dots, X_6$ dispersijos. Faktorius F_1 yra reikšmingai susijęs su kintamaisiais X_1 , X_2 ir X_3 - pavadintas „Dabartinių santykių vertė“. Kintamieji X_4 , X_5 ir X_6 yra reikšmingai susiję su faktoriumi F_2 , kuris pavadintas „Alternatyvos vertė“.

Sprendžiant tą patį uždavinį su STATA pasirenkame meniu: **Statistics → Multivariate analysis → Factor and principal component analysis → Factor analysis**. Atsidariusiame lenge (8.18 pav.) įkeliamė kintamuosius $X_1, X_2, X_3, X_4, X_5, X_6$.



8.18 pav. STATA faktorinės analizės langas



8.19 pav. STATA faktorių išskyrimo metodų ir parametrų nustatymo langas

Factor analysis/correlation
 Method: principal-component factors
 Rotation: (unrotated)

Number of obs = 50
 Retained factors = 2
 Number of params = 11

Factor	Eigenvalue	Difference	Proportion	Cumulative
Factor1	3.29706	1.69397	0.5495	0.5495
Factor2	1.60309	1.18278	0.2672	0.8167
Factor3	0.42031	0.09243	0.0701	0.8867
Factor4	0.32788	0.08004	0.0546	0.9414
Factor5	0.24784	0.14401	0.0413	0.9827
Factor6	0.10383	.	0.0173	1.0000

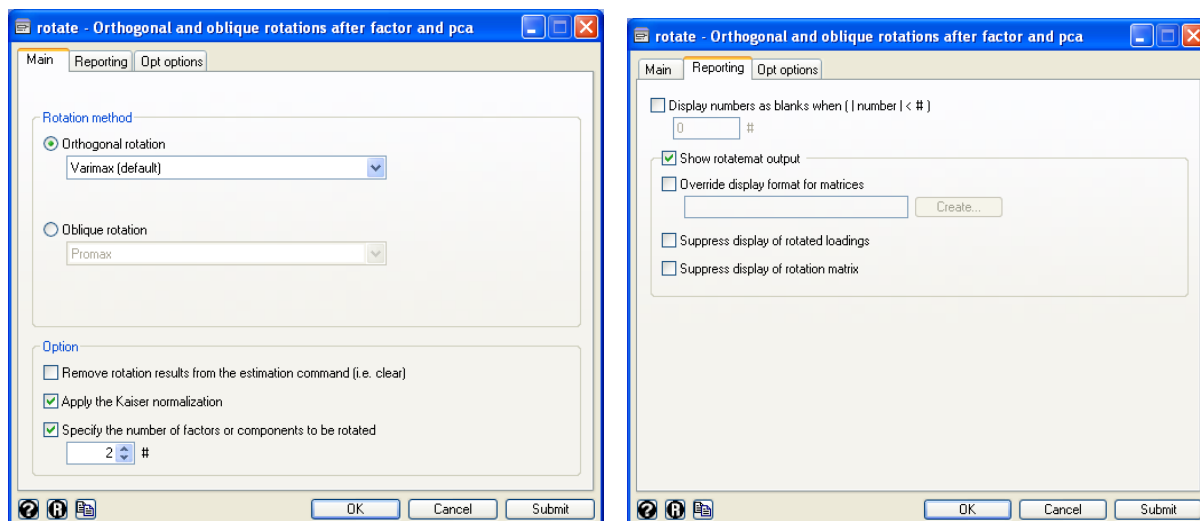
LR test: independent vs. saturated: $\chi^2(15) = 187.58$ Prob> $\chi^2 = 0.0000$

Factor loadings (pattern matrix) and unique variances

variable	Factor1	Factor2	Uniqueness
X_1	0.8648	0.3557	0.1256
X_2	0.7797	0.5645	0.0733
X_3	0.7969	0.4491	0.1632
X_4	-0.6530	0.5385	0.2836
X_5	-0.7126	0.5191	0.2227
X_6	-0.6098	0.6299	0.2313

8.20 pav. STATA faktorių išskyrimo rezultatai

Atliekame faktorių sukimą: **Statistics → Multivariate analysis → Factor and principal component analysis → Postestimation → Rotate loadings**. Pasirenkame **Varimax** faktorių sukimo metodą su Kaizerio normalizacija ir nurodome, kad būtų sukami 2 faktoriai. Gauti rezultatai tokie patys kaip ir skaičiuojant su SPSS pateikti lentelėse (8.22 pav.).



8.21 pav. STATA faktorių sukimo metodo parinkimo meniu

Factor analysis/correlation
 Method: principal-component factors
 Rotation: orthogonal varimax (Kaiser on)

Number of obs = 50
 Retained factors = 2
 Number of params = 11

Factor	Variance	Difference	Proportion	Cumulative
Factor1	2.61145	0.32276	0.4352	0.4352
Factor2	2.28870	.	0.3814	0.8167

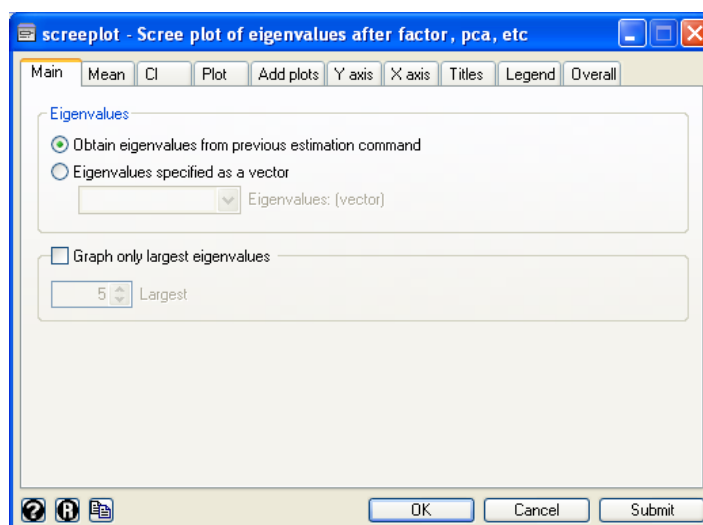
LR test: independent vs. saturated: $\chi^2(15) = 187.58$ Prob> $\chi^2 = 0.0000$

Rotated factor loadings (pattern matrix) and unique variances

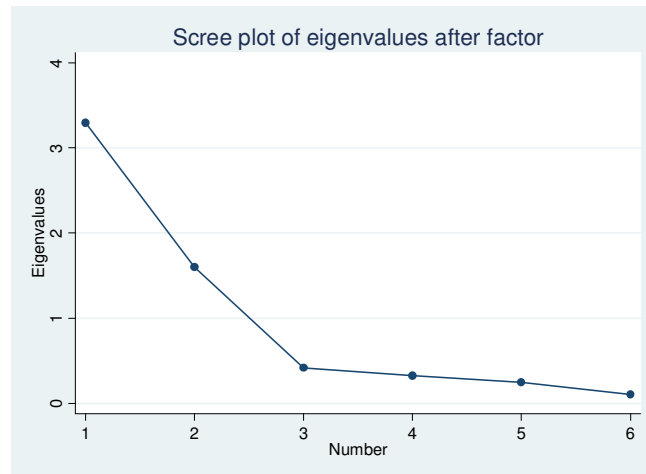
Variable	Factor1	Factor2	Uniqueness
X_1	0.8935	-0.2758	0.1256
X_2	0.9607	-0.0605	0.0733
X_3	0.9006	-0.1605	0.1632
X_4	-0.1612	0.8309	0.2836
X_5	-0.2196	0.8538	0.2227
X_6	-0.0697	0.8740	0.2313

8.22 pav. STATA faktorių išskyrimo rezultatai po sukimo

Nubraižome faktorių tikrinių reikšmių grafiką: **Statistics → Multivariate analysis → Factor and principal component analysis → Postestimation → Scree plot of eigenvalues**



8.22 pav. STATA tikrinių reikšmių grafikas (Scree plot) braižymo meniu



8.23 pav. STATA faktorių tikrinių reikšmių grafikas (Scree plot)

Gautų rezultatų interpretacija ta pati, kuri buvo pateikta aptariant SPSS faktorinės analizės pavyzdį, todėl čia jos nekartosime.

9. DAUGIAMAČIŲ SKALIŲ METODAS (MDS)

9.1. Įvadas

Metodai, kurie leidžia pavaizduoti daugiamačius duomenis mažesnės dimensijos erdvėje vadinami projekcijos metodais. Šie metodai gali būti taikomi ir kaip dimensijos mažinimo metodai. Jų tikslas – pateikti daugiamačius duomenis mažesnės dimensijos erdvėje taip, kad būtų kiek galima tiksliau išsaugota tam tikra duomenų struktūra. Projekcijos metodai gali būti naudojami ir daugiamačiams duomenims vizualizuoti, kai pasirinkta pakankamai maža projektavimo dimensija. Yra išskiriami tiesiniai ir netiesiniai projekcijos metodai. Pagrindinis skirtumas tarp jų yra tas, kad tiesiniai projekcijos metodai ieško tiesinio poerdvio (tokio, kaip tiesė ar plokštuma), o netiesiniai metodai ieško netiesinio poerdvio.

Daugiamačių skalių metodas (MDS) – tai grupė metodų, plačiai naudojamų daugiamačių duomenų analizėje įvairiose šakose, kurių pagalba n -mačiai vektoriai projektuojami į mažesnės dimensijos erdvę siekiant išlaikyti atstumus, panašumus arba kitus artumo matavimus tarp analizuojamos aibės objektų. Dažniausiai siekiama transformuoti artumo duomenis taip, kad jų dimensija sumažėtų iki 2-3 (maksimaliai išlaikant pradinį artumą tarp objektų arba savybių proporcijas bei tvarką). Jeigu dimensija sumažinama iki 2-3, tai duomenis galima vizualizuoti, t.y. pamatyti ir analizuoti duomenų išsidėstymą bei struktūrą. MDS ieško panašumų ar skirtingumų tarp analizuojamų duomenų objektų. Analizės rezultate gautuose grafikuose tie objektai, kurie yra panašūs, yra vaizduojami arčiau vieni kitų, o mažiau panašūs – toliau vieni nuo kitų. Ryšiais tarp aibės elementų gali būti Euklido atstumai. Tačiau, bendru atveju, tai nebūtinai turi būti atstumai. Vienas MDS pavyzdys galėtų būti toks. Tarkime turime matricą, sudarytą iš atstumų tarp pagrindinių šalies miestų. MDS analizės rezultate gautume miestų išdėstymą žemėlapyje, t.y. dvimatėje plokštumoje. Kitas MDS matricos pavyzdys yra koreliacijų tarp duomenų parametrų matrica. Jei tie duomenys traktuojami kaip panašumai, MDS algoritmu stipriai koreliuoti parametrai atvaizduojami arti vieni kitų, silpnai koreliuoti – toliau vieni nuo kitų. Vienas MDS tikslų yra rasti optimalią daugiamačių duomenų konfigūraciją mažo matavimo erdvėje. Yra daugybė skirtingų MDS variantų su skirtingomis paklaidų funkcijomis (STRESS funkcijomis) ir jas optimizuojančiais algoritmais. Pagal analizuojamus duomenis MDS algoritmai gali būti skirstomi į metrinis ir nemetrinius [35,36]. Pirmasis MDS algoritmas metriniais duomenimis buvo pasiūlytas 1930 metais, vėliau MDS algoritmai buvo taikyti ir nemetriniams duomenims. Metriniai MDS algoritmai, dar vadinami klasikiniiais, naudojami tada, kai įmanoma rasti Euklido atstumus tarp analizuojamų duomenų elementų, t.y. analizuojami metriniai duomenys. Pagrindinis šių algoritmų tikslas – pavaizduoti daugiamačius taškus 1-3 matavimų erdvėje taip, kad atstumai tarp gautų vektorių būtų kiek galima artimesni atstumams tarp daugiamačių vektorių.

MDS pirmiausia buvo pritaikytas psichometrijoje – žmonių nuomonių apie tam tikrų objektų panašumą analizei. 1952 m. Torgesonas (*Torgeson*) pirmasis savo darbuose apibrėžė MDS sąvoką ir pasiūlė pirmąjį šio tipo analizės modelį – kiekybinį Euklido modelį (*Euclidean model*), kuris buvo taikomas duomenims išmatuotiems intervalų arba santykių skalėse [38].

Antrasis svarbus žingsnis MDS raidoje buvo kokybinių modelių atsiradimas, kurie taikomi ranginiams kintamiesiems. 1962–1964 metais Šepardas (*Shepard*) ir Kruskalas (*Kruskal*) paskelbė eilę straipsnių kokybinių MDS modelių tema. Toliau sekė individualių skirtumų (*Individual Differences*) MDS modelių atsiradimas. 1970 m. Carroll ir Chang pristatė INDSCAL programą, realizuojančią svorinį MDS modelį (*Weighted model*). 1977 m. Takane, Young ir de Leeuw pristatė vieną populiariausių MDS algoritmų – ALSCAL, kuris realizuoja kiekybinius ir kokybinius

modelius. Šiuo metu MDS metodai plačiai taikomi marketinge, sociologiniuose tyrimuose, politikos moksluose ir kt.

9.2. MDS modeliai

Matrica “**objektai – savybės**” (matavimo $n \times p$) transformuojama į **artumų matricą** (matavimo $m \times m$) nurodžius artumo *matą*, pagal kurį apskaičiuojamas artumas tarp objektų. Kai $m = n$ poromis lyginami objektai, kai $m = p$ poromis lyginamos savybės. Klasikinis MDS modelis naudoja Euklido atstumą (*Euclidian*)

$$d_{ij} = \sqrt{\sum (x_{ia} - x_{ja})^2},$$

čia x_{ia} ir x_{ja} žymi taškų i ir j koordinates a -oje dimensijų erdvėje. Daugumoje MDS modelių dviejų objektų O_i ir O_j artumas ρ_{ij} ir atsumas tarp jų MDS erdviniame žemėlapyje d_{ij} yra susieti sąryšiu:

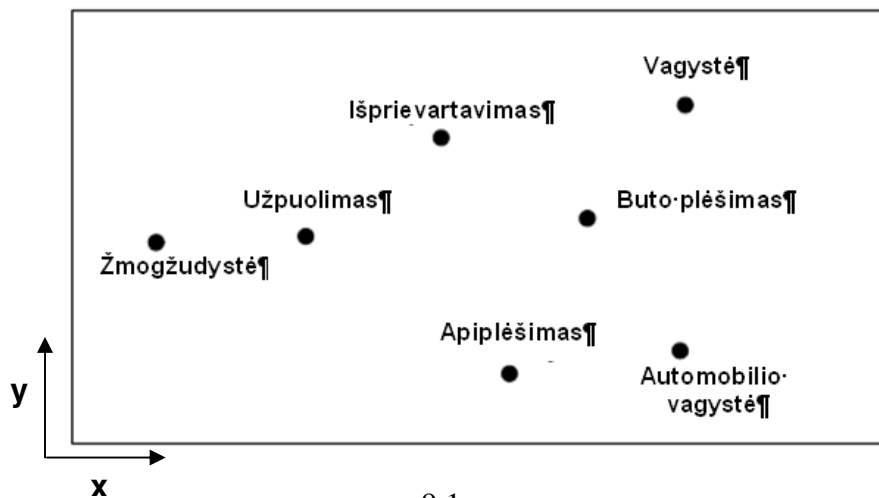
$$\rho_{ij} = f(d_{ij}).$$

Kaip randamas erdvinis žemėlapis, geriausiai atitinkantis artumus? Šis uždavinys sprendžiamas naudojant optimizavimo metodus. MDS išdėsto taškus, kurių kiekvienas žymi objektą, p -matėje erdvėje tam tikra tvarka. Kiekvienai taškų porai yra skaičiuojamas Euklido atstumas. Apskaičiuoti atstumai erdviniame žemėlapyje yra lyginami su pradiniais artumais ir skaičiuojama Stress funkcijos reikšmė, kuri parodo kaip blogai modelis atitinka pradinius duomenis (Stress funkciją detaliau nagrinėsime vėlesniuose skyreliuose). Taškų koordinatės p -dimensijų erdvėje transformuojamos taip, kad Stress funkcijos reikšmė būtų minimali. Iteracinis procesas kartojamas tol, kol Stress reikšmę gauname mažesnę už nurodytą.

1 pavyzdys. Tarkime mus domina, kokiais kriterijais žmonės vadovavosi lygindami tam tikrų automobilių modelius. Kiekvieno respondento buvo prašoma palyginti automobilius nurodytose skalėse. Turėdami tokius lyginamuosius duomenis, ir panaudodami MDS, galime nustatyti apklausos rezultatus aprašančią struktūrą. Pavyzdžiui, nustatytume, kad respondentų rezultatus geriausiai charakterizuoja dvimatė erdvė, kurios viena dimensija yra automobilio kaina, o kita – dydis. Taigi, respondentai lygindami automobilius, vadovavosi automobilio kainos ir dydžio kriterijais. Priklausomai nuo to, kaip automobilių modeliai išsidėsto kainos ir dydžio erdvėje, automobilius galime skirstyti į kategorijas, bei atlikti tolimesnę duomenų analizę.

9.1 lentelė

Nusikaltimas	Nr.	1	2	3	4	5	6	7
Žmogžudystė	1	1,00	0,52	0,34	0,81	0,28	0,06	0,11
Išprievartavimas	2	0,52	1,00	0,55	0,70	0,68	0,60	0,44
Apiplėšimas	3	0,34	0,55	1,00	0,56	0,62	0,44	0,62
Užpuolimas	4	0,81	0,70	0,56	1,00	0,52	0,32	0,33
Buto plėšimas	5	0,28	0,68	0,62	0,52	1,00	0,80	0,70
Vagystė	6	0,06	0,60	0,44	0,32	0,80	1,00	0,55
Automobilio vagystė	7	0,11	0,44	0,62	0,33	0,70	0,55	1,00



9.1 pav.

2 pavyzdys. 9.1 lentelėje pateikti korelacijų koeficientai tarp įvairių nusikaltimų lygių. Tarkime mus domina, kaip žmogžudysčių skaičiaus augimas priklauso nuo butų plėšimų skaičiaus augimo. Dalinai į šį klausimą galime atsakyti pasinaudododę korelacijų matrica (9.1 lentelė). Tačiau korelacijų matrica neparodo bendros struktūros, kuri aprašo nusikaltimų tarpusavio priklausomybę. Raskime dvimatį 9.1 lentelėje pateiktų korelacijos koeficientų erdvinį žemėlapi panaudodami MDS (9.1 pav.). Kiekvienas nusikaltimas atinka tašką dvimatėje erdvėje. Taškai erdvėje yra išdėstomi taip, kad atstumai tarp jų kuo tiksliau atitiktų nusikaltimų tarpusavio korelaciją. Atstumas tarp dviejų taškų yra mažas (pvz. žmogžudystė ir užpuolimas), jeigu nusikaltimų tarpusavio korelacija yra didelė. Priešingai, atstumas yra didelis (pvz. užpuolimas ir vagystė), jeigu korelacija yra maža. Tarkime, kad duomenis apdorojome su MDS programa, kurios rezultatas – optimaliausias sprendinys. Sprendinio tikslumas gali būti patikrintas labai paprastai – reikia išmatuoti atstumus tarp taškų erdvėje ir palyginti juos su duota korelacijų matrica. Jeigu atstumai ir korelacijų koeficientai stipriai koreliuoja tarpusavyje, tuomet gauta erdvinė konfigūracija **X** gerai atitinka pradinę korelacijų matricą tiesine prasme. Mūsų atveju pastebėsime, kad atstumai 9.1 pav. gana gerai atitinka korelacias pateiktas 9.1 lentelėje.

Ką mums parodo 9.1 pav.? Pagal horizontalią ašį (*x* ašį) nusikaltimus galėtume grupuoti į smurtinius (žmogžudystė, išprievartavimas ir užpuolimas) ir turtinius (vagystė, buto plėšimas ir automobilio vagystė). Iš grafiko matome, kad aplėšimą galėtume priskirti tiek smurtiniams, tiek turtiniams nusikaltimams pagal horizontalią ašį. Be to, turtiniai nusikaltimai yra labiau nutolę vienas nuo kito nei smurtiniai nusikaltimai, t.y. labiau „išsibarstę“ erdvėje. Pagal vertikalią ašį (*y* ašį) nusikaltimus galima būti grupuoti į „paslėptus“ ir „gatvės“. Dažnai sprendžiant MDS uždavinį reikalinga ne tik interpretuoti erdvės dimensijas, bet ir nustatyti, kokias grupes ar klasterius suformuoja stebimi objektai. MDS rezultatai taip pat gali parodyti, kad objektai suformuoja tam tikrą sritį ar paviršius erdvėje. Taigi, interpretuodami erdvinio žemėlapio dimensijas, nustatome stebimus objektus vienijančią struktūrą.

MDS modeliai yra klasifikuojami pagal keletą požymių:

- 1) Pagal analizuojamų duomenų tipą:
 - **Kiekybiniai modeliai** analizuoja intervalų arba santykių skalėje išmatuotus duomenis (pvz. geometriniai atstumai tarp miestų);
 - **Kokybiniai modeliai** analizuoja kokybinius duomenis;
- 2) Pagal artumų matricų skaičių:
 - **Klasikinis modelis** (*Classical MDS*): viena artumų matrica,
 - **Atkartojimo modelis** (*Replicated MDS*): keletas artumų matricų,

- **Svorinis modelis** (*Weighted MDS/Individual Difference Scaling*): keletas artumų matricių, kurių kiekvienos svarba apibūdinama jos svoriu.

MDS taikymo problemos:

- dimensijų interpretavimas yra neapibrėžtas ir labai priklauso nuo tyrėjo, pirminės informacijos ir kitų išorinių faktorių;
- jeigu nagrinėjamoje erdvėje yra daugiau negu trys dimensijos, tuomet interpretavimas yra sudėtingas.

9.3. MDS etapai

MDS vykdymo etapų schema pateikta 9.2 pav. Sekančiose skyreliuose aptarsime MDS etapus detaliau.



9.2 pav. MDS pagrindiniai etapai

9.3.1. Duomenų rinkimas

MDS duomenys gali būti renkami įvairiais metodais:

1. Subjektyvūs metodai:

- **Pirmenybių metodas** (*Preference method*). Respondentų paprastai klusiam tokio pobūdžio klausimų: „Ar variantas A yra geriau už variantą B ar variantą C?“.
- **Porinių palyginimų metodas** (*Paired comparisons method*). Respondentų prašome palyginti visas galimas objektų poras: „Palyginkite variantus A ir B skalėje nuo 1 iki 10, kur 1 = visiškai skirtingi, o 10 = visiškai panašūs“.

- **Tiesioginio rangavimo metodas** (*Direct ranking method*). Respondentų prašome suranguoti objektus nuo 1 iki n , kur n – objektų skaičius.
2. **Objektyvūs metodai** (*Objective method*). Šis metodas apima tiesiogiai išmatuojamus duomenis: fiziniai atstumai (*distances*) tarp objektų (pvz., atstumai tarp siuntinių pristatymo vietų), dažniai (*frequencies*) ir pan.

9.3.2. Modelio parinkimas

Kaip jau minėjome, MDS modeliai yra klasifikuojami pagal analizuojamų duomenų tipą (kiekybiniai ar kokybiniai), bei artumų matricų skaičių.

Klasikinis modelis (Classical MDS)

Klasikinis daugiamačių skalių modelis CMDS (*classical MDS*) yra taikomas, kai turime tik vieną artumų matricą. Artumams tarp objektų įvertinti dažniausiai yra naudojamas Euklido atstumas (*Euclidean distance*):

$$d_{ij} = \sqrt{\sum (x_{ia} - x_{ja})^2}.$$

Artumas gali būti įvertintas ir naudojant Minkovskio atstumą (*Minkowski distance*):

$$d_{ij} = \left[\sum_a |x_{ia} - x_{ja}|^u \right]^{1/u},$$

kur u ($u \leq 1$) reikšmė yra parenkama priklausomai nuo sprendžiamo uždavinio.

Nepriklausomai nuo to, kaip vertinsime artumą, erdviniam žemėlapyje bus n taškų, kurių kiekvienas atitiks vieną iš stebimų objektų.

MDS uždaviniams spręsti naudojama matricų algebra. Jeigu turime p -matę erdvę, tuomet visų taškų x_{ia} koordinatės galime pateikti $n \times p$ matrica $\mathbf{X}$. Atstumai d_{ij} sudaro $n \times n$ simetrinę matricą $\mathbf{D}$, o artumai ρ_{ij} - $n \times n$ matricą ρ .

Kiekybinių duomenų atveju, ieškoma tokių atstumų $\mathbf{D}$, kurie kuo geriausiai atitiktų artumus ρ . Praktikoje naudojama keletas metodų šiai atitikčiai nustatyti, o populiariausias iš jų yra mažiausiųjų kvadratų metodas. Taikant šį metodą ieškoma funkcija

$$L(\rho) = \mathbf{D} + \mathbf{E},$$

čia $L(\rho)$ - tiesinė artumų transformacija, o $\mathbf{E}$ – liekamųjų paklaidų matrica. Atstumas d_{ij} gali būti išreikštas koordinačių funkcija, t.y. $d_{ij} = d_{ij}(x)$, todėl CMDS tikslas rasti tokią transformaciją $L(\rho)$ ir koordinačių matricą $\mathbf{X}$, kad liekamųjų paklaidų $\mathbf{E}$ kvadratų sumą būtų minimali.

Jeigu duomenys yra kokybiniai (pavyzdžiui, išmatuoti tvarkos skalėje), ieškoma funkcija:

$$M(\rho) = D + E,$$

čia $M(\rho)$ - monotoninė artumų transformacija. Šiuo atveju CMDS tikslas yra rasti tokią monotoninę transformaciją $M(\rho)$ ir koordinačių matricą $\mathbf{X}$, kad liekamųjų paklaidų $\mathbf{E}$ kvadratų sumą būtų minimali.

Iš esmės, tiek kokybinių, tiek kiekybinių duomenų atveju yra sprendžiama optimizavimo uždavinys, kurios tikslas – minimizuoti skirtumą tarp pradinės artumų matricos ρ ir rezultatų matricos $\mathbf{D}$.

Atkartojimo modelis (*Replicated MDS*)

Atkartojimo MDS modelis - RMDS (*replicated MDS*) yra taikomas, kai turime keletą artumų matricų, t.y. nagrinėjame m artumų matricų ρ , kurių kiekviena aprašo k -tos respondentų grupės ($k = 1, \dots, m$) duomenis. RMDS modelio matematiniai algoritmai yra identiški CMDS modelio algoritmams, tik šiuo atveju jie taikomi daugiau neigui vienai artumų matricai ρ_k . Jeigu analizės duomenys yra kiekybiniai, tuomet ieškoma transformacija

$$L_k(S_k) = D + E_k,$$

čia $L_k(\rho_k)$ - tiesinė k -osios artumų matricos transformacija, kuri geriausiai atitinka atstumus $\mathbf{D}$.

Šiuo atveju RMDS algoritmas minimizuoja visų liekamųjų paklaidų matricų E_k kvadratų sumas.

Kokybiniais duomenimis naudojama tokia pati transformacija kaip ir CMDS modelyje:

$$M_k(\rho_k) = D + E_k,$$

čia $M_k(\rho_k)$ - monotoninė artumų matricos transformacija, kuri geriausiai atitinka atstumus D .

Pastebėsime, kad RMDS modelyje kiekviena artumų matrica ρ_k , yra vienareikšmiškai tiesiškai arba monotoniškai susijusi su atstumais $\mathbf{D}$. Kadangi k kinta nuo 1 iki m , todėl turime rasti m skirtingų tiesinių ar monotoninių transformacijų kiekvienai artumų matricai ρ_k . RMDS modelyje visos artumų matricos yra susijusios tarpusavyje per atstumų matricą $\mathbf{D}$, kuri yra bendra visoms transformacijoms $M_k(\rho_k)$.

Svorinis modelis (*Weighted MDS*)

Svorinis MDS metodas WMDS (*weighted MDS*) taikomas keletui artumų matricų ρ_k , kurių kiekvienos svarba yra nevienoda ir apibrėžiama tam tikru svorio koeficientu. Dėl šios priežasties WMDS yra dažnai vadinamas individualių skirtumų MDS modeliu – INDSCAL. Artumui tarp objektų įvertinti dažniausiai yra naudojamas svorinis Euklido atstumas (*weighted Euclidean distance*):

$$d_{ij} = \sqrt{\sum_a w_{ka} (x_{ia} - x_{ja})^2},$$

čia w_{ka} - k -osios artumų matricos ρ_k svorio koeficientas a -osios dimensijos atžvilgiu. Svorio koeficientas parodo a -osios dimensijos svarbą nagrinėjamam objektui. WMDS modelis dalinai panašus į RMDS modelį: WMDS modelyje yra generuojama k skirtingų atstumų matricų D_k , o RMDS modelyje – viena atstumų matrica D , kuri yra bendra visoms artumų matricoms ρ_k . Taikant WMDS modelį, ieškoma atitikties funkcija tarp visų matricų D_k ir ρ_k .

WMDS modelis kiekybiniais duomenimis:

$$L_k(\rho_k) = D_k + E_k,$$

WMDS modelis kokybiniais duomenimis:

$$M_k(\rho_k) = D_k + E_k.$$

Taigi WMDS pagrindinis tikslas yra toks pat kaip ir ankščiau minėtų MDS modelių – nustatyti tokias atitiktis tarp matricių D_k ir ρ_k , kad visų liekanų matricių E_k kvadratų suma būtų minimali.

9.3.3. Erdvės dimensijų skaičiaus nustatymas

Visų MDS metodų pagrindinis tikslas yra rasti tokią erdvinio žemėlapių koordinačių matricę $\mathbf{X}$, kuri geriausiai tenkintų sąlygą:

$$f(\rho_{ij}) \approx d_{ij}(X).$$

Tačiau, pavyzdžiui, jeigu erdvinis žemėlapis yra 8 dimensijų, tuomet rezultatus interpretuoti yra sudėtinga. Praktikoje optimalus dimensijų skaičius parenkamas pagal keletą kriterijų.

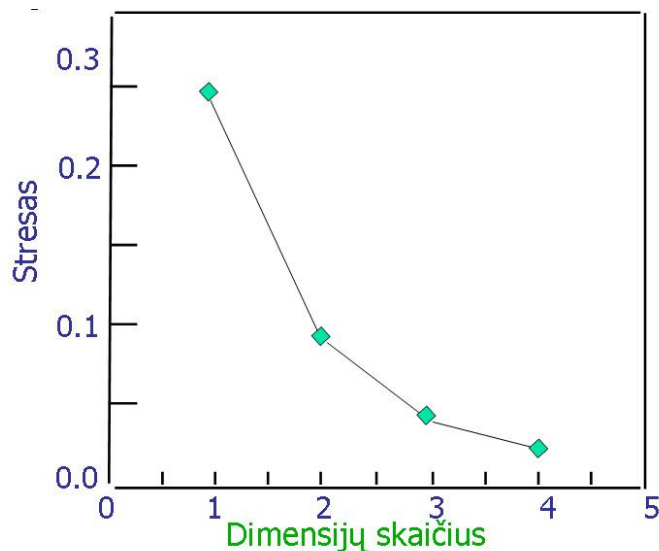
Pirminės informacijos analizė. Dažnai iš teorijos ar ankstesnių tyrimų yra žinoma, kiek dimensijų reikia tam tikram reiškiniui aprašyti.

Erdvinio žemėlapių (*spatial map*) interpretavimo sudėtingumas. Tai vienas iš pagrindinių kriterijų. Paprastai daugiau negu trijų dimensijų žemėlapių interpretacija yra sudėtinga, todėl rekomenduojama apsiriboti 2 arba 3 dimensijomis.

Paklaidų funkcijos, kuri dažnai vadinama Stress funkcija reikšmių grafiko analizė. Stress funkcijos reikšmės apskaičiuojamas pagal formulę:

$$STRESS = \sqrt{\frac{\sum \sum (f(x_{ij}) - d_{ij})^2}{\sum \sum d_{ij}^2}}.$$

Stress funkcijos grafikas padeda nustatyti, ar MDS modelis analizuojamiems duomenims tinka, koks yra optimalus dimensijų skaičius.



9.3. pav. Stress funkcijos grafikas

9.3 pav. pateiktas Stress funkcijos grafiko pavyzdys. Rekomenduojama pasirinkti tokį dimensijų skaičių, kad Stress funkcijos (paklaidų funkcijos) reikšmė būtų ne didesnė už 0,1. Praktikoje dažniausiai ieškoma kompromiso tarp dimensijų skaičiaus ir Stress funkcijos reikšmės dydžio. 9.3 pav. matome, kad mažesnę Stress funkcijos reikšmę atitinka didesnis dimensijų skaičius. Šiuo atveju optimaliausia būtų pasirinkti dviejų dimensijų erdvę.

9.3.4. Rezultatų interpretavimas

Erdvinio žemėlapių dimensijų interpretavimas yra sudėtingas procesas, nes labai priklauso nuo paties tyrėjo kvalifikacijos, respondentų atsakymų kokybės ir pan. Interpretuojant dimensijas reikėtų vadovautis tokiais principais:

- respondentams pateikus savo atsakymus apie stebimą reiškinį, reikėtų siekti išsiaiškinti, kokiai kriterijais vadovaudamiesi jie vertino stebėtą reiškinį;
- jeigu įmanoma, kiekvienam tyrimo dalyviui reikėtų pateikti MDS analizės metu gautą erdvinį žemėlapi, ir paprašyti jį interpretuoti;
- jeigu tyrimo metodu buvo naudoti objektyvūs (metriniai) vertinimo kriterijai, pavyzdžiui, greitis, aukštis, spalva ar pan., tuomet lengviau interpretuoti erdvinį žemėlapi.

9.3.5. Modelio tinkamumo tyrimas

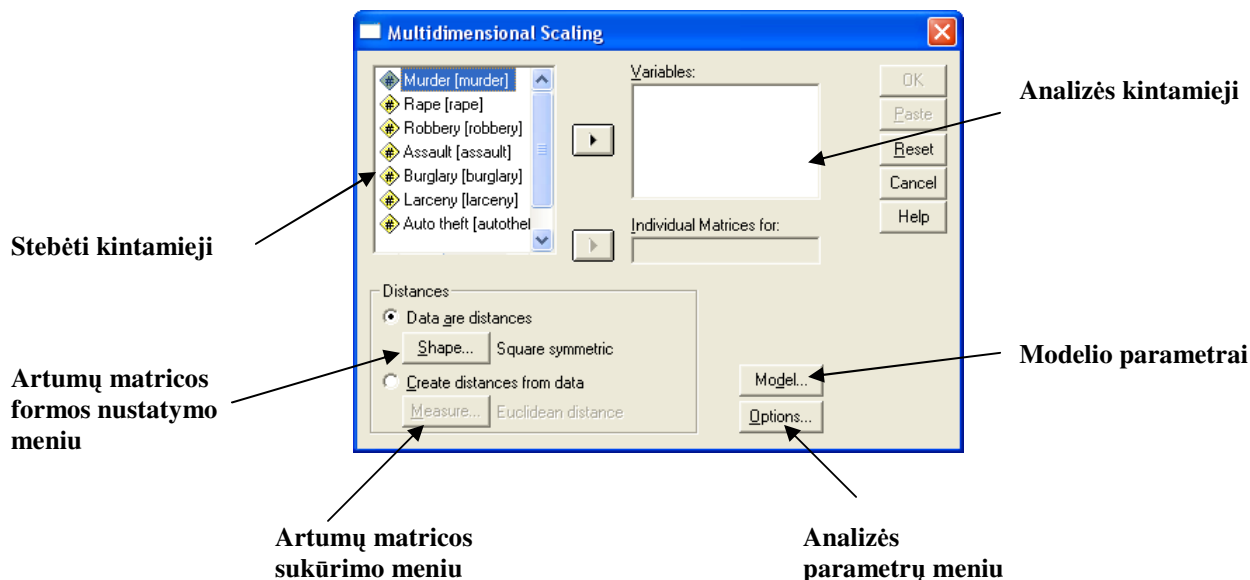
Pagrindiniai modelio tinkamumo kriterijai yra:

- **Apibrėžtumo koeficientas R^2 (*R-Square*)** parodo, kokią pradinių kintamųjų dispersijos dalį paaiškina taikytas MDS modelis. Praktikoje priimama, kad R^2 reikšmė turėtų būti didesnė už 0,60.
- **Stress koeficientas** taip pat parodo gauto modelio kokybę.– kaip blogai modelis tinka duomenims. Kitaip tariant stress reikšmė parodo, kokia pradinių kintamųjų dispersijos dalis nepaaiškinama modeliu. Rekomenduojama, kad stress reikšmė būtų mažesnė už 0,15, o geriausia kai ji neviršija 0,1.
- **Šepardo diagrama (*Shepard Diagram*)** – tai taškų sklaidos diagrama, kuri parodo kaip atstumai tarp objektų mažesnio matavimo erdvėje (MDS metodu gauti atstumai) atitinka pradinius artumus tarp objektų. Diagramoje artumai tarp objektų atidedami X-ašyje, o MDS metodu gauti atstumai atidedami Y-ašyje. Jeigu naudojami metriniai artumo matai ir diagramos taškai išsidėsto arti tiesės, tuomet mažesnio matavimo erdvėje gerai išlaikyti pradiniai artumai tarp objektų. Jeigu naudojami nemetriniai artumo matai, tuomet priklausomybė turi būti monotoninė (monotoniškai didėjanti arba mažėjanti).

9.4. MDS pavyzdys

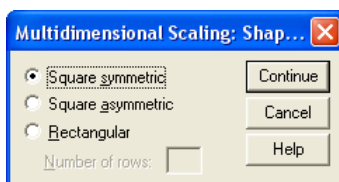
Pateiksime uždavinio sprendimo su SPSS ir STATA pavyzdžius. Pradžioje detaliau paaiškinsime, kaip sprendžiami MDS uždaviniai su SPSS. 9.4 pav. pateiktas pagrindinis SPSS daugiamačių

skalių metodo langas, kuriame parenkami analizės kintamieji ir nustatomi parametrai (*Analyze ➔ Scale ➔ Multidimensional Scaling...*).



9.4 pav. Pagrindinio MDS analizės langas

Artumų matricos formos nustatymo meniu langas (9.5pav.) atveriamas MDS analizės pagrindiniame lange (9.4 pav.) pažymėjus **Data are distances** (*Duomenys yra atstumai*) laukelį ir nuspaudus mygtuką **Shape...** (*duomenų matricos forma*).

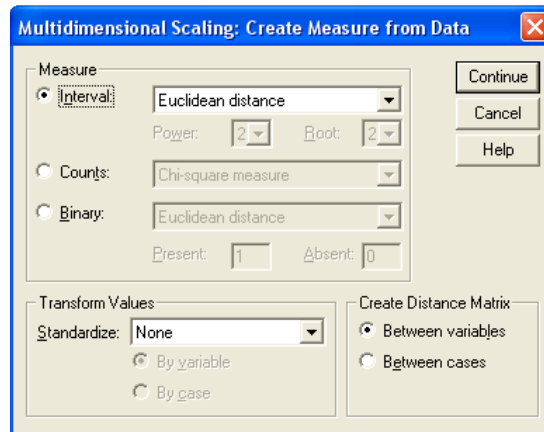


9.5 pav. Artumų matricos formos nustatymo langas

MDS duomenų matricos, priklausomai nuo sprendžiamo uždavinio, gali būti kelių formų:

- **Square symmetric** (*kvadratinė simetrinė*). Ši forma nustatyta pagal nutylėjimą. Eilutės ir stulpeliai atitinka tuos pačius objektus.
- **Square asymmetric** (*kvadratinė asimetrinė*). Eilutės ir stulpeliai atitinka tuos pačius objektus, tačiau atstumas tarp dviejų objektų A ir B yra nebūtinai lygus atstumui tarp B ir A.
- **Rectangular** (*stačiakampė*). Duomenų matrica yra nestandartinės formos. Jeigu duomenų matricą sudaro kelios matricos, reikia nurodyti kiek eilučių sudaro vieną matricą. Pastebėsime, kad eilučių skaičius turi būti didesnis už 4, ir visų matricų formatas vienodas.

Artumų matricos sukūrimo meniu langas (9.6 pav.) atveriamas pagrindiniame lange (9.4 pav.) pažymėjus **Create distances from data** (*Sukurti atstumus pagal duomenis*) ir nuspaudus mygtuką **Measure...**



9.6 pav. Artumų matricos sukūrimo langas

Šiame meniu nustatomi artumų matricos sudarymo parametrai. Artumo mato parinkimas priklausomai nuo nuo pardinių duomenų matavimo skalės :

- **Interval** (kintamieji išmatuoti intervalų arba santykių skalėje). **Euclidean distance** (Euklido atstumas) nustatytas pagal nutylėjimą. Kiti matai: Euklido stdtumo kvadratas (Squared Euclidean distance), Čebyšovo atstumas (Chebychev distance) ir kt.
- **Count** (Dažniai) **Chi-square measure** nustatytas pagal nutylėjimą.
- **Binary** (kintamieji išmatuoti dvejetainėje skalėje). **Euclidean distance** nustatytas pagal nutylėjimą.

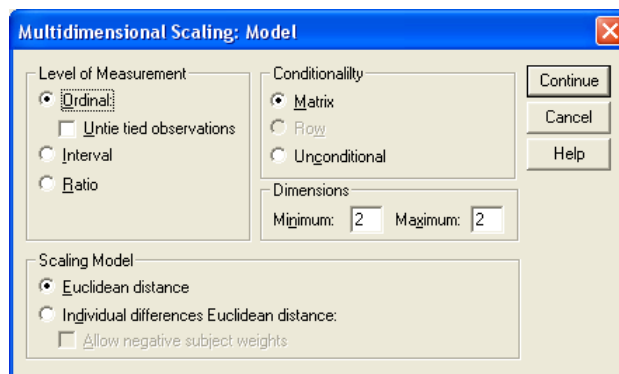
Kuriant artumų matricą (**Create Distance Matrix**) nurodoma kokie artumai skaičiuojami - tarp objektų ar tarp savybių:

- **Between variables** (tarp kintamųjų, t.y. tarp objektų savybių). Artumai skaičiuojami tarp kintamųjų, t.y. tarp duomenų matricos stulpelių.
- **Between cases** (tarp atvejų, t.y. tarp objektų). Atstumai skaičiuojami tarp atvejų, t.y. tarp duomenų matricos eilučių. Šiuo atveju, artumai skaičiuojami tarp objektų.

Duomenų transformavimas (**Transform values**) atliekamas, jeigu duomenų matavimo skalės yra skirtingos. Pagal nutylėjimą duomenys standartizuojami į **Z scores** (*z-reikšmės*). Tokiu būdu standartizuotų duomenų vidurkis yra lygus 0, o dispersija – 1. Kiti standartizavimo būdai: intervalas nuo -1 iki 1 (Range -1 to 1), intervalas nuo 0 iki 1 (Range 0 to 1) ir kt. Standartizavimą galima atlikti:

- **By variable** (pagal kintamuosius).
- **By cases** (pagal atvejus (objektus)).

Modelio parametrų meniu langas (9.7 pav.) atveriamas pagrindiniame analizės lange (9.4 pav.) nuspaudus mygtuką **Model...** .



9.7 pav. Modelio parametrų nustatymo langas

Šiame meniu yra nustatomi MDS modelio parametrai.

Artumų matavimo skalės nustatymas (**Level of Measurement**):

- **Ordinal** (tvarkos skalė). Jeigu pažymimas laukelis **Untie tied observations**, tuomet tvarkos skalėje išmatuoti duomenys yra laikomi nesuietais).
- **Interval** (intervalų skalė).
- **Ratio** (santykių skalė).

Palyginimų reikšmingumo nustatymas (**Conditionality**):

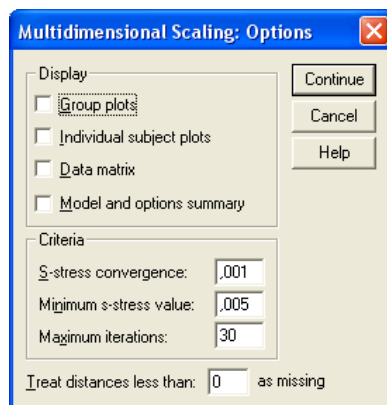
- **Matrix** (matrica). Nustatyta pagal nutylėjimą.
- **Row** (eilutė)
- **Unconditional** ()

Dimensijų skaičiaus nustatymas (**Dimensions**). Nustatoma, kiek dimensijų sudarys MDS analizės sprendinį. Kiekvienam dimensijų skaičiui, patenkančiam į nurodytą intervalą (nurodytam **Minimum** ir **Maximum** laukeliuose), skaičiuojamas atskiras sprendinys. Maksimalus dimensijų skaičius – 6, minimalus – 1. Tačiau 1 dimensiją galima nurodyti tik tuo atveju, jeigu analizei naudojamas **Euclidean distance** modelis. Norint gauti vienintelį sprendinį, reikia įrašyti tą patį skaičių **Minimum** ir **Maximum** laukeliuose.

Scaling model

- **Euclidean distance** (Euklido modelis). Nustatytas pagal nutylėjimą. Apjungia CMDS ir RMDS modelius (žiūr. 9.3.3.).
- **Individual differences Euclidean distance** (individualių skirtumų Euklido modelis). Svorinis WMDS modelis (žiūr. 9.3.3.).

Analizės parametrų meniu langas (9.8 pav.) atveriamas pagrindiniame analizės lange (9.4 pav.) nuspaudus mygtuką **Options...** (Parametrai).



9.8 pav. Analizės parametrų nustatymo langas

Šiame meniu nustatomi MDS analizės ir rezultatų pateikimo parametrai.

Rezultatų pateikimo užsakymas (**Display**):

- **Group plots.** Rezultatų lange bus pateiktas: **MDS map** (erdvinis žemėlapis), **Scatterplot of linear fit** (modelio tinkamumo duomenims grafikas – Šepardo sklaidos diagrama). Kokybiniais modeliams papildomai generuojami **Plot of nonlinear fit** ir **Plot of transformation**.
- **Individual subject plots.** Jeigu analizės duomenys yra išmatuoti tvarkos skalėje ir **Model** meniu nustatyta **Matrix**, pateikiamas grafikas kiekvienam subjektui.
- **Data matrix** (duomenų matrica). Rezultatų lange bus pateikta pradinė duomenų matrica ir MDS atstumų matrica, kurios elementai – objektų koordinatės erdviniam žemėlapyje.
- **Model and options summary** (Modelio ir parametrų suvestinė). Rezultatų lange pateikiama informacija apie analizės duomenis, rezultatus, naudotą modelį ir pan.

Iteracinio algoritmo konvergavimo parametrai nustatomi langeliuose (**Criteria**). Jeigu norime neitraukti į analizę artumų mažesnių už tam tikrą skaičių, turime tą skaičių įrašyti į langelį **Treat distances less than n as missing**.

Pateiksime MDS taikymo pavyzdį. Faile `automobiliai.sav` pateiktos 6 charakteristikos kiekvienam automobiliui:

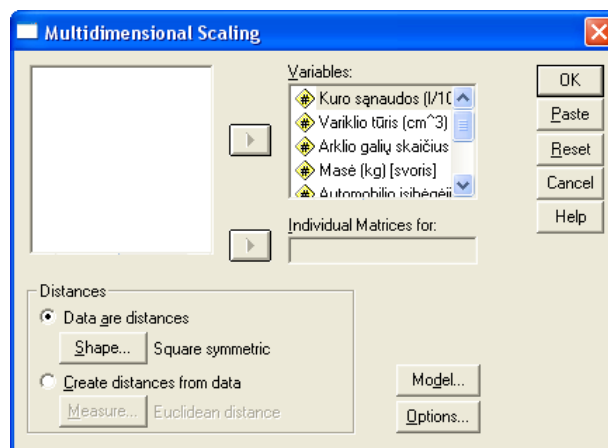
- *sanaudos* = „Kuro sąnaudos (l/100 km)“,
- *variklis* = „Variklio tūris (cm³)“,
- *galia* = „Arklio galių skaičius“,
- *svoris* = „Masė (kg)“,
- *akceleracija* = „Įsibėgėjimo nuo 0 iki 100 km/h laikas (s)“,
- *cilindrai* = „Cilindrų skaičius“.

Duomenų failo fragmentas pateiktas 9.9 pav. Panaudoję daugiamačių skalių analizę (MDS) sumažinsime dimensijų skaičių.

	sanaudos	variklis	galia	svoris	akceleracija	cilindrai	var
1	13,07	1049	130	1589	12	8	
2	15,68	5735	165	1675	12	8	
3	13,07	5211	150	1559	11	8	
4	14,70	4982	150	1557	12	8	
5	13,84	4949	140	1564	11	8	
6	15,68	7030	198	1969	10	8	
7	16,80	7440	220	1975	9	8	
8	16,80	7210	215	1956	9	8	
9	16,80	7456	225	2007	10	8	
10	15,68	6391	190	1746	9	8	
11	.	2179	115	1402	18	4	
12	.	5735	165	1879	12	8	
13	.	5752	153	1830	11	8	
14	.	6276	175	1890	11	8	
15	.	5899	175	1746	11	8	
16	15,68	6276	170	1616	10	8	
17	16,80	5572	160	1637	8	8	
18	.	4949	140	1521	8	8	
19	15,68	6555	150	1706	10	8	
20	16,80	7456	225	1400	10	8	
21	9,80	1852	95	1076	15	4	

9.9 pav. Duomenų failo fragmentas

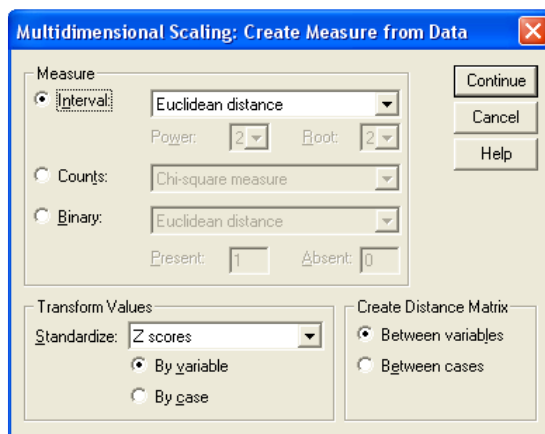
Atverkime pagrindinį SPSS daugiamačių skalių (MDS) analizės langą (*Analyze ➔ Scale ➔ Multidimensional Scaling*).



9.10 pav. MDS analizės pagrindinis langas

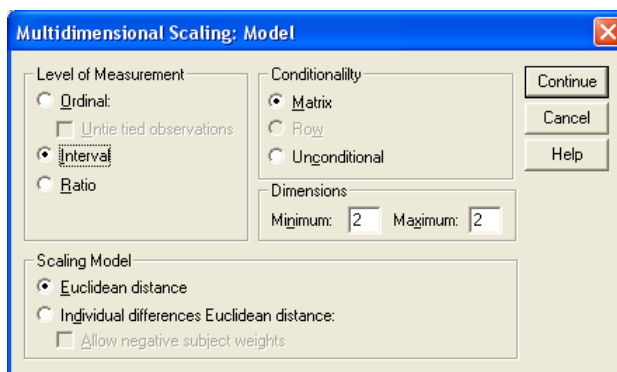
Kintamuosius *sanaudos*, *variklis*, *galia*, *svoris*, *akceleracija* ir *cilindrai* perkeliame į langelį **Variables** (MDS analizės kintamieji) (9.10 pav.).

Pagrindiniame analizės lange (9.10 pav.) pažymėkime **Create distances from data** (artumus apskaičiuoti pagal duomenis) ir atverkime **Measure...** (artumo mato parinkimo) langą.



9.11 pav. Artumo mato parinkimo langas

Artumo mato parinkimo lange (9.11 pav.) pažymėkime **Interval** (kintamieji išmatuoti intervalų arba santykių skalėje) ir pasirinkime **Euclidean distance** (Euklido atstumas). Nagrinėsime artumus tarp objektų savybių, todėl pažymėkime **Between variables** (artumai skaičiuojami tarp kintamųjų, t.y. tarp duomenų matricos stulpelių). Analiziuojami kintamieji yra skirtingos eilės dydžiai, todėl juos reikia standartizuoti. **Standardize** laukelyje nurodykite **Z scores** (standartizuotų duomenų vidurkis yra lygus 0, dispersija – 1) ir pažymėkime **By variable** (standartizuoti pagal kintamuosius).

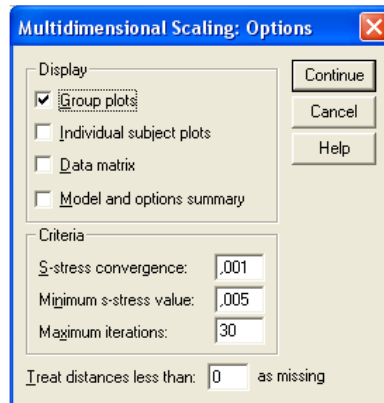


9.12 pav. MDS modelio parametrų langas

Model... (MDS modelio parametrai) meniu (9.12 pav.) pažymėkime:

- **Interval** (artumai išmatuoti intervalų skalėje),
- **Matrix** (matrica),
- **Euclidean distance** (Euklido atstumo modelis).

Šiame pavyzdyje nagrinėsime tik dviejų dimensijų erdvinį žemėlapi, todėl įrašykime **2** į **Minimum** ir **Maximum** laukelius.



9.13 pav. Analizės parametrų nustatymo langas

Options (analizės parametrai) meniu (9.13 pav.) pažymėkime:

- **Group plots** (Rezultatų lange bus matomi: MDS map (erdvinis žemėlapis), Scatterplot of linear fit (modelio tinkamumo duomenims grafikas – Šepardo diagrama).

Criteria (iteracinio algoritmo konvergavimoparametrai) laukeliuose visas reikšmes paliekame nepakeistas.

Nustatę visus parametrus, pagrindiniame MDS analizės lange (9.10 pav.) spauskime **OK**. Gauti rezultatai pateikti 9.14-9.16 pav.

Iteration history for the 2 dimensional solution (in squared distances)

Young's S-stress formula 1 is used.

Iteration	S-stress	Improvement
1	,01104	
2	,00723	,00381
3	,00561	,00162
4	,00505	,00056

Iterations stopped because

S-stress improvement is less than ,001000
Stress and squared correlation (RSQ) in distances

RSQ values are the proportion of variance of the scaled data (disparities)

in the partition (row, matrix, or entire data) which is accounted for by their corresponding distances.

Stress values are Kruskal's stress formula 1.

For matrix
Stress = ,03246 RSQ = ,99808

9.14 pav. MDS rezultatai (1 dalis)

Daugiamačių skalių analizės sprendiniui gauti prireikė 3 iteracijų (9.14 pav.). Kiekvienoje iteracijoje paklaida (Stress) buvo minimizuojamas taip, kad atstumai tarp savybių mažesnio matavimo erdvėje kuo tiksliau atitiktų pradinius artumus tarp savybių. Paklaidos (Stress) yra 0,03246 ($< 0,1$), o apibrėžtumo koeficientas **RSQ** = 0,99808 ($> 0,6$), todėl modelis labai gerai tinka

analizuojamiems duomenims. Didelė apibrėžtumo koeficiento reikšmė parodo, kad gautas modelis paaiškina beveik visą pradinį kintamųjų dispersiją.

Configuration derived in 2 dimensions

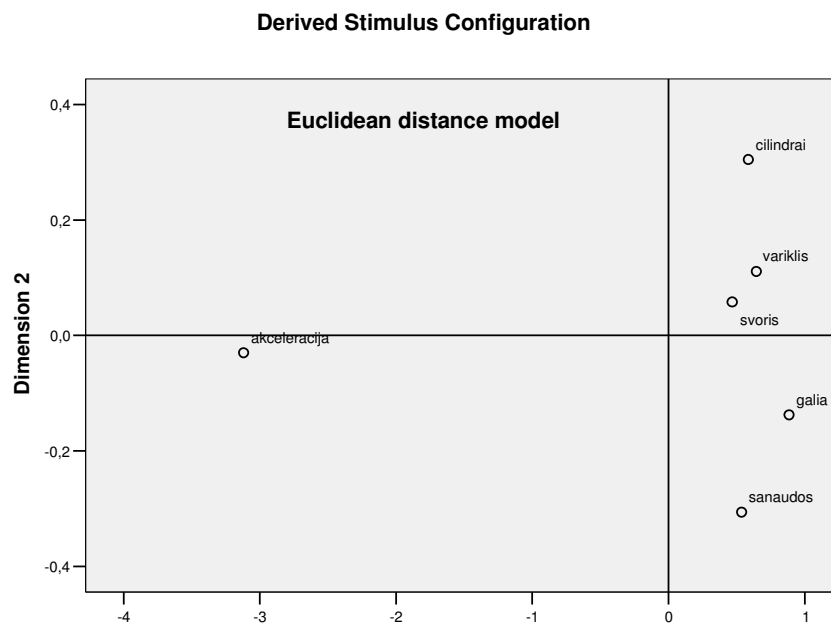
Stimulus Coordinates

Stimulus Number	Stimulus Name	Dimension	
		1	2
1	sanaudos	,5364	-,3062
2	variklis	,6439	,1109
3	galia	,8855	-,1375
4	svoris	,4674	,0579
5	akcelera	-3,1192	-,0299
6	cilindra	,5861	,3047

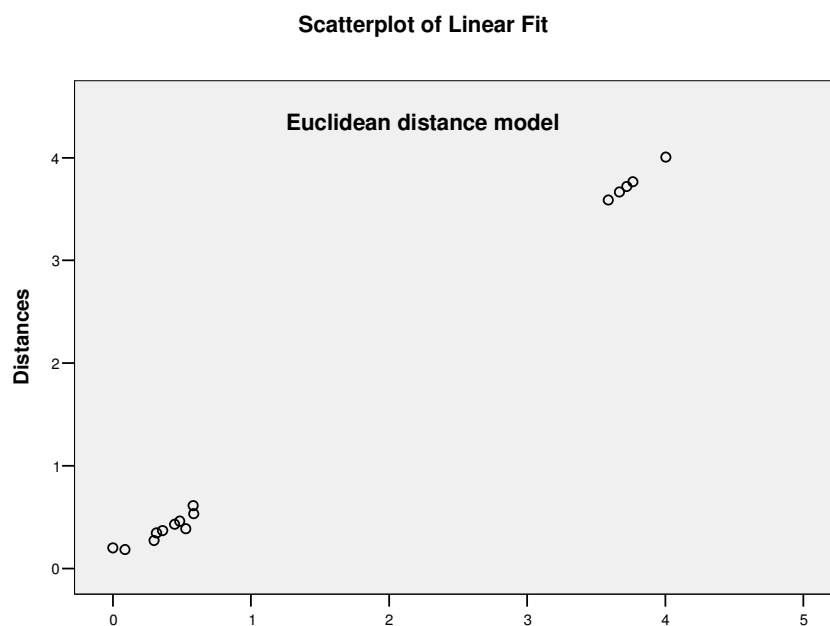
Abbreviated Name	Extended Name
akcelera	akceleracija
cilindra	cilindrai

9.15 pav. MDS rezultatai (2 dalis)

9.15 pav. pateiktos MDS metodu apskaičiuotos savybių koordinatės (*Stimulus Coordinates*) dvimatėje erdvėje. Taškus galime pavaizduoti grafiškai nubraižydami dvimatį erdvinį žemėlapi (9.16 pav.)



9.16 pav. Erdvinis žemėlapis



9.16 pav. Šepardo diagrama

Dviejų dimensijų sprendinys gerai atitinka pradinę duomenų struktūrą. Tą patvirtina Šepardo diagrama (9.16 pav.) Šepardo diagramoje taškai išsidėstę arti tiesės, todėl galime teigti, kad atstumai tarp savybių dvimatėje erdvėje labai gerai atitinka jų atstumus pradinėje, 6 matavimų erdvėje.

LITERATŪRA

1. Howell D.C.(1999) Fundamental Statistics for the Behavioral Sciences.-New York: Duxbury Press.
2. Vogt, W. Paul (2005) Dictionary of statistics & methodology: a nontechnical guide for the social sciences. Thousand Oaks : Sage Publications.
3. Norušis Marija J. (2007). SPSS 15.0 Guide to Data Analysis. Prentice Hall.
4. Norušis Marija J. (2007). SPSS 15.0 Statistical Procedures Companion. Prentice Hall.
5. McClave, James T. (2006) Statistics for business and economics. Upper Saddle River : Pearson Prentice Hall.
6. Gaur, Ajai S (2007) Statistical methods for practice and research: a guide to data analysis using SPSS. New Delhi: Response Books.
7. SPSS Base 15.0 User's Guide (2006). Chicago: SPSS Inc.
8. SPSS 15.0 Algorithms (2006). Chicago: SPSS Inc.
9. Base Reference Manual. STATA 10 (2007). Stata Press.
10. Multivariate Statistics Reference Manual. STATA 10 (2007). Stata Press.
11. User's Guide STATA 10 (2007). Stata Press.
12. Wagner III, William E (2007) Using SPSS for social statistics and research methods. Thousand Oaks : Pine Forge Press.
13. Aksomaitis, Algimantas (2000). Tikimybių teorija ir statistika : vadovėlis aukštųjų mokyklų studentams / Algimantas Aksomaitis. - Kaunas : Technologija.
14. Čekanavičius V., Murauskas G. (2000). Statistika ir jos taikymai. D.1-Vilnius: TEV.
15. Čekanavičius V., Murauskas G. (2004). Statistika ir jos taikymai. D.2-Vilnius: TEV.
16. Statistika ir duomenų analizės programinė įranga. Distancinio mokymo kursas. Projekto vadovas V. Janilionis.- Kaunas: KTU, 1999-2001.-(<http://fmf.ktu.lt/janil/stat1.htm>).
17. Janilionis V.; Aksomaitis A., Jokimaitis A. ir kt. (2003). Tikimybių teorija ir statistika 2. Nuotolinio mokymo kursas. Projekto vadovas V. Janilionis. Kaunas: KTU. (<http://webct.liedm.lt>).
18. Janilionis V., Aksomaitis A., Jokimaitis A. ir kt. (2001). Tikimybių teorija ir statistika 1. Nuotolinio mokymo kursas. Projekto vadovas V. Janilionis. Kaunas: KTU. (<http://distance.ktu.lt:4949>).
19. Bačinskas A., Janilionis V., Jokimaitis (2001). Tikimybių teorija ir statistika. Praktikumai. Kaunas: Technologija.
20. Janilionis V. , Aksomaitis A. (1993) Sistema STATGRAPHICS 5.0. Matematinės statistikos metodai. Kaunas: Technologija.
21. SPSS. <http://www.spss.com> [žiūrėta 2008 03 10].
22. STATA. <http://www.stata.com> [žiūrėta 2008 03 14].
23. SAS. <http://www.sas.com> [žiūrėta 2008 03 12].
24. Anderson, David R. (2007). Statistics for business and economics.Thomson.

25. Huizingh, Eelko (2007) Applied statistics with SPSS. London :Sage publications.
26. Stock, James H. and Mark W.Watson (2007), Introduction to Econometrics. Addison Wesley, Pearson.
27. Evans, James R. (2003). Statistics, data analysis, and decision modeling. Upper Saddle River : Prentice Hall.
28. Сошникова Л. А., Тамашевич В. Н., Уебе Г., Шефер М. (1999) Многомерный статистический анализ в экономике.- Москва:Юнити-ДАНА.
29. Johnson, Richard A. (2002). Applied multivariate statistical analysis. Upper Saddle River : Pearson Education International.
30. Lomax R. G. (1992) Statistical Concepts: A Second Course for Education and the Behavioral sciences. London: Loughman Group Ltd.
31. . Nachmias Ch. F., Nachmias D.(1992) Reserch Methods in the Social Sciences. Fourth edition.- New York: Edward Arnold.
32. Sirkin R. Mark (1995) Statistics for the social sciences.-SAGE Publications Inc.
33. Jain A.K., DubesR.C. (1988). Algorithms for Clustering data. New Jersey: Prentice-Hall Inc.
34. Johnson, Richard A. (2002). Applied multivariate statistical analysis. Upper Saddle River : Pearson Education International.
35. Abdi, H. (2007). Metric multidimensional scaling. In Salkind, N.J., ed.. Encyclopedia of measurement and statistics. Thousand Oaks (CA): Sage.
36. Borg, I., & Groenen, P. (1997). Modern multidimensional scaling. Theory and applications. New York: Springer.
37. Cox, M.F. & Cox, M.A.A., (2001), Multidimensional scaling. London: Chapman and Hall.
38. Torgerson, W. S. (1958). Theory and methods of scaling. NY: Wiley.
39. Young, Forrest W. (1999). Multidimensional scaling. Retrieved 11/17/06 from <http://forrest.psych.unc.edu/teaching/p208a/mds/mds.html> [žiūrėta 2008 04 15].

**III DALIS. STATISTINĖS ANALIZĖS PAVYZDŽIŲ
NAUDOJANT PAVYZDINĘ SKAITMENINĘ
DUOMENŲ BAZĘ MEDŽIAGA**

Autorius:

Dr. Vaidas Morkevičius

(parašas)

KAUNAS, 2008

1. ĮVADAS

Statistinės analizės pavyzdžiai naudojant pavyzdinę skaitmeninę kiekybinių tyrimų duomenų bazę pateikiami pagal įvadinio kurso į statistinę analizę temas. Kiekvienai temai pateikiama pavyzdžių ir nurodomos procedūros, kuriomis sekant vartotojas gali pats atlikti aptariamą analizę su pavyzdinėje skaitmeninėje duomenų bazėje esančiomis laikmenomis. Prieš pateikiant pavyzdį nurodoma laikmena, kurioje esantys duomenys bus naudojami. Kiekvienam pavyzdžiui nurodomos procedūros, kaip analizę atlikti ir SPSS, ir Stata paketuose, jei tai įmanoma. Statistinės analizės atlikimo procedūros iliustruojamos grafiškai ir pateikiama tekstinė informacija paaiškinanti atliktą analizės procedūrą.

Statistinės analizės pavyzdžių naudojant pavyzdinę skaitmeninę kiekybinių tyrimų duomenų bazę medžiagą sudaro 9 skyriai:

Aprašomoji statistika

Pasikliautiniai intervalai

Hipotezių tikrinimas

Vidurkių palyginimas. Dispersinė analizė (ANOVA)

Požymių priklausomumo tyrimas. Koreliacinė analizė

Daugialypė regresinė analizė

Klasterinė analizė

Faktorinė analizė

Daugiamačių skalių analizė (MDS)

Pavyzdžių medžiagos bendra apimtis 188 puslapiai.

2. APRAŠOMOJI STATISTIKA

Jei žiūrėtume apibendrintai, statistika yra mokslas apie **duomenų apie masinius reiškinius** rinkimą, aprašymą ir sisteminimą bei analizavimą ir interpretavimą. Taigi pirmasis statistinės analizės žingsnis jau surinkus duomenis apie tam tikrą masinį reiškinį¹ yra aprašomoji statistika. Kaip teigia Vaitkevičius ir Saudargienė, „vienas iš svarbiausių statistikos uždavinių – aprašyti tam tikro kintamojo reikšmių aibę trumpa ir aiškia forma, išreiškiant svarbiausias tos reikšmių aibės savybes“ (Vaitkevičius ir Saudargienė 2006, p. 46). Paprastaiariant, **aprašomoji statistika – „tai duomenų sisteminimo ir grafinio vaizdavimo metodai“** (Čekanavičius ir Murauskas 2000, p. 25). Tačiau aprašomoji statistika turi ir kitą svarbų aspektą. Išsamiai aprašius ir grafiškai pavaizdavus duomenis dažnai jau galima daryti preliminarias išvadas apie tam tikras nagrinėjamos populiacijos charakteristikas².

Aprašomoji statistika remiasi **skaitinėmis** – grupuotų variacinių sekų (dažnių ir santykinių dažnių lentelių) sudarymas, duomenų padėties, sklaidos ir formos charakteristikų (įvairių indeksų) skaičiavimas – bei **grafinėmis** – informaciją apibendrinančių grafikų braižymas – **priemonėmis**. Tų priemonių naudojimas konkrečioje sistuacijoje priklauso nuo to, kokie statistiniai duomenys yra surinkti. Visų pirma, reikia atsižvelgti į tai, koks turimų kintamųjų tipas: kokybiniai (kategoriniai) ar kiekybiniai kintamieji. Be to, svarbu, kokia skale išmatuoti šie kintamieji. Apibendrintai pasirinkimo galimybės susistemintos 2.1 lentelėje. Toliau kiekviena iš jų bus trumpai apibūdinta ir pademonstruota, kaip ją suskaičiuoti arba pavaizduoti praktiškai naudojant statistinės analizės paketus SPSS ir Stata.

2.1 lentelė. Aprašomosios statistikos priemonės pagal kintamųjų tipus ir matavimo skalę.

Kintamojo tipas	Matavimo skalė	Variacinės sekos	Padėties charakteristikos	Sklaidos charakteristikos	Formos charakteristikos	Grafinės priemonės
Kokybinis	Pavadinimų (nominal)	Dažniai ir santykiniai dažniai	Moda (mode)	Įvairovės (diversity) ir kokybinės variacijos (qualitative variation) indeksai ³	-	Skritulinė (pie chart) ir stulpelinė (bar chart) diagrama
	Rangų arba tvarkos (ordinal)		Mediana (median) ir moda			
Kiekybinis	Intervalinė (interval)	Intervaliniai dažniai, intervaliniai santykiniai dažniai,	Vidurkis (mean), mediana ir moda	Duomenų aibės plotis (range), minimali ir maksimali reikšmė, dispersija (variance), standartinis nuokrypis (standard deviations)	Asimetrija (skewness) Ekscesas (kurtosis)	Histograma (histogram) ir stačiakampė (boxplot) diagrama
	Santykių (ratio)	kvartilai, procentiliai				

¹ Pirmasis statistinio tyrimo žingsnis – duomenų rinkimas – šioje statistinės analizės pavyzdžių medžiagoje neaptariamas, nes tai mažai su kompiuterine statistine analize susijęs dalykas. Tiesa, abu čia aptariami statistinės analizės paketai – SPSS ir Stata – leidžia apskaičiuoti imties dydį ar sudaryti simuliacines imtis, tačiau tai jau platesnio nei įvadinis statistinės analizės kurso objektas.

² O išvadų apie populiaciją, remiantis „tinkamai“ surinktais imties duomenimis, darymas, kaip žinia, yra trečiasis statistinės analizės etapas.

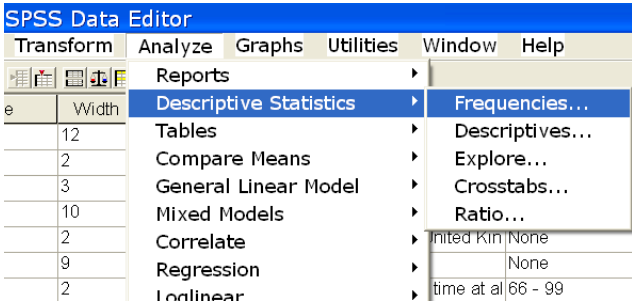
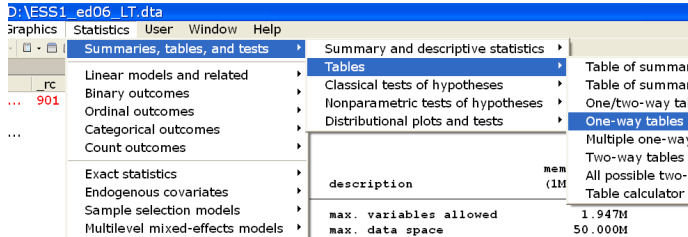
³ Pastarųjų indeksų neskaičiuoja nei SPSS, nei Stata patektai, todėl jie toliau nenagrinėjami, o tik pažymima galimybė tokius indeksus skaičiuoti apskritai. Beje, tą gana nesunku padaryti „rankomis“.

Pateikdami aprašomosios statistinės analizės su SPSS ir Stata programiniais paketais pavyzdžius naudosime Europos socialinio tyrimo (ESS) pirmosios bangos (2002 m.) duomenis (SPSS laikmena – „ESS1_ed06_LT.sav“, Stata laikmena – „ESS1_ed06_LT.dta“). Šio tyrimo ir jame tyrinėjamų reiškinių bei kintamųjų aprašas yra pateiktas pirmojoje studijų paketo dalyje.

2.1. Dažniai, intervaliniai dažniai ir santykiniai dažniai

Pirmiausiai aptarsime **kokybinių kintamųjų** skaitinį statistinį apibendrinimą. Tam iliustruoti panagrinėsime **europiečių⁴ religingumo** rodiklius. Buvo klausta, ar respondentai įvairiose Europos šalyse priskiria save kokiai nors religijai ar denominacijai (atsakymų variantai: taip, ne) ir jei taip tai – kokiai. Kaip matome, abu klausimai matuoja religingumą **pavadinimų skalėje**, t. y. siekiama sužinoti tik religingumo tipus. Kad apibendrinti tokią informaciją ir **parodyti bendras tendencijas**, paprastai sudaromos **dažnių ir santykinų dažnių lentelės⁵**. Pavyzdys, kaip tai padaryti su SPSS ir Stata, yra pateiktas 2.2 lentelėje.

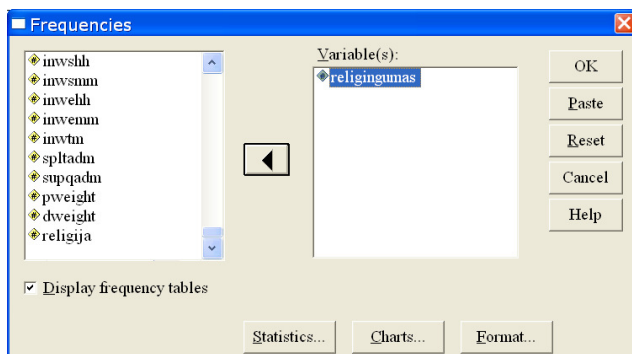
2.2 lentelė. Kokybinių kintamųjų reikšmių pasiskirstymo bendros tendencijos skaitinis vaizdavimas: dažnių ir santykinų dažnių lentelės (su paaiškinimais, kaip jas nukopijuoti).

Sprendimas su SPSS 15.0	Sprendimas su Stata SE 10.0
Parengiamieji veiksmai	
Nėra	
Sintaksės komandos	
Nėra	
Analizės veiksmai	
Meniu punktai: Analyze → Descriptive Statistics → Frequencies 	Meniu punktai: Statistics → Summaries, tables, and tests → Tables → One-way tables 
Atsidariusiame lange „Frequencies“ į dešinėje pusėje esantį langelį „Variables(s)“ įkeliam kintamąjį religingumas ir spaudžiame „OK“.	Atsidariusiame lange „tabulate1 – One-way tables“ parenkame analizės kintamąjį ties „Categorical variable“ religingumas ir spaudžiame „OK“.

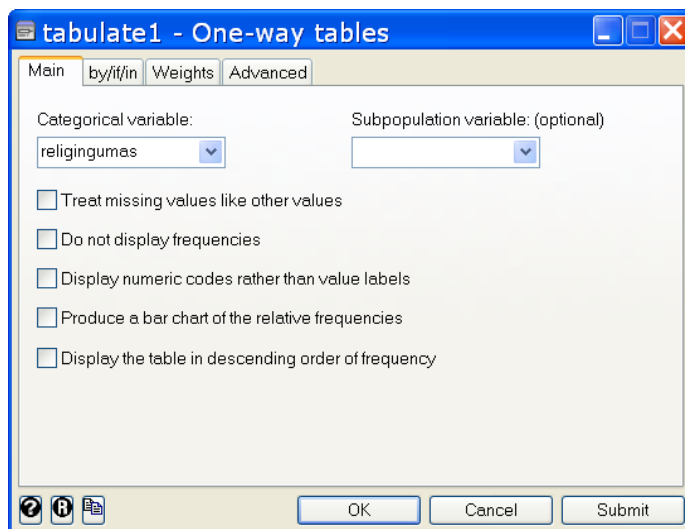
⁴ Į ESS pirmąją bangą įtrauktų šalių sąrašas pateiktas pirmoje studijų paketo dalyje (arba žr. http://www.europeansocialsurvey.org/index.php?option=com_content&task=view&id=42&Itemid=74). Taigi terminas „europiečiai“ šiuo atveju apima ne visas Europos šalis, o tik tas, kuriose tyrimas buvo vykdytas 2002 m.

⁵ Be to, taip pateikta informacija leidžia pastebėti įvairias duomenų savybes (duomenų sklaidos mastą, dažniausiai pasikartojančias reikšmes ir pan.)

Sprendimas su SPSS 15.0



Sprendimas su Stata SE 10.0



Sintaksės komandos

```
FREQUENCIES VARIABLES = religingumas
/ORDER = ANGLYSIS .
```

```
tabulate religingumas
```

Gauname tokius rezultatus **rezultatų išvesties lange**:

Savęs priskyrimas kokiai nors religijai ar denominacijai					
		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	Priskiria	27081	63,9	64,2	64,2
	Nepriskiria	15070	35,6	35,8	100,0
	Total	42151	99,5	100,0	
Missing	Atsisakė nurodyti	37	,1		
	Nežino	127	,3		
	Neatsakė	44	,1		
	Total	208	,5		
Total		42359	100,0		

Savęs priskyrimas kokiai nors religijai ar denominacijai			
	Freq.	Percent	Cum.
Priskiria	27,081	64.25	64.25
Nepriskiria	15,070	35.75	100.00
Total	42,151	100.00	

Dažnių ir santykinųjų dažnių lenteles galima **nukopijuoti ir įkelti į tekstų redagavimo programas**

Kad **nukopijuoti lentelę**, reikia ją pažymėti ir paspausti dešinį pelės klavišą. Pasirodys kortelė.

Pasirinkus „**Copy**“ lentelė bus nukopijuota **lenteliniu** formatu, o pasirinkus „**Copy objects**“ – kaip **paveikslas** (t. y. jos negalima bus redaguoti, tačiau jos formatavimas neišsikraipys).

Kad **nukopijuoti lentelę**, reikia ją pažymėti ir paspausti dešinį pelės klavišą. Pasirodys kortelė.

Pasirinkus „**Copy Text**“ lentelė bus nukopijuota **tekstiniu** formatu (t. y. ji nebus „įprasta lentelė“, o tik tam tikru būdu sutvarkytas tekstas, kad „atrodytų kaip lentelė“), pasirinkus „**Copy Table**“ – taip pat bus nukopijuota **tekstiniu** formatu tik kitaip sutvarkyta (taip sutvarkytą lentelę galima perkelti į skaičiuokles), pasirinkus „**Copy Table as HTML**“ – **lenteliniu** formatu, o pasirinkus „**Copy as Picture**“ – kaip **paveikslas** (t. y. jos negalima bus redaguoti, tačiau jos formatavimas neišsikraipys).

Sprendimas su SPSS 15.0

Savęs priskyrimas kokiai nors religijai ar denominacijai

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	Priskiria	27081	63,9	64,2	64,2
	Nepriskiria	15070	35,6	35,8	100,0
	Total	42151	99,5	100,0	
Missing	Atsisakė nurodyti	37	1		
	Nežino				
	Neatsakė				
	Total				

What's This?

Cut

Copy

Copy objects

Paste After

Sprendimas su Stata SE 10.0

Savęs priskyrimas kokiai nors religijai ar denominacijai			
ai	Freq.	Percent	Cum.
Priskiria	27,081	63.93	63.93
Nepriskiria	15,070	35.60	99.53
.a			
.b			
.c			
Total			

Copy Text Ctrl+C

Copy Table

Copy Table as HTML

Copy as Picture

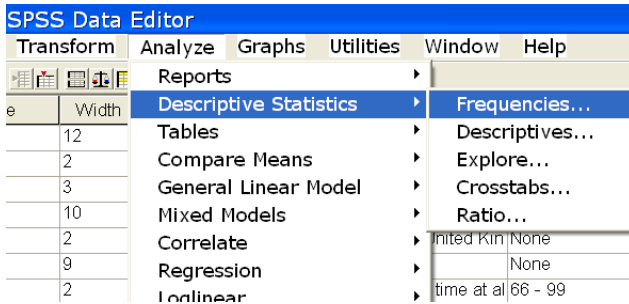
Kaip matome, rezultatai pateikiami šiek tiek skirtingai: standartinėje dažnių lentelėje SPSS paketas pateikia informaciją apie trūkstamąsias (*missing*) reikšmes, o Stata paketas – ne (tiesa, tą galima nurodyti pažymint varnele pasirinkimą „*Treat missing values like other values*“). Be to, SPSS pateikia ne tik visų (įskaitant trūkstamąsias) reikšmių santykinius dažnius visų apklaustųjų ir atsakiusiųjų (*valid*) reikšmių santykinius dažnius. Tačiau šie skirtumai nėra esminiai, reikia tik žinoti, kuri informacija reikalinga ir ją surasti rezultatų lange. Taigi matome, kad **apie du trečdalius (63,9 %⁶) europiečių priskiria save kokiai nors religijai, o 35,6 % – nepriskiria**. Norėdami sužinoti, **kokioms religijoms europiečiai save priskiria**, turime atlikti analogišką analizę, tik įkelti kitą kintamąjį. Pavyzdys, kaip tai padaryti su SPSS ir Stata, yra pateiktas 2.3 lentelėje.

2.3 lentelė. Kokybinių kintamųjų reikšmių pasiskirstymo bendros tendencijos skaitinis vaizdavimas: dažnių ir santykinų dažnių lentelės.

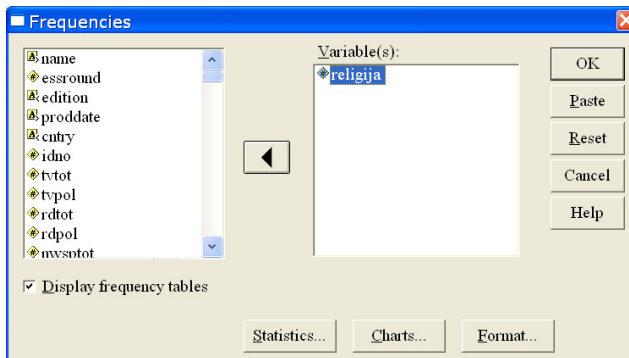
Sprendimas su SPSS 15.0	Sprendimas su Stata SE 10.0
Parengiamieji veiksmai	
Nėra	
Sintaksės komandos	
Nėra	
Analizės veiksmai	
Meniu punktai: Analyze → Descriptive Statistics → Frequencies	Meniu punktai: Statistics → Summaries, tables, and tests → Tables → One-way tables

⁶ Čia reikia atkreipti dėmesį į tai, kad SPSS standartiškai pateikia skaičius suapvalintus iki vieno skaičiaus po kablelio, o Stata – iki dviejų skaičių po kablelio.

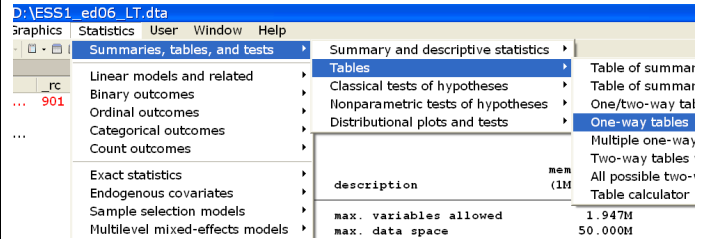
Sprendimas su SPSS 15.0



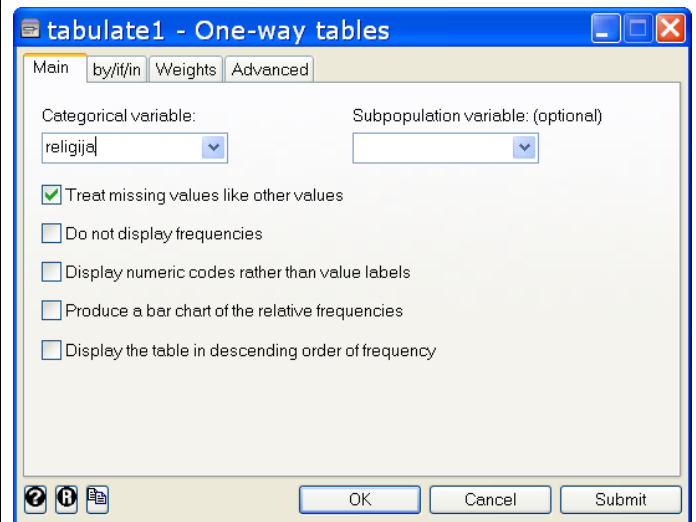
Atsidariusiame lange „*Frequencies*“ į dešinėje pusėje esantį langelį „*Variables(s)*“ įkeliam kintamąjį **religija** ir spaudžiame „*OK*“.



Sprendimas su Stata SE 10.0



Atsidariusiame lange „*tabulate1 – One-way tables*“ parenkame analizės kintamąjį „*Categorical variable*“ **religija**, nurodome įtraukti į analizę trūkstamąsias reikšmes pažymėdami „*Treat missing values like other values*“ ir spaudžiame „*OK*“.



Sintaksės komandos

`FREQUENCIES VARIABLES = religija /ORDER = ANALYSIS .`

`tabulate religija, missing`

Gauname tokius rezultatus **rezultatų išvesties** lange:

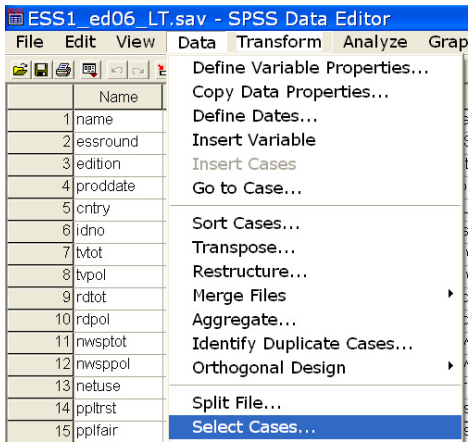
Religija ar denominacija, kuriai priskiria save šiuo metu				
		Frequency	Percent	Cumulative Percent
Valid	Katalikų	14446	34,1	53,6
	Protestantų	6727	15,9	78,5
	Rytų ortodoksų	2531	6,0	87,9
	Kita krikščionių denominacija	762	1,8	90,7
	Žydų	1365	3,2	95,8
	Islamo	841	2,0	98,9
	Rytų religijos	128	,3	99,4
	Kita nekrikščionių religija	169	,4	100,0
	Total	26969	63,7	100,0
Missing	Netaikytina	15234	36,0	
	Atsisakė nurodyti	69	,2	
	Neatsakė	87	,2	
	Total	15390	36,3	
Total		42359	100,0	

Religija ar denominacija, kuriai priskiria save šiuo metu			
	Freq.	Percent	Cum.
Katalikų	14,446	34.10	34.10
Protestantų	6,727	15.88	49.98
Rytų ortodoksų	2,531	5.98	55.96
Kita krikščionių denominacija	762	1.80	57.76
Žydų	1,365	3.22	60.98
Islamo	841	1.99	62.97
Rytų religijos	128	0.30	63.27
Kita nekrikščionių religija	169	0.40	63.67
.a	15,234	35.96	99.63
.b	69	0.16	99.79
.c	87	0.21	100.00
Total	42,359	100.00	

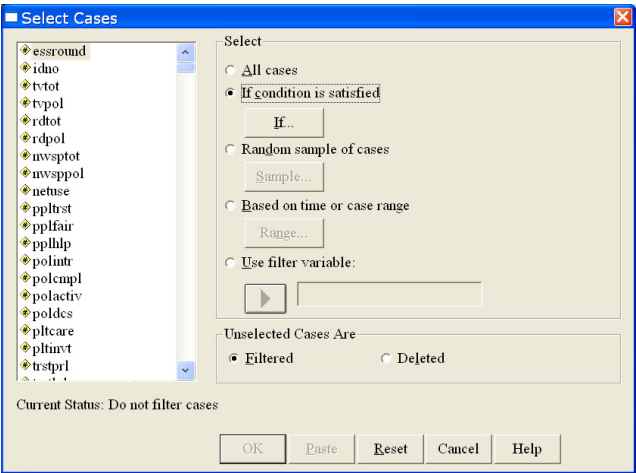
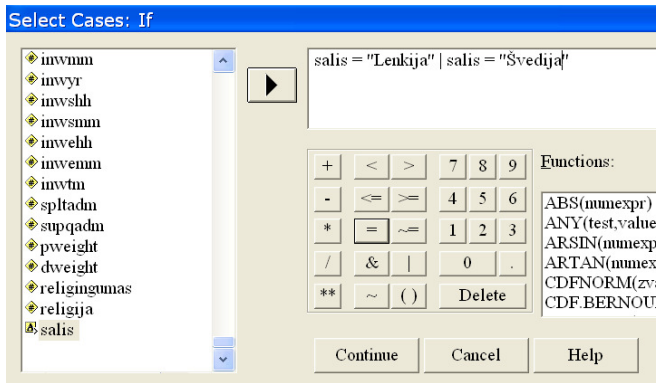
Taigi matome, kad **europiečių religinės tapatybės yra labai įvairios, o išsiskiria katalikiška (34 %) ir protestantiška (16 %) tapatybė**. Reikėtų pastebėti, kad Stata pakete trūkstamosios reikšmės neturi (ir negali turėti) pavadinimų. Taigi jos iš tikrųjų yra „trūkstamosios“. Norėdami sužinoti, kiek respondentų atsisakė nurodyti religinę tapatybę, kiek jų neatsakė į šį klausimą ir pan., turėtume tuos pasirinkimus laikyti įprastais atsakymų variantais ir koduoti kaip tinkamas reikšmes.

Atskleidus bendrąsias europiečių religingumo tendencijas, įdomu **palyginti tas tendencijas skirtingose šalyse**. Empiriniai tyrimai rodo, kad religingumas yra silpnesnis Šiaurės šalyse – Skandinavijoje, o stipresnis – Rytų Europoje. Taigi pabandytume palyginti religingumo tendencijas Švedijoje ir Lenkijoje. Tam atlikti paprastai sudaromos **kryžminės dažnių ir santykinų dažnių lentelės**⁷. Pavyzdys, kaip tai padaryti su SPSS ir Stata, yra pateiktas 2.4 lentelėje.

2.4 lentelė. Kokybinių kintamųjų reikšmių pasiskirstymo bendrų tendencijų palyginimo skaitinis vaizdavimas: kryžminės dažnių ir santykinų dažnių lentelės.

Sprendimas su SPSS 15.0	Sprendimas su Stata SE 10.0
Parengiamieji veiksmai	Parengiamieji veiksmai
Prieš sudarant kryžminę lentelę SPSS pakete reikia iš anksto nurodyti, kad bus analizuojami tik tam tikri (šiuo atveju Švedijos ir Lenkijos) duomenys. <i>Meniu punktai: Data → Select Cases</i>  Atsidariusiame lange „<i>Select Cases</i>“ pasirenkame „<i>If condition is satisfied</i>“ ir spaudžiame „<i>If</i>“.	Nėra, nes Stata pakete pasirinkti duomenų aibes, su kuriomis bus dirbama, galima atliekant kiekvieną konkrečią analizę.

⁷ Išsamiau apie kryžmines lenteles bus kalbama penktajame skyriuje. Čia tik pateikiami bendrieji kryžminių lentelių sudarymo principai, bei grafinis vaizdavimas.

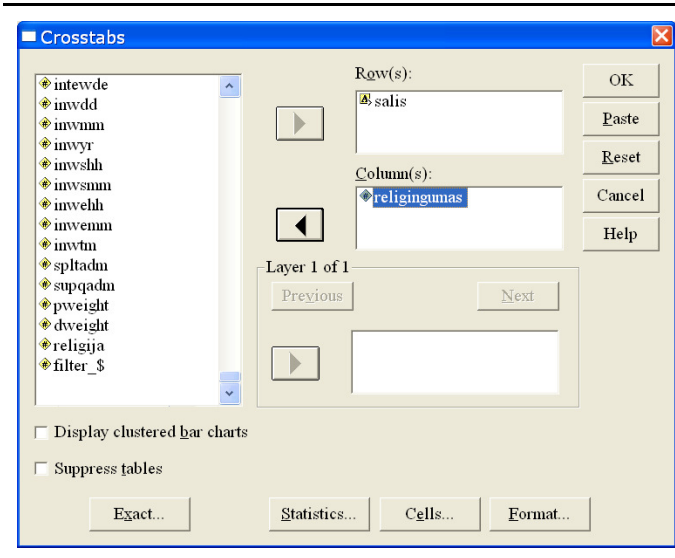
Sprendimas su SPSS 15.0	Sprendimas su Stata SE 10.0
 The image shows the 'Select Cases' dialog box in SPSS 15.0. On the left, a list of variables is shown, including 'essround', 'idno', 'tvrtot', 'tvpot', 'rdtot', 'rdpot', 'mvspot', 'mvspot', 'netuse', 'ppltrst', 'pplfair', 'pplhlp', 'polintr', 'polcmpl', 'polactiv', 'poldes', 'pltrare', 'pltrvrt', and 'trstprl'. The 'Select' section on the right has four options: 'All cases' (unselected), 'If condition is satisfied' (selected), 'Random sample of cases' (unselected), and 'Based on time or case range' (unselected). Below these are buttons for 'If...', 'Sample...', and 'Range...'. The 'Unselected Cases Are' section has two options: 'Filtered' (selected) and 'Deleted' (unselected). At the bottom, there are buttons for 'OK', 'Paste', 'Reset', 'Cancel', and 'Help'. The 'Current Status' at the bottom left says 'Do not filter cases'.	
Atsidariusiame lange „Select Cases: If“ nurodome sąlygą: <code>salis = "Lenkija" salis = "Švedija"</code> ir spaudžiame „Continue“, o sugrįžus į „Select Cases“ langą spaudžiame „OK“.	
 The image shows the 'Select Cases: If' dialog box in SPSS 15.0. On the left, a list of variables is shown, including 'invmm', 'invyr', 'invshh', 'invsmm', 'invvehh', 'invnem', 'invtm', 'spltdm', 'supqadm', 'pweight', 'dweight', 'religingumas', 'religija', and 'salis'. The 'If' text box contains the condition <code>salis = "Lenkija" salis = "Švedija"</code>. Below the text box is a button with a right-pointing arrow. To the right of the arrow is a numeric keypad with buttons for '+', '<', '>', '7', '8', '9', '-', '<=', '>=', '4', '5', '6', '*', '=', '~=', '1', '2', '3', '/', '&', ' ', '0', '.', '**', '~', '()', and 'Delete'. To the right of the keypad is a list of functions: 'ABS(numexpr)', 'ANY(test,value)', 'AR SIN(numexp', 'ARTAN(numex', 'CDFNORM(zv', and 'CDF.BERNOU'. At the bottom are buttons for 'Continue', 'Cancel', and 'Help'.	
Sintaksės komandos	
<pre>USE ALL. COMPUTE filter_\$=(salis = "Lenkija" salis = "Švedija"). VARIABLE LABEL filter_\$ 'salis = "Lenkija" salis = "Švedija" (FILTER)'. VALUE LABELS filter_\$ 0 'Not Selected' 1 'Selected'. FORMAT filter_\$ (f1.0). FILTER BY filter_\$. EXECUTE .</pre>	Nėra
Analizės veiksmiai	
Meniu punktai: Analyze → Descriptive Statistics → Crosstabs	Meniu punktai: Statistics → Summaries, tables, and tests → Tables → Two-way tables with measures of association

The screenshot shows the SPSS Data Editor interface. The 'Analyze' menu is open, and the path 'Analyze > Descriptive Statistics > Crosstabs...' is highlighted. The background data table is partially visible, showing columns for 'ESS2002/2003', 'assround', 'editon', and several numerical columns.

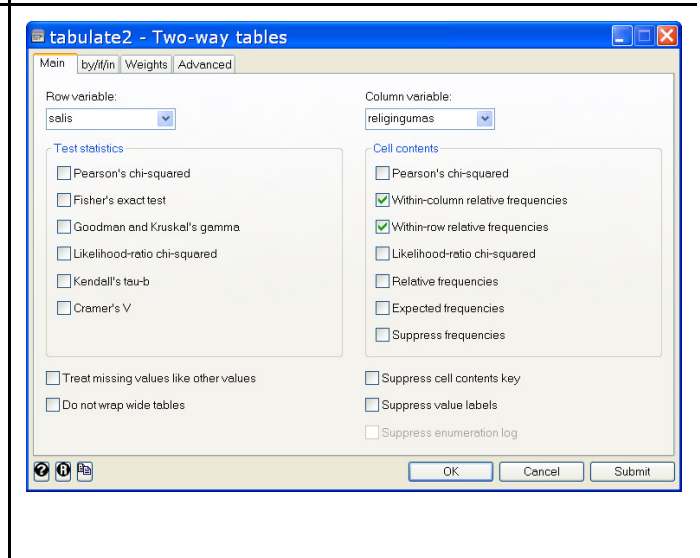
The screenshot shows the Stata SE 10.0 interface. The 'Statistics' menu is open, and the path 'Statistics > Summaries, tables, and tests > Tables' is highlighted. The command window shows the command 'salis=="I' and the command line shows 'salis=="Lenkija"|salis=="Švedija", c'.

Atsidariusiame lange „**Crosstabs**“ įkeliam analizei kintamuosius: **salis** (į eilučių langelį (**rows**)) ir **religingumas** (į stulpelių langelį (**columns**)).

Atsidariusiame lange „*tabulate2 – Two-way tables*“ parenkame analizės kintamuosius: **salis** (į eilučių langelį (*Row variable*)) ir **religingumas** (į stulpelių langelį (*Column variable*)) ir nurodome, kad būtų skaičiuojami santykiniai dažniai pagal eilutes pažymėdami „*Within-row relative frequencies*“ bei stulpelius pažymėdami „*Within-column relative frequencies*“.



Paspaudę mygtuką „**Cells...**“ atsidariusiame lange „**Crosstabs: Cell Display**“ nurodome, kad būtų skaičiuojami santykiniai dažniai pagal eilutes



Atidarę kortelę „by/if/in“ nurodome sąlygą ties „If: (expression)“: <code>salis = "Lenkija" salis</code>
--

Sprendimas su SPSS 15.0	Sprendimas su Stata SE 10.0
(<i>Percentages: row</i>) bei stulpelius (<i>Percentages: column</i>) ir spaudžiame „Continue“, o sugrįžus į „Crosstabs“ langą spaudžiame „OK“.	= "Švedija" ir spaudžiame „OK“.

Sintaksės komandos

CROSSTABS /TABLES = salis BY religingumas /FORMAT = AVALUE TABLES /CELLS = COUNT ROW COLUMN TOTAL /COUNT ROUND CELL . ⁸	tabulate salis religingumas if salis == "Lenkija" salis == "Švedija", column row ⁹
--	---

Gauname tokius rezultatus rezultatų išvesties lange:

Šalis * Savęs priskyrimas kokiai nors religijai ar denominacijai Crosstabulation

			Savęs priskyrimas kokiai nors religijai ar denominacijai		Total
			Priskiria	Nepriskiria	
Šalis	Lenkija	Count	1944	156	2100
		% within Šalis	92,6%	7,4%	100,0%
		% within Savęs priskyrimas kokiai nors religijai ar denominacijai	76,6%	10,0%	51,3%
	Švedija	Count	593	1402	1995
		% within Šalis	29,7%	70,3%	100,0%
		% within Savęs priskyrimas kokiai nors religijai ar denominacijai	23,4%	90,0%	48,7%
Total	Count	2537	1558	4095	
	% within Šalis	62,0%	38,0%	100,0%	
	% within Savęs priskyrimas kokiai nors religijai ar denominacijai	100,0%	100,0%	100,0%	

Key

frequency
row percentage
column percentage

Savęs priskyrimas kokiai nors religijai ar denominacijai			
Šalis	Priskiria	Nepriskir	Total
Lenkija	1,944	156	2,100
	92.57	7.43	100.00
	76.63	10.01	51.28
Švedija	593	1,402	1,995
	29.72	70.28	100.00
	23.37	89.99	48.72
Total	2,537	1,558	4,095
	61.95	38.05	100.00
	100.00	100.00	100.00

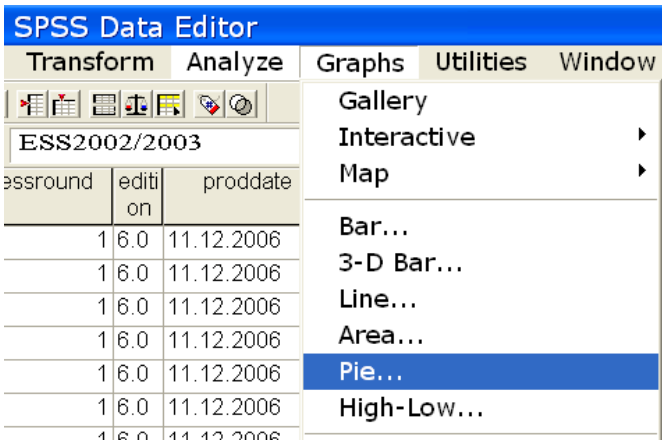
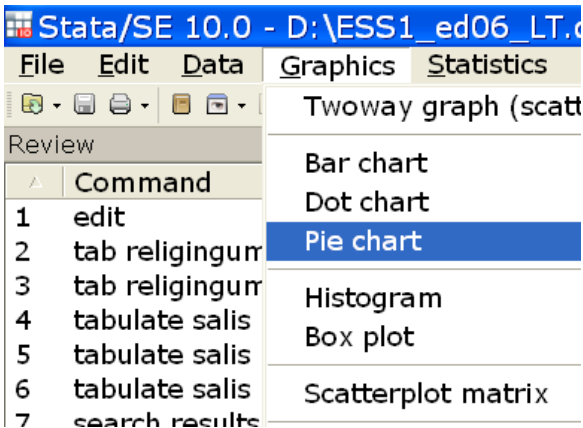
⁸ Reikia pažymėti, kad kintamasis esantis prieš BY visada tampa eilučių kintamuoju, o esantis po BY – stulpelių.

⁹ Reikia pažymėti, kad pirmas kintamasis nurodomas po komandos tabulate visada tampa eilučių kintamuoju, o antrasis – stulpelių.

Taigi matome, kad kaip ir tikėtasi, Lenkijos gyventojai yra religingesni nei Švedijos¹⁰: net 92,6 % Lenkijos gyventojų priskiria save kokiai nors religijai, kai tuo tarpu Švedijoje – tik 29,7 %.

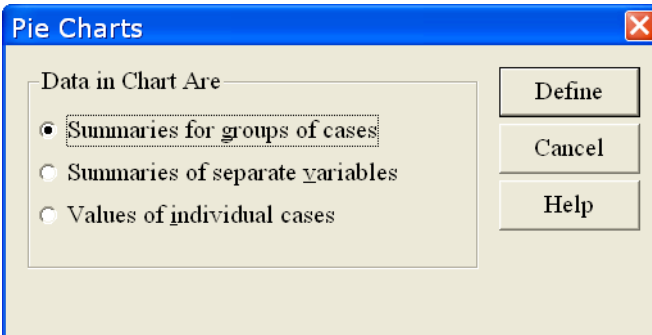
Žinoma, skaitinis vaizdavimo būdas nėra pats efektyviausias. Informatyvumo ir pagaulumo prasme jis gerokai nusileidžia **grafinio vaizdavimo** būdams. Kalbant apie statistinį **kokybinių kintamųjų** apibendrinimą ir siekiant **grafiškai pavaizduoti bendrąsias reikšmių pasiskirstymo tendencijas**, yra galimybė nubraižyti **skritulines (pie) ir stulpelines (bar) diagramas**. Iliustruodami, kaip tai padaryti, panagrinėsime **europiečių domėjimąsi politika**. Buvo klausta, kiek respondentai įvairiose Europos šalyse domisi politika (atsakymų variantai: labai domisi, kažkiek domisi, mažai domisi ir visiškai nesidomi). Kaip matome, domėjimasis politika matuojamas **ranginėje skalėje**, t. y. siekiama sužinoti domėjimosi lygį daugmaž rangine prasme: kurie respondentai domisi politika daugiau, o kurie – mažiau. Kad apibendrinti tokią informaciją ir **grafiškai pavaizduoti bendras tendencijas**, paprastai braižomos **skritulinės ir stulpelinės diagramos**. Pavyzdys, kaip tai padaryti su SPSS ir Stata, yra pateiktas 2.5 lentelėje.

2.5 lentelė. Kokybinių kintamųjų reikšmių pasiskirstymo bendrų tendencijų grafinis vaizdavimas: skritulinės ir stulpelinės diagramos.

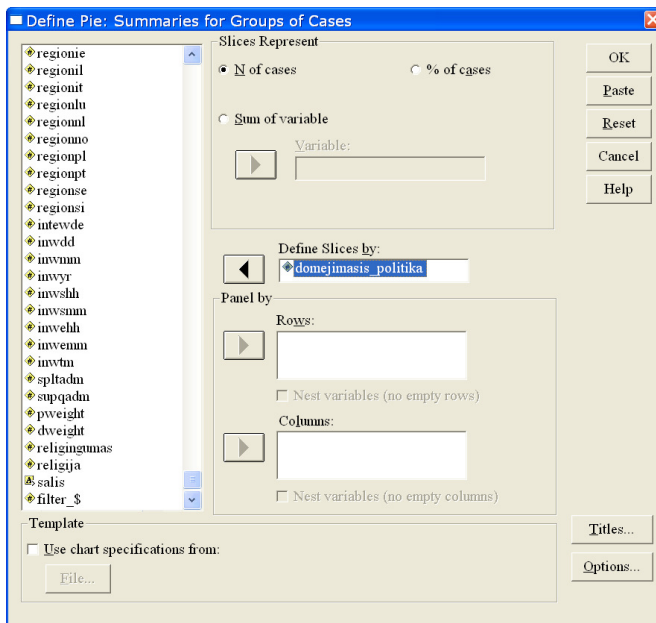
Sprendimas su SPSS 15.0	Sprendimas su Stata SE 10.0
Parengiamieji veiksmai	
Nėra	
Sintaksės komandos	
Nėra	
Analizės veiksmai: skritulinė diagrama	
<i>Meniu punktai:</i> Graphs → Pie...	<i>Meniu punktai:</i> Graphics → Pie chart
	
Atsidariusiame lange „ Pie Charts “ nurodome, kad būtų vaizduojamas atskirų kintamojo reikšmių pasiskirstymo grafikas pažymėdami „ Summaries of group of cases “ ir spaudžiame „ Define “.	Atsidariusiame lange „ graph pie – Pie charts “ parenkame analizės kintamąjį ties „ Categorical variable “ domejimasis_politika ir spaudžiame „ OK “.

¹⁰ Tiesa, statistiškai reikšmingas skirtumas ir to skirtumo dydis gali būti nustatytas tik atliekant testus bei skaičiuojant indeksus. Apie tai kalbama penktajame skyriuje. Čia pateikiamas tik preliminarus rezultatų aptarimas.

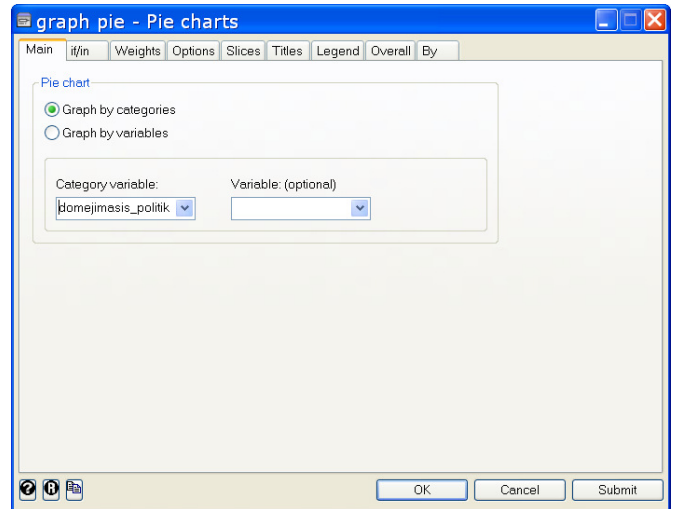
Sprendimas su SPSS 15.0



Atsidariusiame lange „*Define Pie: Summaries for Groups of Cases*“ į dešinėje esantį langelį „*Define Slices by*“ įkeliam kintamąjį **domejimasis_politika** ir spaudžiame „*OK*“.



Sprendimas su Stata SE 10.0



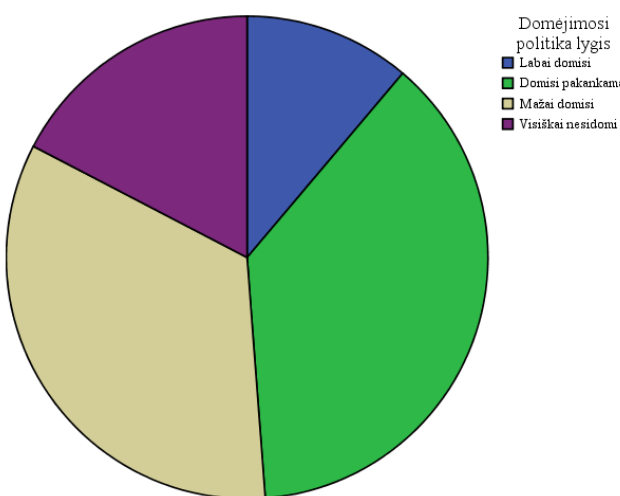
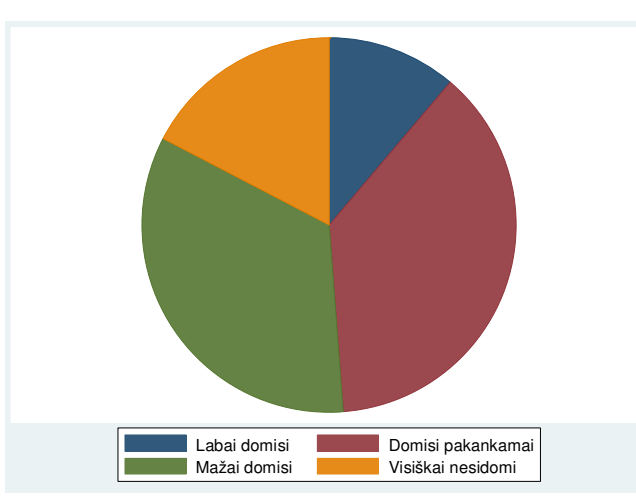
Sintaksės komandos

```
GRAPH /PIE = COUNT BY
domejimasis_politika .
```

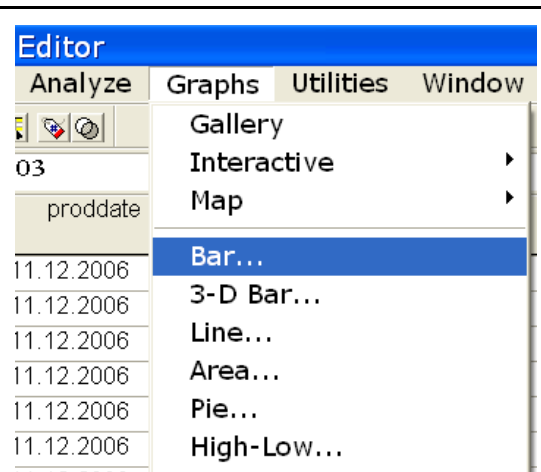
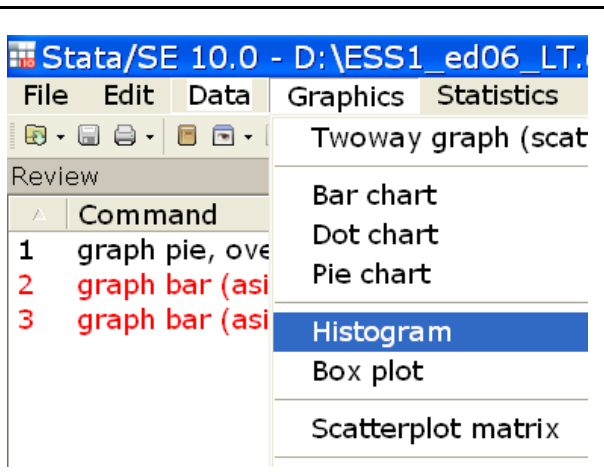
```
graph pie, over(domejimasis_politika)
```

Gauname tokius rezultatus **rezultatų išvesties lange**¹¹:

¹¹ Abiejuose paketuose gautus grafikus galima interaktyviai redaguoti, tačiau šioje medžiagoje tai nėra aptariama. Paminėsime tik tai, kad norėdami redaguoti nubraižytą grafiką, turėtumėte du kartus spustelėti kairiąją pelės klavišą ant grafiko. Be to, Stata pakete galima nurodyti grafiko braižymo nuostatas iš anksto, tam skirtos įvairios kortelės pagrindiniame grafiko braižymo lange. Bet reikia prisiminti ir tai, kad gera diagrama turi būti „ne graži, o informatyvi“. Keltą gerų patarimų grafikų braižymo klausimais duoda Vaitkevičius ir Saudargienė (2006, p. 109-120).

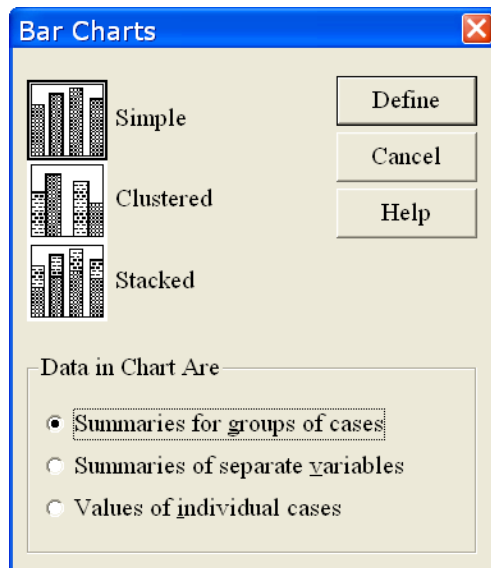
Sprendimas su SPSS 15.0	Sprendimas su Stata SE 10.0
	

Analizės veiksmas: stulpelinė diagrama

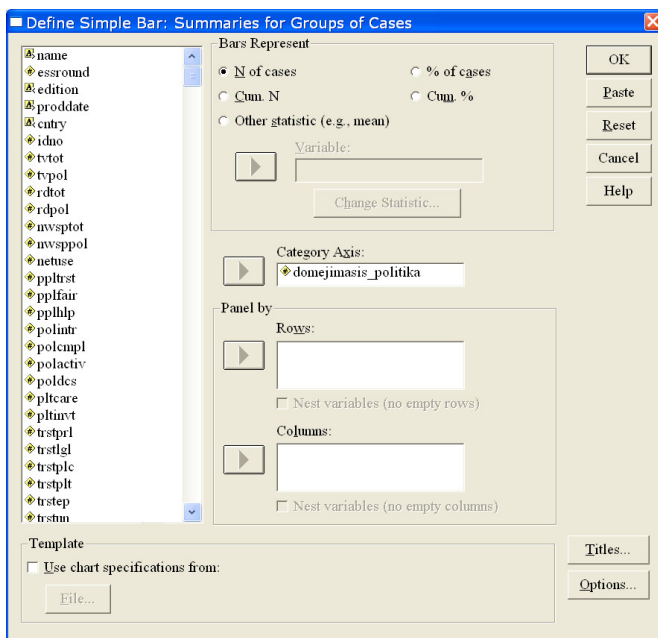
Meniu punktai: Graphs → Bar...	Meniu punktai: Graphics → Histogram ¹²
	
Atsidariusiame lange „Bar Charts“ nurodome, kad būtų vaizduojamas paprastas atskirų kintamojo reikšmių pasiskirstymo grafikas pažymėdami „Simple“ ir „Summaries of group of cases“, o po to spaudžiame „Define“.	Atsidariusiame lange „histogram – Histograms for continuous and categorical variables“ nurodome, kad būtų vaizduojamas kokybinio kintamojo reikšmių pasiskirstymo grafikas pažymėdami „Data are discrete“ ir parenkame analizės kintamąjį ties „Variable“ domejimasis politika. Po to, nurodome, kad grafiko stulpeliai vaizduotų reikšmių dažnius pažymėdami „Frequency“ ties „Y axis“ ir spaudžiame „OK“.

¹² Atkreipiame dėmesį į tai, kas Stata pakete kategoriniams (kokybiniams) kintamiesiems sudaromos histogramos yra prilygintos stulpelinėms diagramoms.

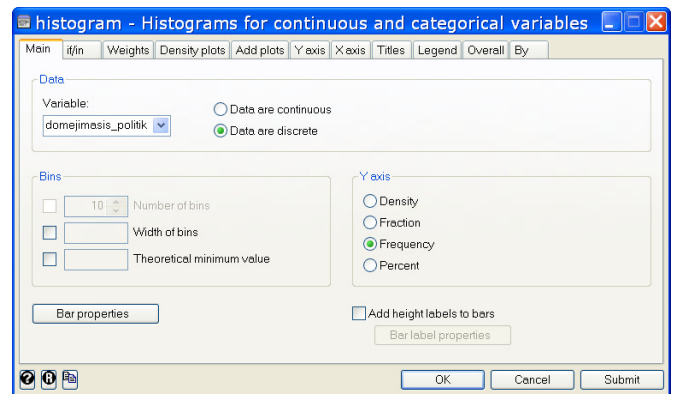
Sprendimas su SPSS 15.0



Atsidariusiame lange „*Define Simple Bar: Summaries for Groups of Cases*“ į dešinėje esantį langelį „*Category Axis*“ įkeliam kintamąjį **domejimasis_politika** ir spaudžiame „*OK*“ (yra galimybė nurodyti, kad grafiko stulpeliai vaizduotų ne atskirų reikšmių dažnius, o santykinius dažnius: ties „*Bars Represent*“ galima pasirinkti „*N of Cases*“ arba „*% of Cases*“).



Sprendimas su Stata SE 10.0



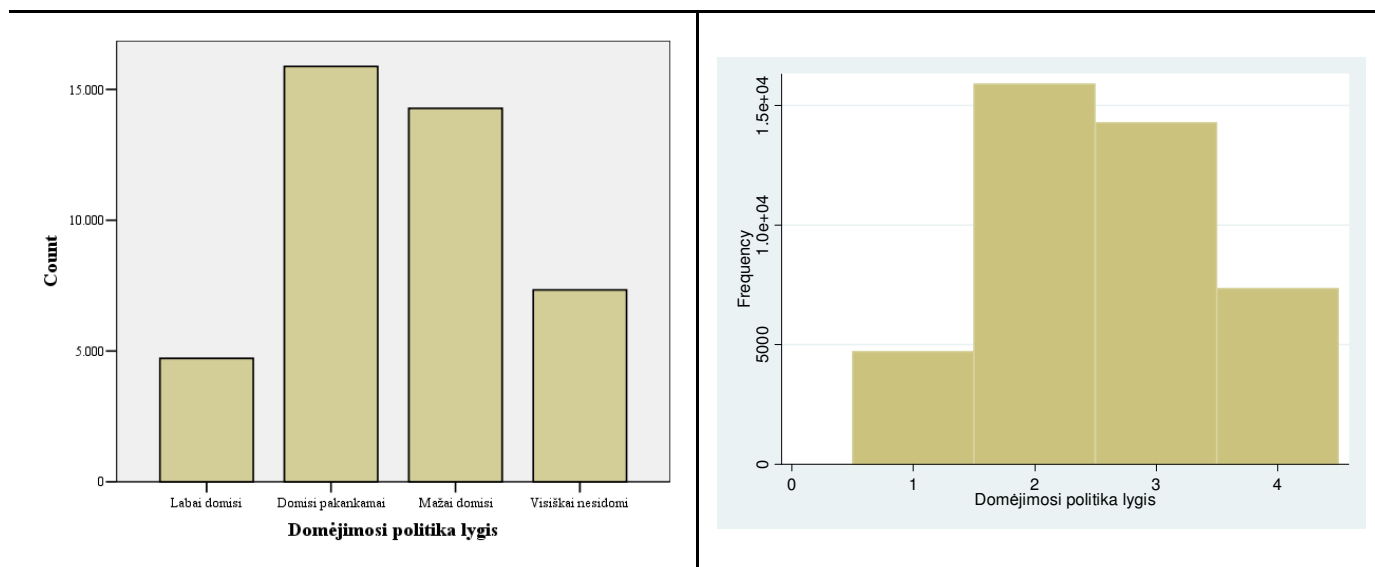
Sintaksės komandos

```
GRAPH /PIE = COUNT BY
domejimasis_politika .
```

```
histogram domejimasis_politika, discrete
frequency
```


Sprendimas su SPSS 15.0	Sprendimas su Stata SE 10.0
-------------------------	-----------------------------

Gauname tokius rezultatus **rezultatų išvesties lange**¹³:



Taigi matome, kad europiečių domėjimasis politika yra gana įvairus: apie 10 % jų arba labai domisi, arba visiškai nesidomi, o dar po 40 % – arba mažai domisi, arba domisi santykinai nemažai. Atskleidus bendrąsias europiečių domėjimosi politika tendencijas, įdomu palyginti tas tendencijas skirtingose šalyse. Empiriniai tyrimai rodo, kad apolitiškumas yra stipresnis Rytų Europos šalyse, o silpnesnis – Vakarų Europoje. Taigi pabandysime palyginti domėjimosi politika tendencijas Vokietijoje ir Lenkijoje. Tam atlikti paprastai braižomos kryžminės stulpelinės diagramos¹⁴. Pavyzdys, kaip tai padaryti mūsų nagrinėjamu atveju su SPSS ir Stata, yra pateiktas 2.6 lentelėje.

2.6 lentelė. Kokybinių kintamųjų reikšmių pasiskirstymo bendrų tendencijų palyginimo grafinis vaizdavimas: kryžminės stulpelinės diagramos.

Sprendimas su SPSS 15.0	Sprendimas su Stata SE 10.0
-------------------------	-----------------------------

Parengiamieji veiksmai

Prieš sudarant kryžminę diagramą SPSS pakete reikia iš anksto nurodyti, kad bus analizuojami tik tam tikri (šiuo atveju Vokietijos ir Lenkijos) duomenys. Kaip tą padaryti, jau buvo nurodyta 2.4 lentelėje, todėl čia nekartosime.	Nėra, nes Stata pakete pasirinkti duomenų aibės, su kuriomis bus dirbama, galima atliekant kiekvieną konkrečią analizę.
--	---

Sintaksės komandos

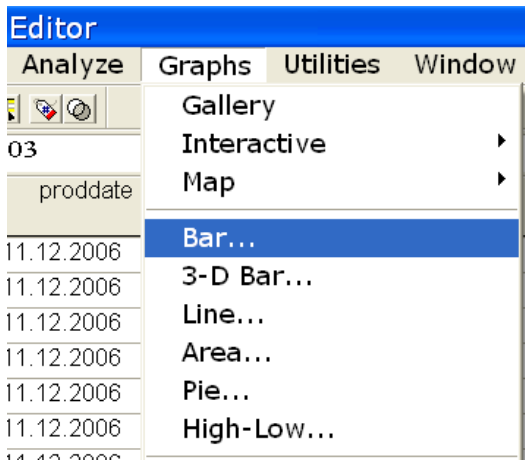
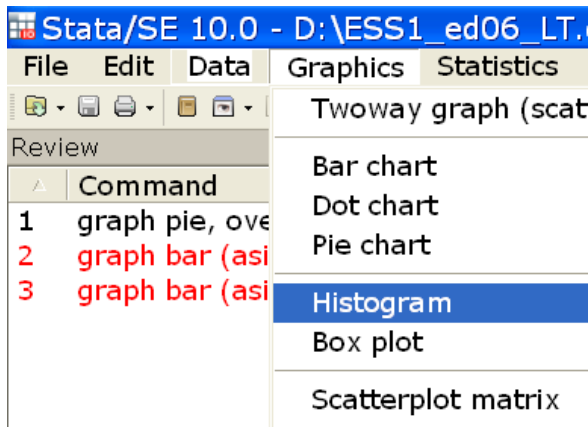
<pre>USE ALL. COMPUTE filter_\$=(salis = "Vokietija" salis = "Lenkija"). VARIABLE LABEL filter_\$ 'salis = "Vokietija" salis = "Lenkija" (FILTER)'. VALUE LABELS filter_\$ 0 'Not Selected' 1</pre>	Nėra.
---	--------------

¹³ Priminsime, kad abiejuose paketuose gautus grafikus galima interaktyviai redaguoti, tačiau šioje medžiagoje tai nėra aptariama. Be to, Stata pakete galima nurodyti grafiko braižymo nuostatas iš anksto, tam skirtos įvairios kortelės pagrindiniame grafiko braižymo lange.

¹⁴ Išsamiau apie požymių priklausomumo analizę bus kalbama penktajame skyriuje. Čia tik pateikiami bendrieji kryžminių grafikų braižymo principai.

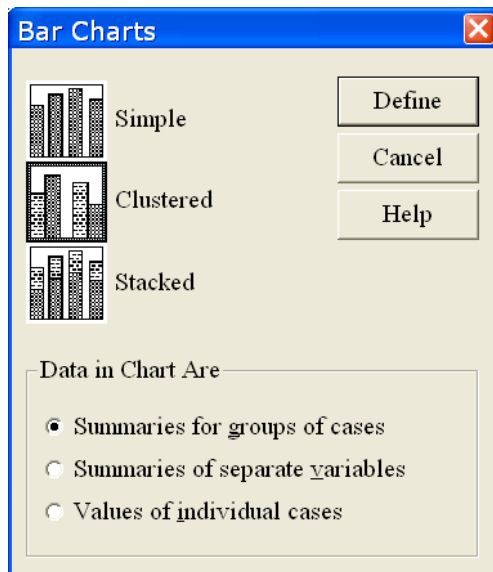
Sprendimas su SPSS 15.0	Sprendimas su Stata SE 10.0
'Selected'. FORMAT filter_\$ (f1.0). FILTER BY filter_\$. EXECUTE .	

Analizės veiksmas

Meniu punktai: Graphs → Bar...	Meniu punktai: Graphics → Histogram ¹⁵
	
Atsidariusiame lange „<i>Bar Charts</i>“ nurodome, kad būtų vaizduojamas bendras skirtingų kintamųjų atskirų reikšmių pasiskirstymo grafikas pažymėdami „<i>Clustered</i>“ ir „<i>Summaries of group of cases</i>“, o po to spaudžiame „<i>Define</i>“.	Atsidariusiame lange „<i>histogram – Histograms for continuous and categorical variables</i>“ nurodome, kad būtų vaizduojamas kokybinio kintamojo reikšmių pasiskirstymo grafikas pažymėdami „<i>Data are discrete</i>“ ir parenkame analizės kintamąjį ties „<i>Variable</i>“ domejimasis politika. Po to, nurodome, kad grafiko stulpeliai vaizduotų reikšmių santykinius dažnius (šiuo atveju, tai būtina sąlyga, nes daromas palyginimas) pažymėdami „<i>Percent</i>“ ties „<i>Y axis</i>“.

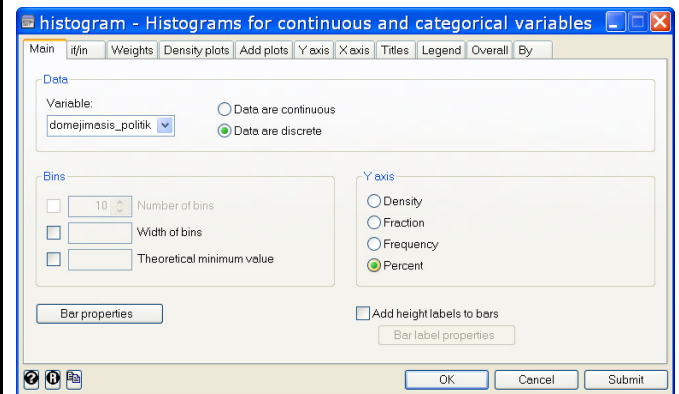
¹⁵ Vėlgį, atkreipiame dėmesį į tai, kas Stata pakete kategoriniams (kokybiniams) kintamiesiems sudaromos histogramos yra prilygintos stulpelinėms diagramoms.

Sprendimas su SPSS 15.0

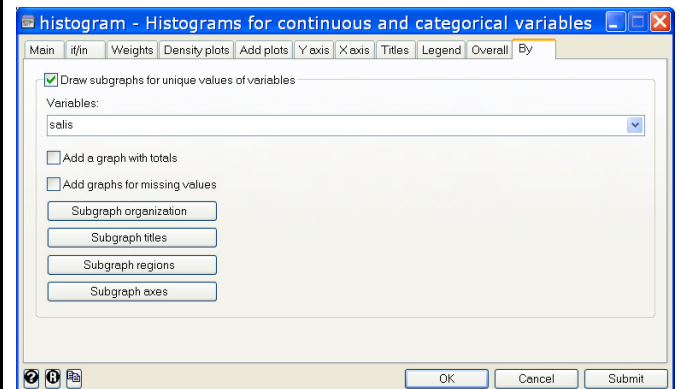
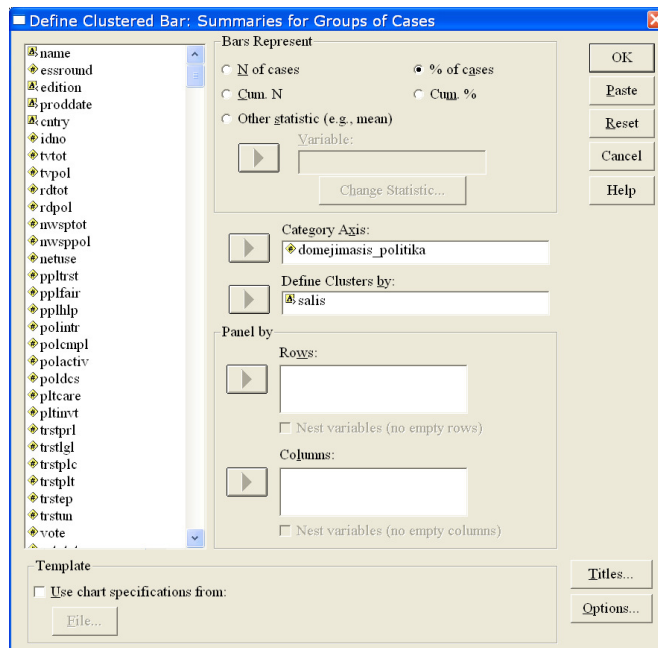


Atsidariusiame lange „**Define Clustered Bar: Summaries for Groups of Cases**“ į dešinėje esantį langelį „**Category Axis**“ įkeliam kintamąjį **domejimas_politika**, o į langelį „**Define Clusters by**“ – **salis**. Po to, ties „**Bars Represent**“ pažymime, kad grafiko stulpeliai vaizduotų ne atskirų reikšmių dažnius, o santykinius dažnius „**% of Cases**“ (šiuo atveju, tai būtina sąlyga, nes daromas palyginimas) ir spaudžiame „**OK**“.

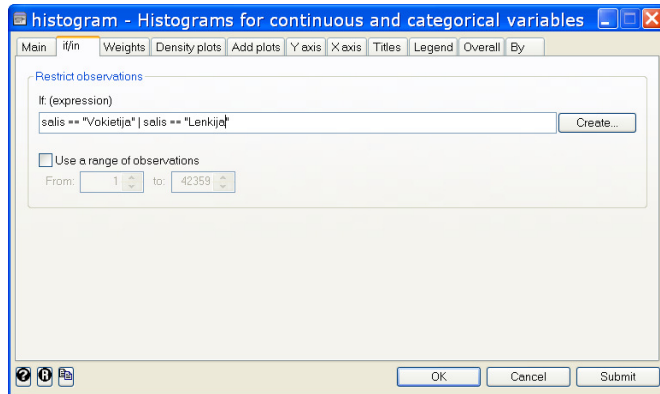
Sprendimas su Stata SE 10.0



Atidarę kortelę „**By**“ nurodome duomenų grupavimo kintamąjį **salis** ties „**Draw subgraphs for unique values of variables**“.



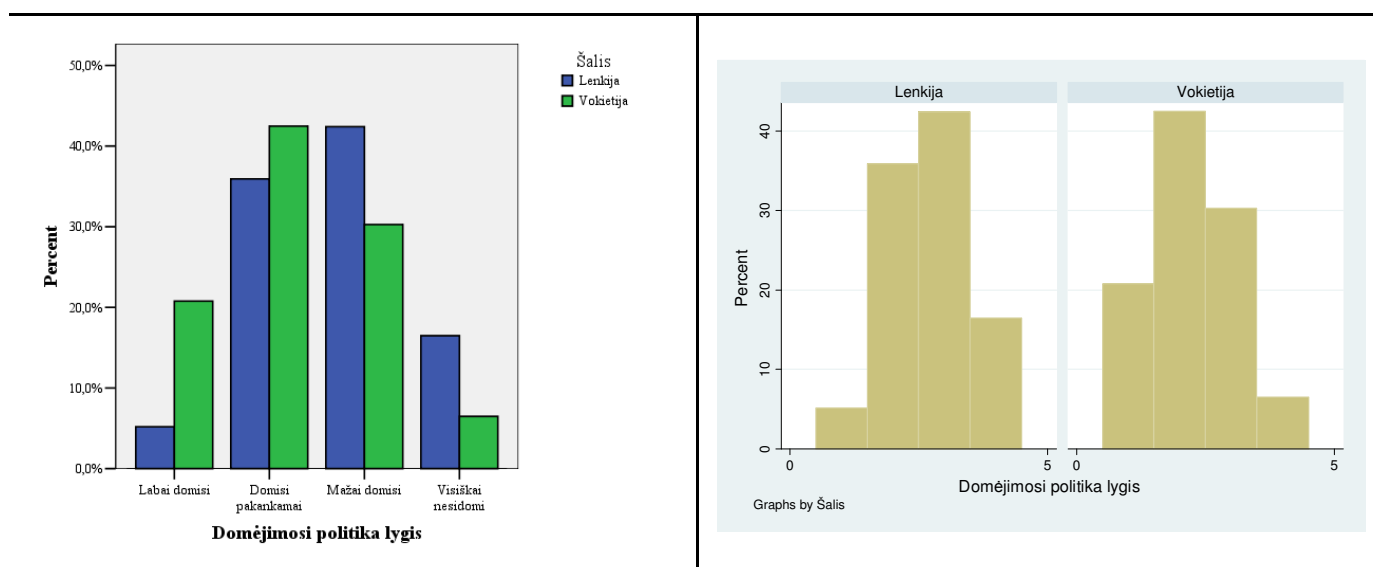
Atidarę kortelę „**if/in**“ nurodome sąlygą ties „**If: (expression)**“: **salis = "Vokietija" | salis = "Lenkija"** ir spaudžiame „**OK**“.

Sprendimas su SPSS 15.0	Sprendimas su Stata SE 10.0
	

Sintaksės komandos

<pre>GRAPH /BAR(GROUPED) = PCT BY domejimsis_politika BY salis .</pre>	<pre>histogram domejimsis_politika if salis == "Vokietija" salis == "Lenkija", discrete percent by(salis)</pre>
--	---

Gauname tokius rezultatus **rezultatų išvesties lange**¹⁶:



Taigi matome, kad kaip ir tikėtasi, daugiau Vokietijos gyventojų nurodė labai arba pakankamai besidomintys politika nei Lenkijos¹⁷. Pavyzdžiui, **net 20,8 % Vokietijos gyventojų labai domisi politika, kai tuo tarpu Lenkijoje – tik 5,2 %**.

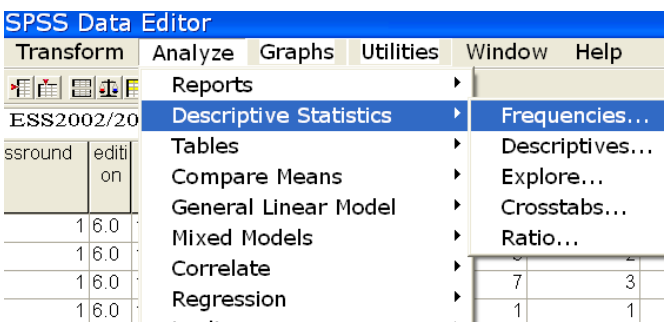
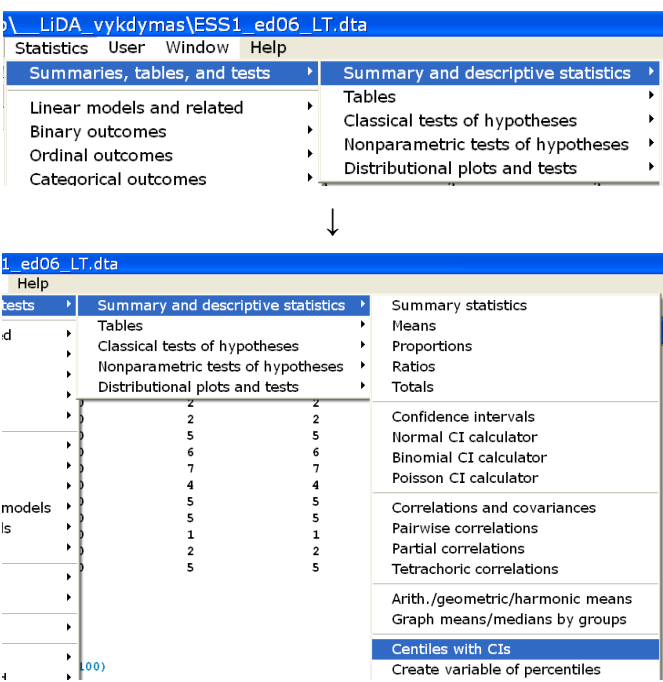
Toliau aptarsime **kiekybinių kintamųjų** skaitinį statistinį apibendrinimą. Tam iliustruoti panagrinėsime **europiečių amžių**. Buvo prašyta, kad respondentai įvairiose Europos šalyse nurodytų savo gimimo metus. Taigi amžius buvo išmatuotas **santykių skalėje**. Kad skaitiniu būdu apibendrinti tokią informaciją ir **pavaizduoti bendrąsias reikšmių pasiskirstymo tendencijas**, gali būti sudaromos **intervalinių dažnių ir santykinų intervalinių dažnių lentelės**. Tiesa, sudaryti intervalinių dažnių lenteles SPSS ir Stata programiniuose paketuose nėra galimybių. Šiose

¹⁶ Dar kartą primename, kad abiejuose paketuose gautus grafikus galima interaktyviai redaguoti, tačiau šioje medžiagoje tai nėra aptariama. to, Stata pakete galima nurodyti grafiko braižymo nuostatas iš anksto, tam skirtos įvairios kortelės pagrindiniame grafiko braižymo lange.

¹⁷ Tiesa, statistiškai reikšmingas skirtumas ir to skirtumo dydis gali būti nustatytas tik atliekant įvairius testus bei skaičiuojant indeksus. Apie tai kalbama penktajame skyriuje. Čia pateikiamas tik preliminarus rezultatų aptarimas.

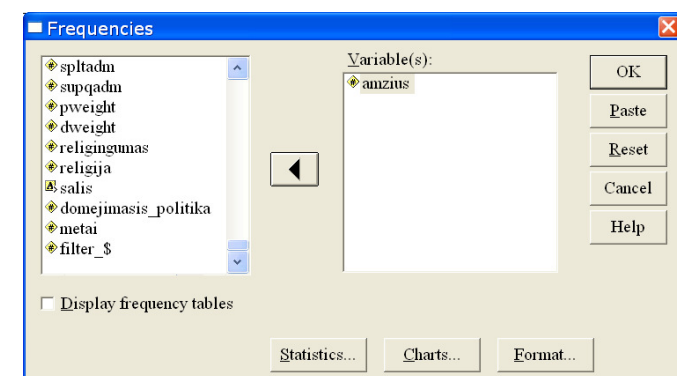
programose galima sudaryti tik santykinų intervalinių dažnių lenteles – **kvantilių („procentilių“)**¹⁸ **lenteles**. Be kvantilių, dažnai dar yra skaičiuojami ir **kvartiliai** (iš esmės jie yra 0,25, 0,5 ir 0,75 eilės kvartiliai), kurie suskirsto visas kiekybinio kintamojo reikšmes į 4 lygias dalis. Taigi sudarysime intervalinių santykinų dažnių lentelę 0,1, 0,2, 0,3 0,4, 0,5, 0,6, 0,7, 0,8, 0,9, 1 eilės kvartiliams (suskirstysime visas amžiaus reikšmes į 10 lygių intervalų) ir kvartiliams (suskirstysime visas amžiaus reikšmes į 4 lygius intervalus). Pavyzdys, kaip tai padaryti su SPSS ir Stata, yra pateiktas 2.7 lentelėje.

2.7 lentelė. Kiekybinių kintamųjų reikšmių pasiskirstymo bendrų tendencijų skaitinis vaizdavimas: intervalinių santykinų dažnių lenteles.

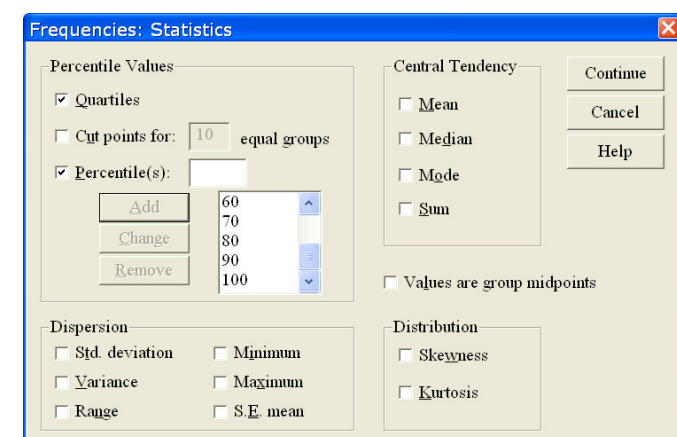
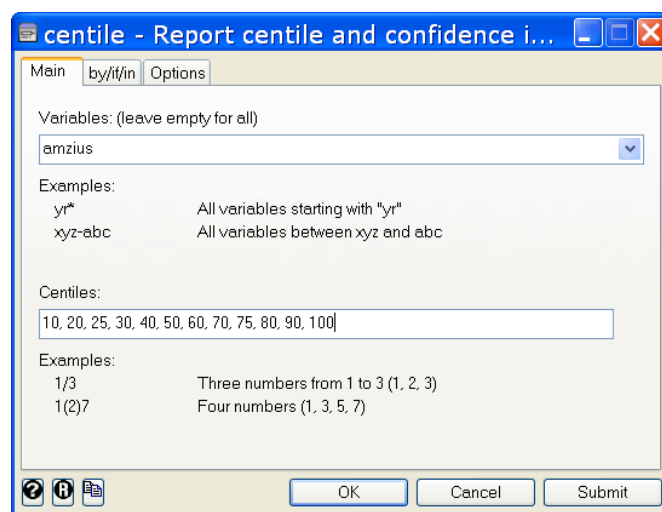
Sprendimas su SPSS 15.0	Sprendimas su Stata SE 10.0
Parengiamieji veiksmai	
Nėra	
Sintaksės komandos	
Nėra	
Analizės veiksmai	
Meniu punktai: Descriptive Statistics → Frequencies... 	Meniu punktai: Statistics → Summaries, tables, and tests → Summary and descriptive statistics → Centiles with CIs 
Atsidariusiame lange „Frequencies“ į dešinėje esantį langelį „Variable(s)“ įkeliam kintamąjį	Atsidariusiame lange „centile – Report centile and confidence intervals“ parenkame analizės

¹⁸ Kadangi procentilis yra 0,01 eilės kvantilis, tai jis šioje medžiagoje nėra atskirai nagrinėjamas. Be to, reiktų atkreipti dėmesį į tai, kad ir SPSS, ir Stata paketuose kvartiliai įvardijami „procentiliais“. Taip yra greičiausiai todėl, kad kvartiliai pateikiami išreikšti procentais.

Sprendimas su SPSS 15.0	Sprendimas su Stata SE 10.0
amzius, ir spaudžiame mygtuką „Statistics...“.	kintamąjį ties „Variables: ...“ amzius, o ties „Centiles“ įvedame 10, 20, 25 ¹⁹ , 30, 40, 50 ²⁰ , 60, 70, 75 ²¹ , 80, 90, 100 ir spaudžiame „OK“.



Atsidariusiame lange „Frequencies: Statistics“ nurodome, kad būtų skaičiuojami kvartilai pažymėdami varnelę ties „Quartiles“ ir įvedame reikiamas kvantilių (išreikštų procentais) reikšmes 10, 20, 30, 40, 50, 60, 70, 80, 90 ir 100 ties „Percentile(s)“ (įvesdami skaičių ir spausdami „Add“). Tada spaudžiame „Continue“, o sugrįžus į „Frequencies“ langą spaudžiame „OK“.



Sintaksės komandos

<pre>FREQUENCIES VARIABLES = amzius /NTILES = 4 /PERCENTILES = 10 20 30 40 50 60 70 80 90 100 /ORDER = ANALYSIS .</pre>	<pre>centile amzius, centile(10, 20, 25, 30, 40, 50, 60, 70, 75, 80, 90, 100)</pre>
---	---

Gauname tokius rezultatus **rezultatų išvesties** lange:

¹⁹ Pirmasis kvartilis.

²⁰ Antrasis kvartilis.

²¹ Trečiasis kvartilis.

Sprendimas su SPSS 15.0			Sprendimas su Stata SE 10.0					
Statistics								
Amžius								
N	Valid	42101	Variable	Obs	Percentile	Centile	— Binom. Interp. —	
	Missing	258					[95% Conf. Interval]	
Percentiles	10	22,00	amzius	42101	10	22	21	22
	20	28,00			20	28	28	29
	25	31,00			25	31	31	32
	30	34,00			30	34	34	35
	40	39,00			40	39	39	40
	50	45,00			50	45	45	45
	60	51,00			60	51	50	51
	70	57,00			70	57	56	57
	75	60,00			75	60	60	60
	80	64,00			80	64	63	64
	90	72,00			90	72	72	72
	100	109,00			100	109	109	109*
			* Lower (upper) confidence limit held at minimum (maximum) of sample					

Taigi matome, kad pavyzdžiui, amžiaus **intervale nuo 28 iki 34 metų** yra dešimt procentų europiečių, o jauniausio europiečių **ketvirtadalio** amžius yra nuo 13²² iki 31 metų.

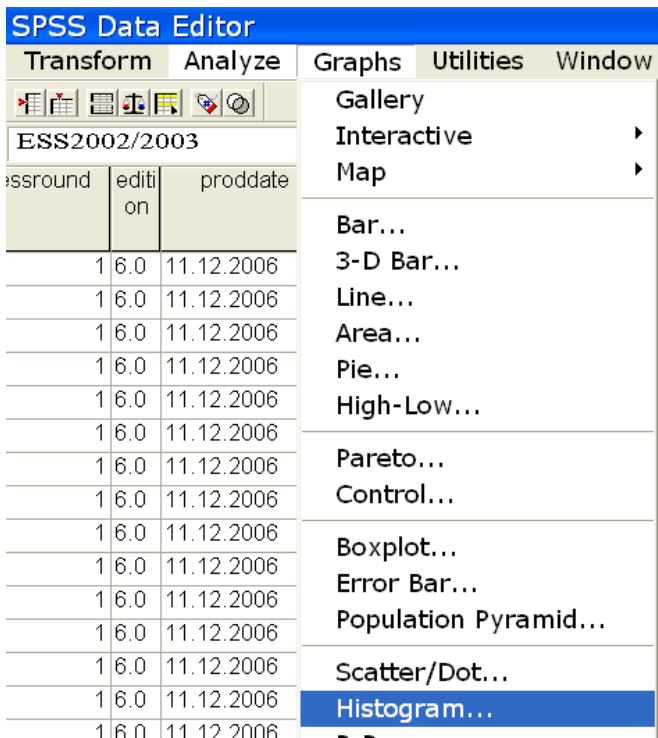
Kaip jau minėta, skaitinis vaizdavimo būdas nėra pats efektyviausias, nes informatyvumo ir pagaulumo prasme jis gerokai nusileidžia **grafinio vaizdavimo** būdams. Kalbant apie statistinį **kiekybinių kintamųjų** apibendrinimą ir siekiant **grafiškai pavaizduoti bendrąsias reikšmių pasiskirstymo tendencijas**, paprastai braižomos **histogramos (histogram)** ir **stačiakampės diagramos (boxplot)**. Histogramoje pateikiama informacija apie intervalinius dažnius arba intervalinius santykinus dažnius, o stačiakampėje diagramoje pateikiami kvartiliai ir kita apibendrinta informacija. Iliustracijai, kaip nubraižyti histogramą ir stačiakampę diagramą, vėlgi panaudosime **europiečių amžiaus kintamąjį**. Pavyzdys, kaip tai padaryti su SPSS ir Stata, pateiktas 2.8 lentelėje.

2.8 lentelė. Kiekybinių kintamųjų reikšmių pasiskirstymo bendrų tendencijų grafinis vaizdavimas: histogramos ir stačiakampės diagramos.

Sprendimas su SPSS 15.0	Sprendimas su Stata SE 10.0
Parengiamieji veiksmai	
Nėra	
Sintaksės komandos	
Nėra	
Analizės veiksmai: histograma	
Meniu punktai: Graphs → Histogram...	Meniu punktai: Graphics → Histogram

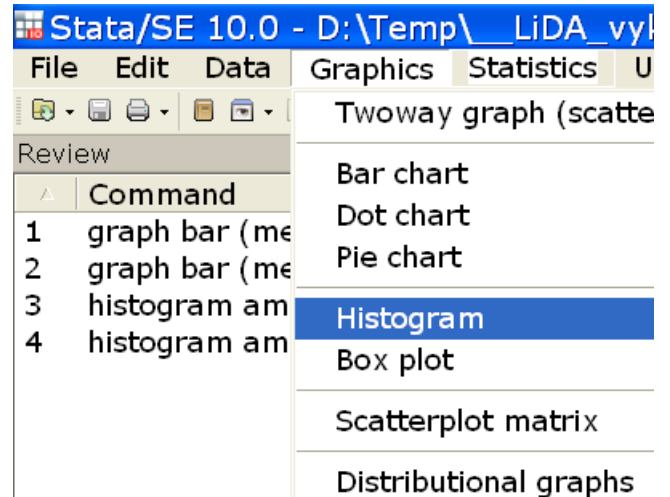
²² Toks yra jauniausio apklausto europiečio amžius. Kaip šį rodiklį suskaičiuoti, bus kalbama poskyryje 2.2., pateikiant pavyzdžius apie mažiausios ir didžiausios reikšmės skaičiavimą.

Sprendimas su SPSS 15.0

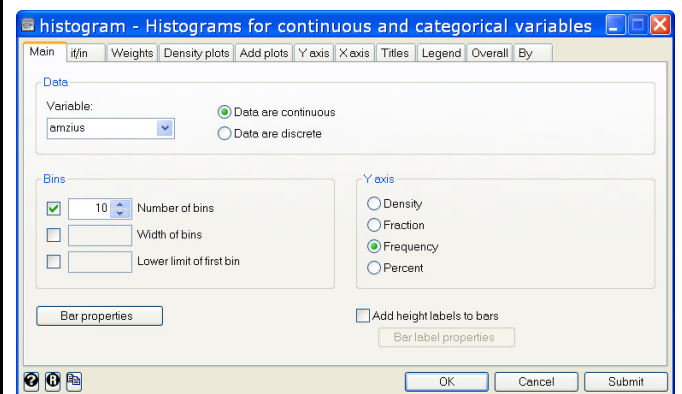
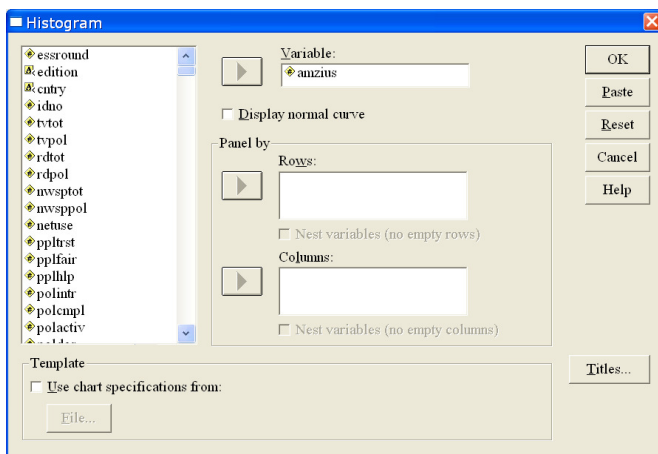


Atsidariusiame lange „**Histogram**“ į dešinėje esantį langelį „**Variable**“ įkeliam kintamąjį **amzius**, ir spaudžiame „**OK**“.

Sprendimas su Stata SE 10.0



Atsidariusiame lange „**histogram – Histograms for continuous and categorical variables**“ parenkame analizės kintamąjį **amzius** ties „**Variable: ...**“, o ties „**Y axis**“ pažymime „**Frequency**“, kad histogramoje būtų vaizduojami intervaliniai dažniai, ir spaudžiame „**OK**“.



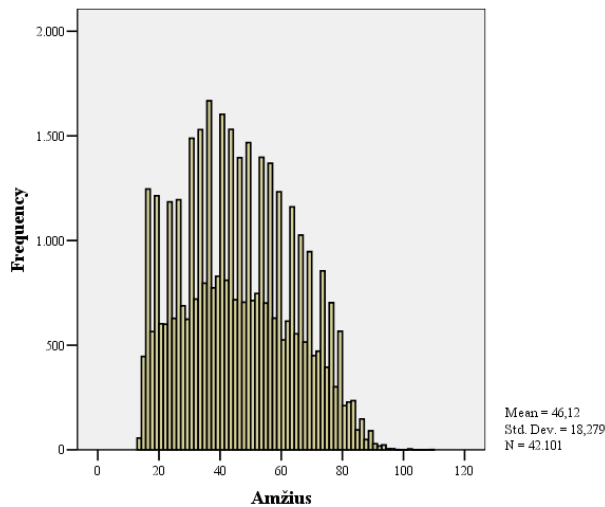
Sintaksės komandos: histograma

GRAPH /HISTOGRAM = amzius .

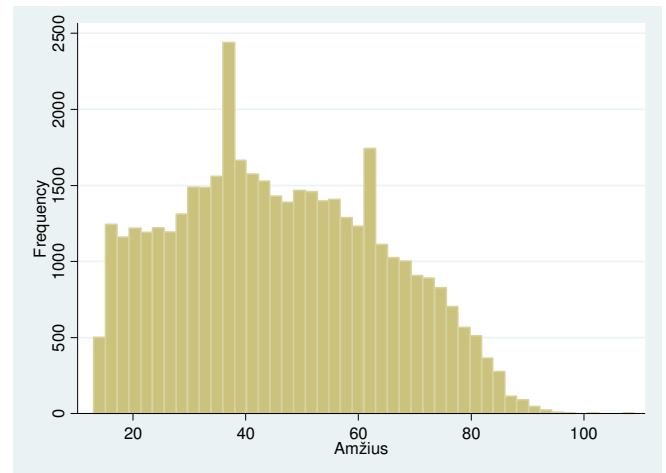
histogram amzius, frequency

Gauname tokius rezultatus **rezultatų išvesties** lange:

Sprendimas su SPSS 15.0



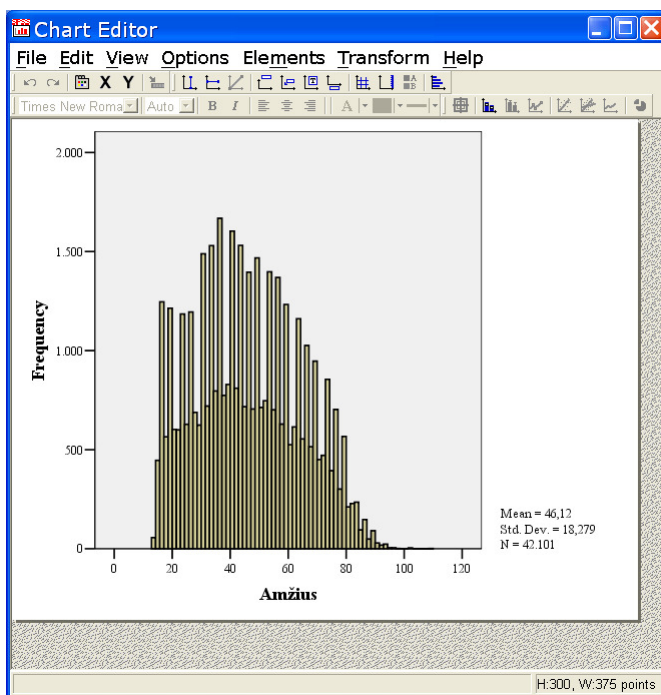
Sprendimas su Stata SE 10.0



Kaip matome, abiem atvejais histogramų intervalų kiekis yra gana didelis, todėl vis dar gana sunku suvokti informaciją. Tokiu atveju yra galimybė sumažinti **intervalų skaičių, t. y. nustatyti intervalų skaičių „rankiniu“ būdu**, o ne automatiškai.

Paspaudus du kartus kairiąją pelės klavišą ant grafiko rezultatų lange, atsidaro grafikų redagavimo langas „**Chart Editor**“.

Viską darome taip pat, tik lange „**histogram – Histograms for continuous and categorical variables**“ ties intervalų parinktimis „**Bins**“ nurodome intervalų skaičių 10 ir spaudžiame „**OK**“. (Kadangi nurodomas intervalų kiekis, pasikeičia ir sintaksės komanda: `histogram amžius, bin(10) frequency`).

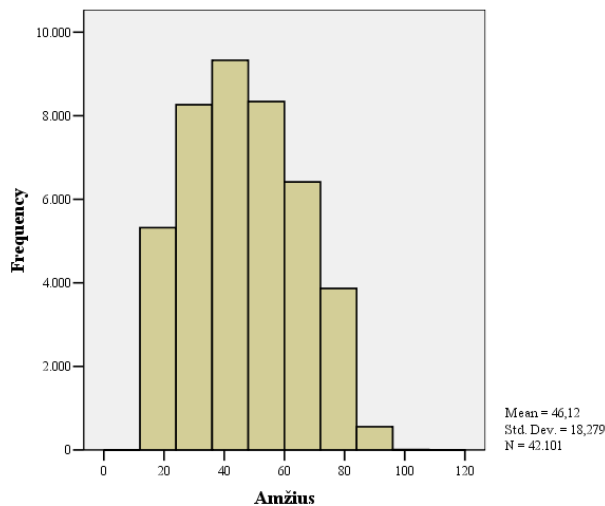


Paspaudus du kartus kairiąją pelės klavišą ant grafiko **ties x ašimi**, atsidaro grafiko parinkčių

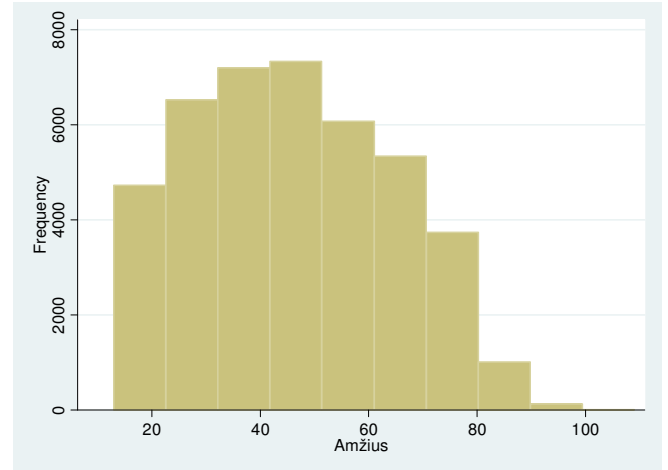
Gauname histogramą su dešimčia intervalų.

Sprendimas su SPSS 15.0

nustatymo „*Properties*“ langas. Šio lango kortelėje „*Histogram Options*“, ties intervalų sudarymo parinktimis „*Bin Sizes*“ nurodome intervalų skaičių: 10, spaudžiame „*Apply*“ ir „*Close*“. Po to, įprastu būdu uždarome grafikų redagavimo langą ir rezultatų lange gauname histogramą su dešimčia intervalų.

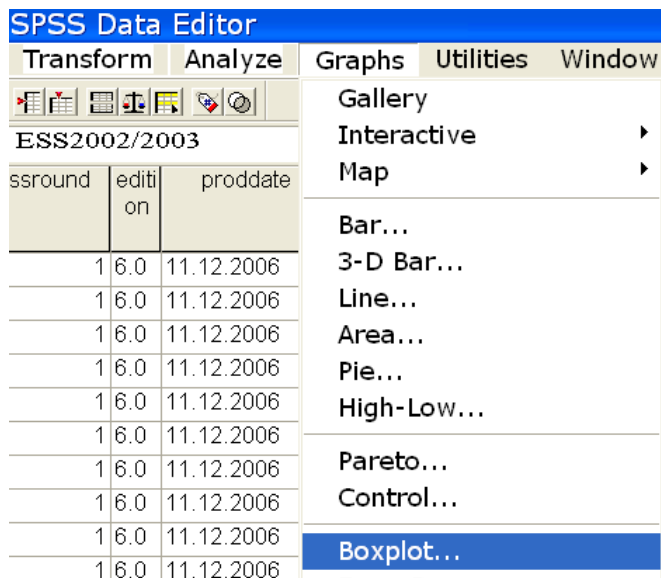


Sprendimas su Stata SE 10.0



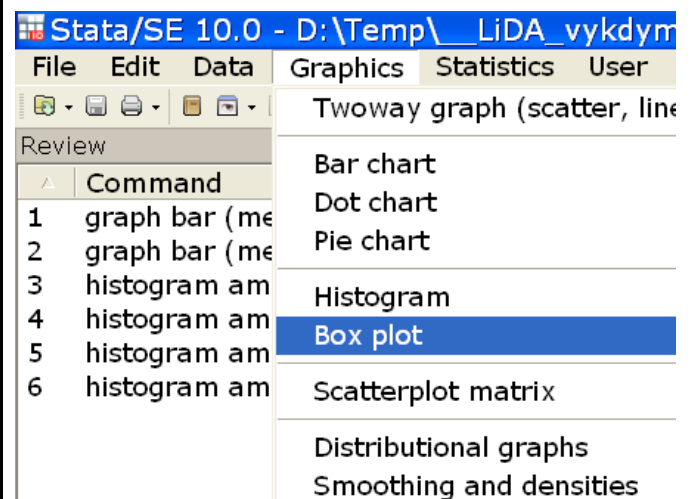
Analizės veiksmas: stačiakampė diagrama

Meniu punktai: Graphs → Boxplot...



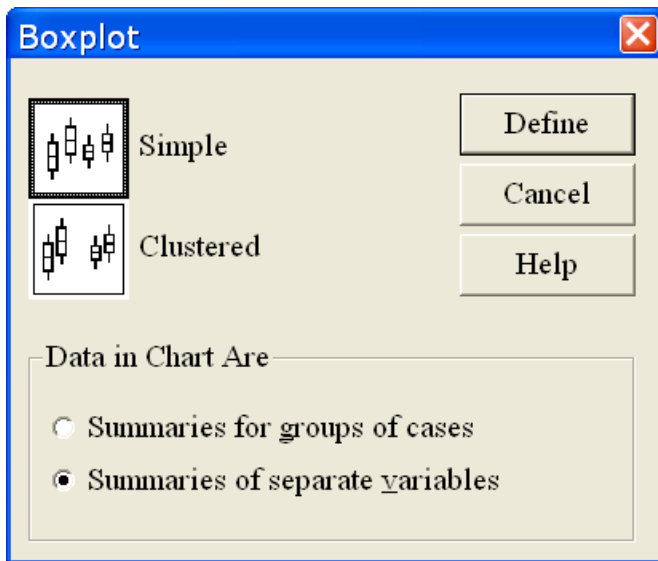
Atsidariusiame lange „*Boxplot*“ nurodome, kad būtų vaizduojamas paprastas atskirų kintamųjų reikšmių apibendrinimas, pažymėdami „*Simple*“ ir „*Summaries of separate variables*“, o po to spaudžiame „*Define*“.

Meniu punktai: Graphics → Box plot

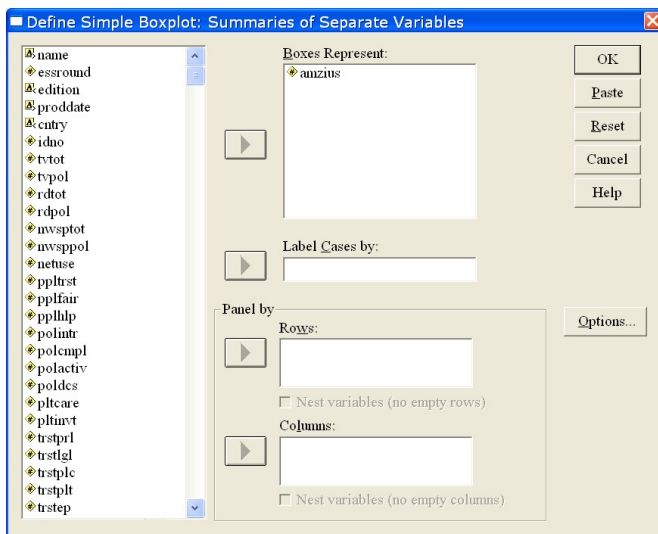


Atsidariusiame lange „*graph box – Box plots*“ parenkame analizės kintamąjį **amžius** ties „*Variables*“ ir spaudžiame „*OK*“.

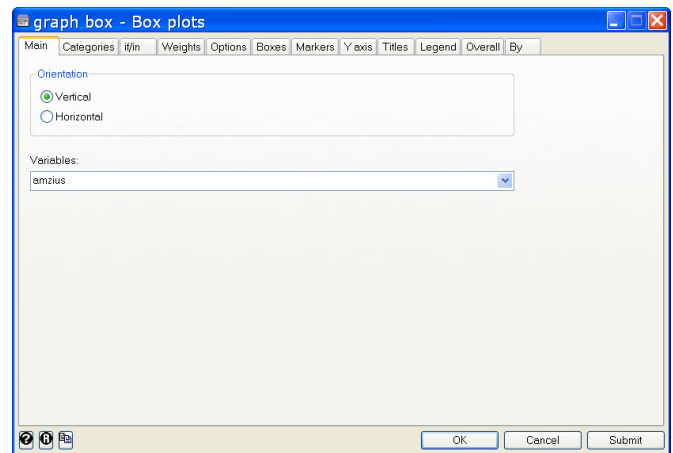
Sprendimas su SPSS 15.0



Atsidariusiame lange „*Define Simple Boxplot: Summaries of Separate Variables*“ į dešinę esantį langelį „*Boxes Represent*“ įkeliam kintamąjį **amzius**, ir spaudžiame „*OK*“.



Sprendimas su Stata SE 10.0

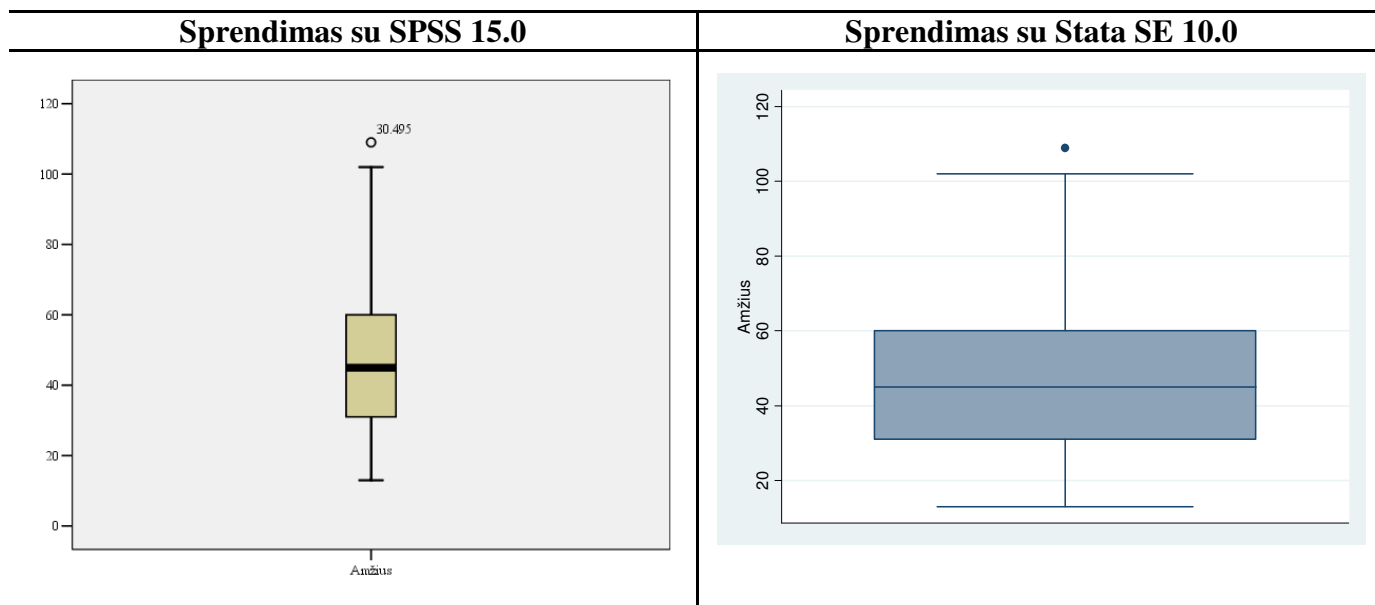


Sintaksės komandos: stačiakampė diagrama

```
EXAMINE VARIABLES = amzius /COMPARE
VARIABLE/PLOT=BOXPLOT/STATISTICS=NONE/NOT
OTAL /MISSING=LISTWISE .
```

```
graph box amzius
```

Gauname tokius rezultatus **rezultatų išvesties lange**:



Taigi histograma ir stačiakampė diagrama²³ leidžia pamatyti bendras kiekybinio kintamojo reikšmių pasiskirstymo tendencijas. Europiečių amžiaus histogramos atveju matome, kad **didžiausia amžiaus reikšmių koncentracija yra ties 40-45 metais**. Iš stačiakampės diagramos matome, kad **pusė visų amžiaus reikšmių (visos vidutinės reikšmės) yra intervale nuo maždaug 30 iki 60 metų**. Taip pat šiame grafike matome ir amžiaus išskirčių (*outliers*), jos signalizuoja tai, kad buvo apklausta žmonių, kurių amžius viršija 100 metų.

Atskleidus bendrąsias **europiečių amžiaus** tendencijas, įdomu **palyginti tas tendencijas skirtingose šalyse**. Empiriniai tyrimai rodo, kad Vakarų Europos šalys labiau nei Rytų Europos šalys susiduria su gyventojų senėjimo problema, taigi Vakarų Europos šalyse gyvena daugiau senyvo amžiaus žmonių nei Rytų Europos šalyse. Todėl pabandysime **palyginti bendrąsias žmonių amžiaus tendencijas Vokietijoje ir Lenkijoje**. Tam atlikti paprastai braižomos **kryžminės stačiakampės diagramos**²⁴. Pavyzdys, kaip tai padaryti su SPSS ir Stata, yra pateiktas 2.9 lentelėje.

2.9 lentelė. Kiekybinių kintamųjų reikšmių pasiskirstymo pogrupiuose bendrų tendencijų palyginimo grafinis vaizdavimas: kryžminės stačiakampės diagramos.

Sprendimas su SPSS 15.0	Sprendimas su Stata SE 10.0
Parengiamieji veiksmai	Parengiamieji veiksmai
Prieš sudarant kryžminę diagramą SPSS pakete reikia iš anksto nurodyti, kad bus analizuojami tik tam tikri (šiuo atveju Vokietijos ir Lenkijos) duomenys. Kaip tą padaryti, jau buvo nurodyta 2.4 lentelėje , todėl čia nekartosime.	Nėra , nes Stata pakete pasirinkti duomenų aibes, su kuriomis bus dirbama, galima atliekant kiekvieną konkrečią analizę.

Sintaksės komandos

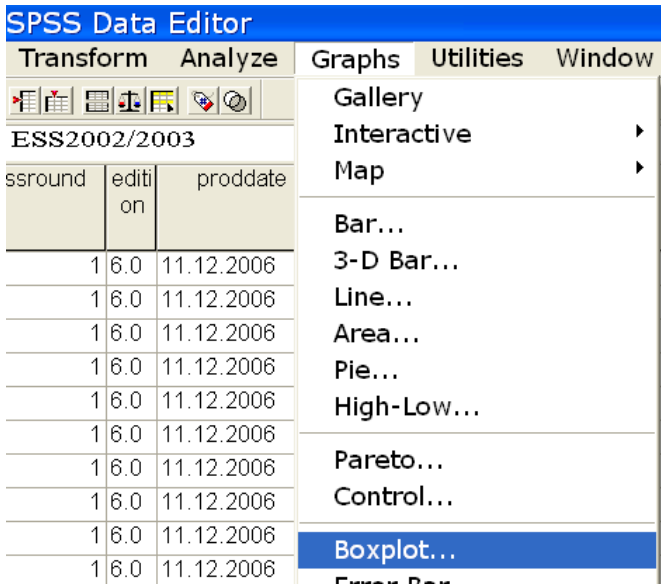
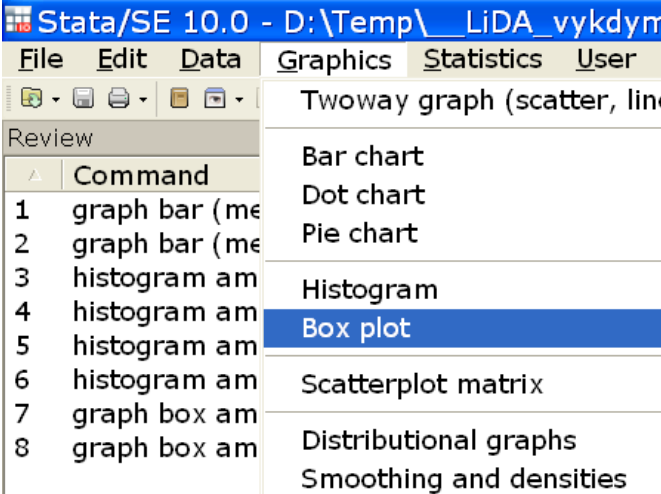
USE ALL. COMPUTE filter_\$=(salis = "Vokietija"	Nėra.
--	-------

²³ Apie tai, ką konkrečiai reiškia stačiakampėje diagramoje vaizduojama informacija, galite pasiskaityti antros dalies 2.4 poskyryje.

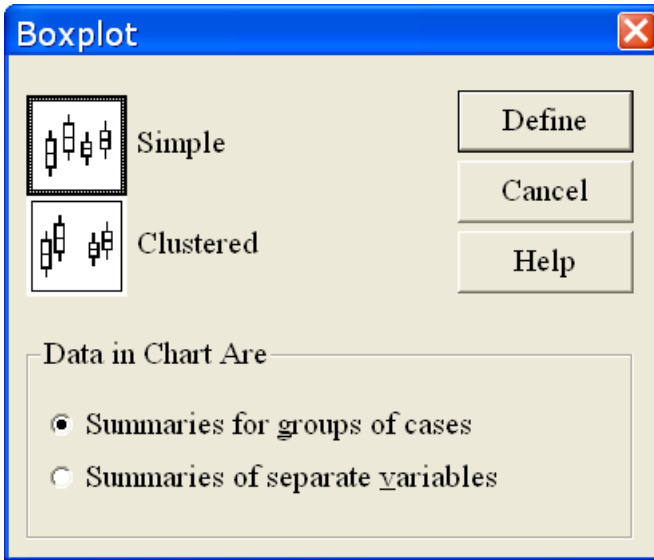
²⁴ Išsamiau apie kiekybinių kintamųjų pogrupių palyginimą bus kalbama šeštajame skyriuje. Čia pateikiami tik bendrieji kiekybinių kintamųjų pogrupių kryžminio grafinio vaizdavimo principai.

Sprendimas su SPSS 15.0	Sprendimas su Stata SE 10.0
<pre> salis = "Lenkija"). VARIABLE LABEL filter_\$ 'salis = "Vokietija" salis = "Lenkija" (FILTER)'. VALUE LABELS filter_\$ 0 'Not Selected' 1 'Selected'. FORMAT filter_\$ (f1.0). FILTER BY filter_\$. EXECUTE . </pre>	

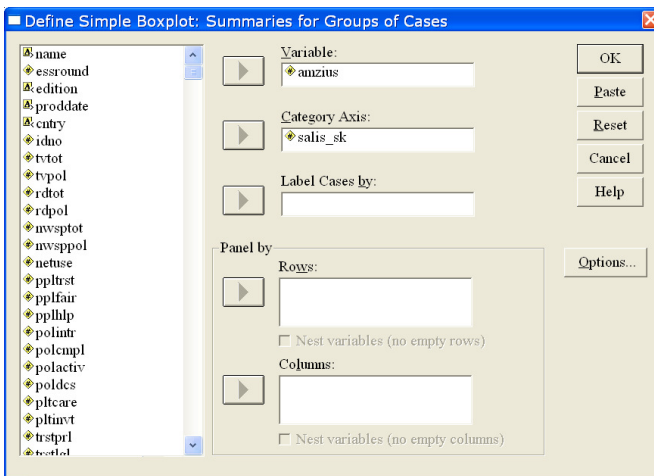
Analizės veiksmiai

Meniu punktai: Graphs → Boxplot...	Meniu punktai: Graphics → Box plot
 The screenshot shows the SPSS Data Editor window with the 'Graphs' menu open. The 'Boxplot...' option is highlighted at the bottom of the menu. The data view shows columns for 'ssround', 'editon', and 'proddate' with values for 11.12.2006.	 The screenshot shows the Stata/SE 10.0 window with the 'Graphics' menu open. The 'Box plot' option is highlighted. The command window shows a list of commands, including 'graph bar', 'graph box', 'histogram', and 'twoway graph'.
Atsidariusiame lange „Boxplot“ nurodome, kad būtų vaizduojamas paprastas kintamojo reikšmių apibendrinimas pagal nustatytas grupes, pažymėdami „<i>Simple</i>“ ir „<i>Summaries for group of cases</i>“, o po to spaudžiame „<i>Define</i>“.	Atsidariusiame lange „graph box – Box plots“ parenkame analizės kintamąjį amzius ties „<i>Variables</i>“.

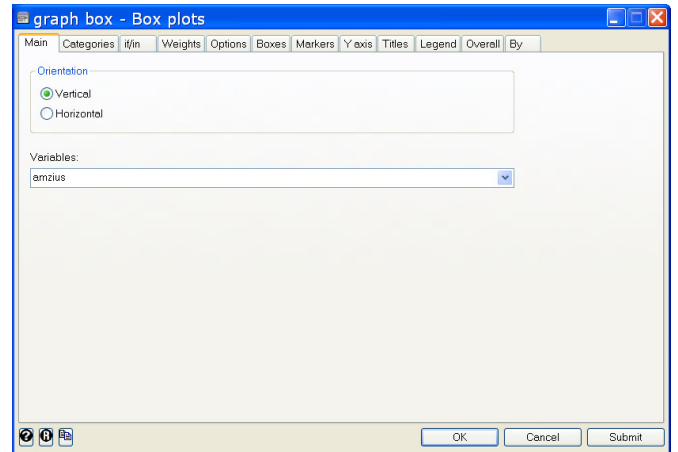
Sprendimas su SPSS 15.0



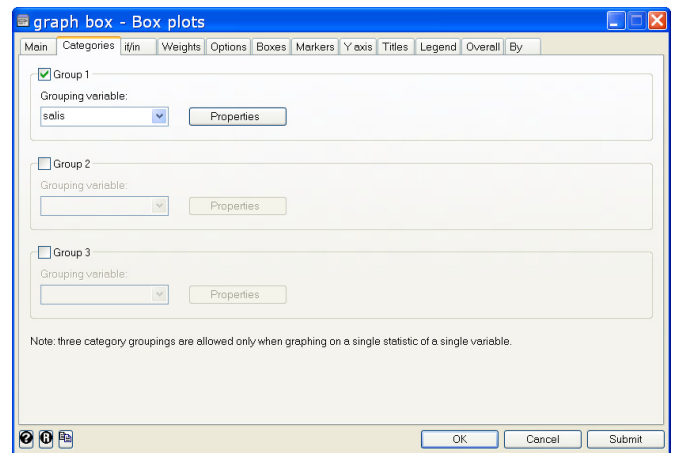
Atsidariusiame lange „**Define Simple Boxplot: Summaries for Groups of Cases**“ į dešinėje esantį langelį „**Variable**“ įkeliam kintamąjį **amzius**, o į langelį „**Category Axis**“ – **salis_sk** ir spaudžiame „**OK**“.



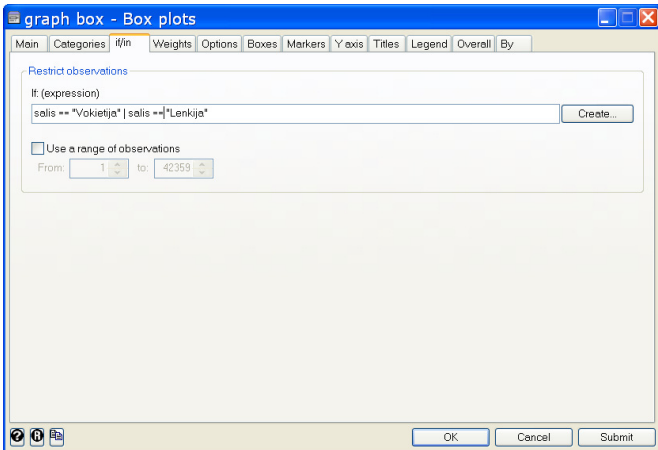
Sprendimas su Stata SE 10.0



Atidarę kortelę „**Categories**“ nurodome, kad duomenys bus grupuojami pažymėdami varnelę ties „**Group 1**“ ir parenkame duomenų grupavimo kintamąjį **salis** ties „**Grouping variable**“.



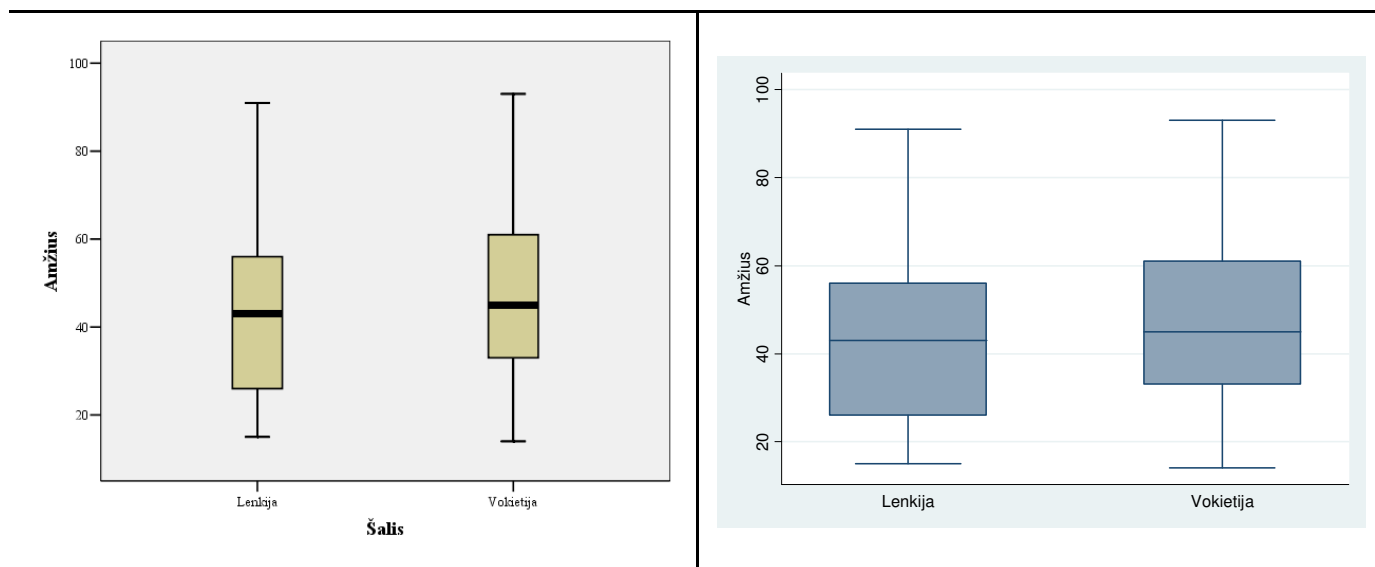
Atidarę kortelę „**if/in**“ nurodome sąlygą ties „**If: (expression)**“: **salis == "Vokietija" | salis == "Lenkija"** ir spaudžiame „**OK**“.

Sprendimas su SPSS 15.0	Sprendimas su Stata SE 10.0
	

Sintaksės komandos

<pre>EXAMINE VARIABLES = amzius BY salis_sk /PLOT = BOXPLOT /STATISTICS = NONE /NOTOTAL .</pre>	<pre>graph box amzius if salis == "Vokietija" salis == "Lenkija", over(salis)</pre>
---	---

Gauname tokius rezultatus **rezultatų išvesties lange**:



Taigi pažiūrėję į stačiakampę diagramą matome, kad kaip ir tikėtasi, Vokietijos gyventojų vidutinės amžiaus skirstinio reikšmės yra didesnės²⁵: **Lenkijoje vidutinės gyventojų amžiaus reikšmės susikongravesios tarp mažiau nei 30 m. ir mažiau nei 60 m., tuo tarpu Vokietijoje – tarp daugiau nei 30 m. ir daugiau nei 60 m.**

2.2. Duomenų sklaidos, padėties ir formos charakteristikos

Pirmame šio skyriaus poskyryje parodėme, kaip atskleisti bendrąsias kiekybinių ir kokybinių kintamųjų reikšmių pasiskirstymo tendencijas. Toliau aptarsime duomenis koncentruotai

²⁵ Tiesa, statistškai reikšmingas skirtumas ir to skirtumo dydis gali būti nustatytas tik atliekant testus. Apie tai kalbama šeštajame skyriuje. Čia pateikiamas tik preliminarus rezultatų aptarimas.

charakterizuojančius rodiklius. Šiuo požiūriu statistikoje egzistuoja **duomenų padėties, sklaidos ir formos charakteristikos**.

Pagrindinės **duomenų aibės reikšmių padėties charakteristikos** yra **moda** (*mode*) – arba dažniausiai pasikartojanti duomenų aibės reikšmė, **mediana** (*median*) – arba vidurinė duomenų aibės (kurios reikšmės išdėstytos nuo mažiausios iki didžiausios) reikšmė ir **vidurkis** (*mean*) – arba vidutinė duomenų aibės reikšmė (taškas, vidutiniškai artimiausias visoms duomenų aibės reikšmėms). Kaip žinia, vidurkis yra labiausiai paplitusi duomenų padėties charakteristika, tačiau reikia prisiminti, kad ji skaičiuojama tik kiekybiniais kintamiesiems. Tuo pačiu reikia vėl priminti, kad moda skaičiuojama tik kokybiniais kintamiesiems, o mediana tinka ir kiekybiniais, ir kokybiniais kintamiesiems (tiesa, tik tiems, kurie išmatuoti tvarkos skalėje)²⁶.

Pagrindinės **duomenų aibės reikšmių sklaidos charakteristikos** yra mažiausia ir didžiausia reikšmė, **duomenų aibės plotis** (*range*) – arba didžiausios ir mažiausios reikšmės skirtumas, **dispersija** (*variance*) – arba kvadratu pakeltų reikšmių nuokrypių nuo vidurkio suma bei **standartinis nuokrypis** (*standard deviation*) – šaknis ištraukta iš dispersijos rodiklio²⁷. Šiuo atveju vėlgi tenka priminti, kad visos šios charakteristikos yra skaičiuojamos **kiekybiniais kintamiesiems ir kokybiniais kintamiesiems išmatuotiems tvarkos skalėje (ir turintiems santykinai daug skirtingų reikšmių, kas leidžia daryti jų panašumo į kiekybinius kintamuosius prielaidą)**.

Pagrindinės **duomenų aibės reikšmių formos charakteristikos** yra **asimetrijos** (*skewness*) – arba duomenų histogramos simetriškumo matas – ir **eksceso** (*kurtosis*) – duomenų histogramos lėkštumo matas – koeficientai. Jos abi skaičiuotinos **kiekybiniais kintamiesiems ir kokybiniais kintamiesiems išmatuotiems tvarkos skalėje bei turintiems santykinai daug reikšmių (tada galima daryti prielaidą, kad jie yra labai panašūs į kiekybinius kintamuosius)**.

Taigi pabandysime trumpai charakterizuoti **europiečių religingumo laipsnį**. Buvo klausta, kiek respondentai įvairiose Europos šalyse mano esą religingi (atsakymų variantai: nuo 1 – visiškai nereliginas, -a iki 10 – labai religingas). Kaip matome, klausimai matuoja religingumo laipsnį **ranginėje skalėje**, t. y. tesiekama sužinoti religingumo laipsnio rangą, kas save laiko daugiau ir kas save laiko mažiau religingu. Kad trumpai apibūdinti ir apibendrinti kintamąjį bei pateikti **jų reikšmių padėties, sklaidos ir formos charakteristikas**, paprastai sudaromos **tokių charakteristikų lentelės**. Pavyzdys, kaip tai padaryti su SPSS ir Stata, yra pateiktas 2.10 lentelėje.

2.10 lentelė. Kintamųjų reikšmių padėties, sklaidos ir formos charakteristikų skaitinis vaizdavimas: aprašomosios statistikos lentelės.

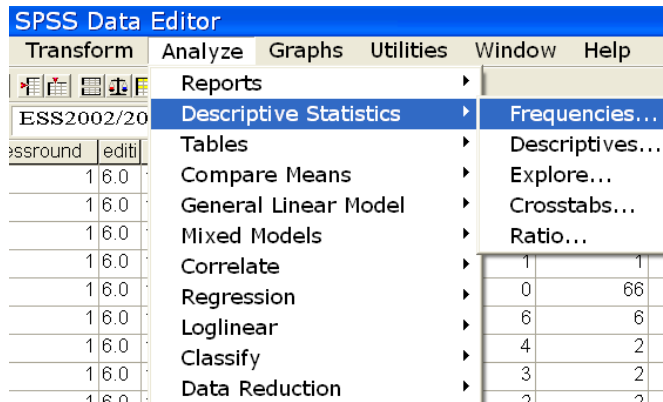
Sprendimas su SPSS 15.0	Sprendimas su Stata SE 10.0
Parengiamieji veiksmai	
Nėra	
Sintaksės komandos	
Nėra	
Analizės veiksmai	
Meniu punktai: Descriptive Statistics →	Meniu punktai: Statistics → Summaries, tables,

²⁶ Kvantilis taip pat yra duomenų padėties charakteristika, tačiau čia ji nebeaptariama, nes apie ją buvo kalbama pirmajame šio skyriaus poskyryje.

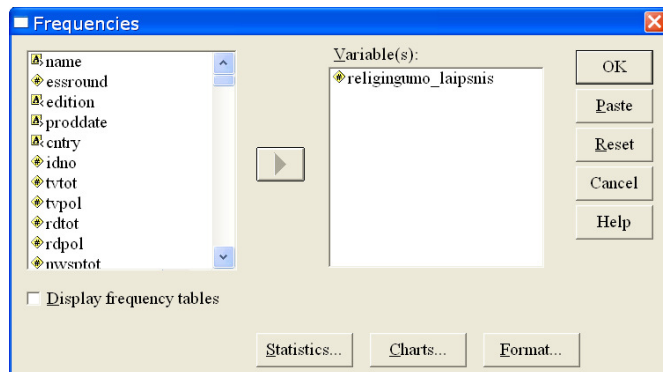
²⁷ Tiesa, kvartilinis plotis taip pat yra duomenų aibės reikšmių sklaidos matas. Čia jis nebeaptariamas, nes jo vaizdavimo būdai buvo pateikti pirmajame šio skyriaus poskyryje, kalbant apie stačiakampių diagramų braižymą.

Sprendimas su SPSS 15.0

Frequencies...



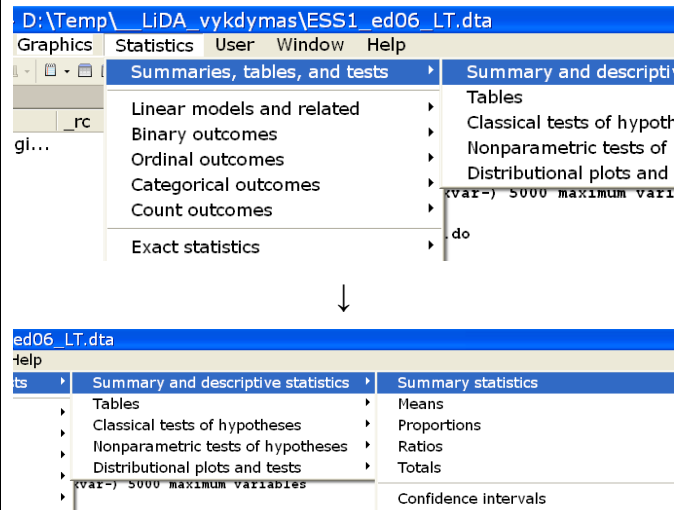
Atsidariusiame lange „**Frequencies**“ į dešinėje esantį langelį „**Variable(s)**“ įkeliam kintamąjį **religingumo_laipsnis**, ir spaudžiame mygtuką „**Statistics...**“.



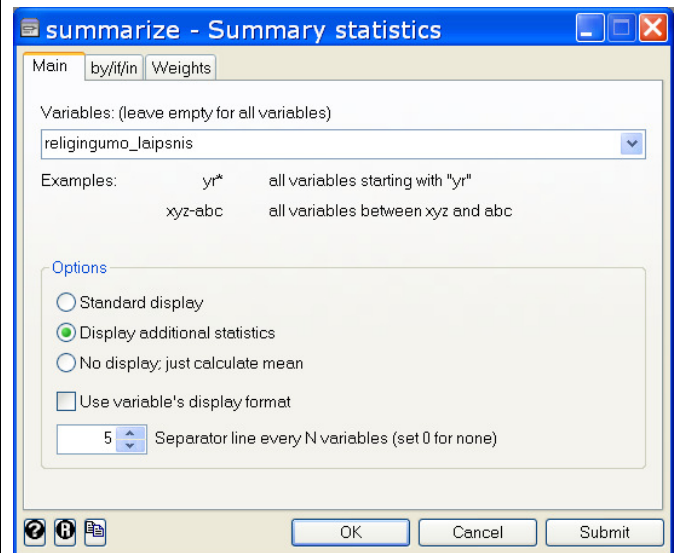
Atsidariusiame lange „**Frequencies: Statistics**“ nurodome, kad būtų skaičiuojamas vidurkis (**Mean**), mediana (**Median**), moda (**Mode**), reikšmių aibės plotis (**Range**), didžiausia (**Maximum**) ir mažiausia (**Minimum**) reikšmė, dispersija (**Variance**) ir standartinis nuokrypis (**Standard deviation**), asimetrijos (**Skewness**) ir eksceso (**Kurtosis**) koeficientai pažymėdami atitinkamas varneles. Tada spaudžiame „**Continue**“, o sugrįžus į „**Frequencies**“ langą spaudžiame „**OK**“.

Sprendimas su Stata SE 10.0

and tests → Summary and descriptive statistics → Summary statistics



Atsidariusiame lange „**summarize – Summary statistics**“ parenkame analizės kintamąjį ties „**Variables: ...**“ **religingumo_laipsnis**, o ties parinktimis „**Options**“ nurodome pateikti išsamią ataskaitą „**Display additional statistics**“ ir spaudžiame „**OK**“.



Sprendimas su SPSS 15.0

Sprendimas su Stata SE 10.0

Sintaksės komandos

```
FREQUENCIES VARIABLES =
religingumo_laipsnis /FORMAT = NOTABLE
/STATISTICS = STDDEV VARIANCE RANGE
MINIMUM MAXIMUM MEAN MEDIAN MODE SKEWNESS
SESKEW KURTOSIS SEKURT /ORDER = ANALYSIS.
```

```
summarize religingumo_laipsnis, detail
```

Gauname tokius rezultatus **rezultatų išvesties lange**:

Statistics		
Religingumo laipsnis		
N	Valid	42067
	Missing	292
Mean		4,95
Median		5,00
Mode		5
Std. Deviation		2,955
Variance		8,733
Skewness		-,162
Std. Error of Skewness		,012
Kurtosis		-,944
Std. Error of Kurtosis		,024
Range		10
Minimum		0
Maximum		10

Religingumo laipsnis					
Percentiles			Smallest		
1%	0		0		
5%	0		0		
10%	0		0	Obs	42067
25%	3		0	Sum of Wgt.	42067
50%	5			Mean	4.947465
			Largest	Std. Dev.	2.955243
75%	7		10		
90%	9		10	Variance	8.733464
95%	10		10	Skewness	-,1618421
99%	10		10	Kurtosis	2.056329

Taigi iš aprašomosios statistikos lentelių matome, kad **vidutinė** religingumo reikšmė yra visiškai ties reikšmių intervalo viduriu – beveik 5. Be to, **vidurinės bei dažniausios reikšmės** įverčiai taip pat yra 5 santykiniai religingumo laipsnio vienetai. Tokiais atvejais yra didelė tikimybė, kad kintamojo skirstinys yra **normalusis**. Tiesa, šią prielaidą kvestionuoja **neigiamas asimetrijos** (šiek tiek daugiau reikšmių koncentruojasi kairėje pusėje nuo vidurkio) ir **eksceso** (santykinai daug reikšmių pasklidę toli nuo vidurkio) koeficientai. Pastarosios išvados hipotetini teisingumą sustiprina ir į tai, kad **standartinis nuokrypis** sudaro beveik 3 vienetus, taigi reikšmės gana „plačiai pasklidusios apie vidurkį“.

Glaustai charakterizavus europiečių religingumo laipsnį, įdomu **palyginti religingumo laipsnio charakteristikas skirtingose šalyse**. Empiriniai tyrimai rodo, kad Vakarų Europos šalys yra labiau sekuliarizuotos nei Rytų Europos šalys, taigi Vakarų Europos šalyse – ypač

Skandinavijoje – žmonės turėtų būti mažiau religingi nei Rytų Europos šalyse. Kad patikrinti šią prielaidą pabandysime **palyginti religingumo laipsnio padėties charakteristikas Švedijoje ir Lenkijoje**. Tam atlikti paprastai braižomos **kryžminės stulpelinės (linijinės ar taškinės) duomenų padėties charakteristikų diagramos**²⁸. Pavyzdys, kaip tai padaryti su SPSS ir Stata, yra pateiktas 2.11 lentelėje.

2.11 lentelė. Kintamųjų reikšmių aibės padėties charakteristikų palyginimo grafinis vaizdavimas: kryžminės stulpelinės²⁹ vidurkių³⁰ diagramos.

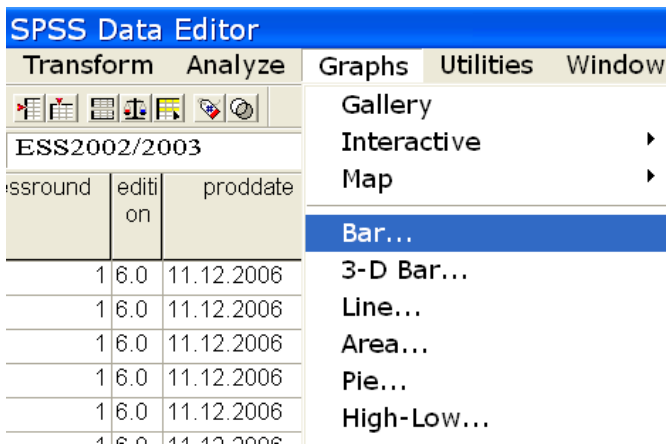
Sprendimas su SPSS 15.0	Sprendimas su Stata SE 10.0
Parengiamieji veiksmai	Parengiamieji veiksmai
Prieš sudarant kryžminę diagramą SPSS pakete reikia iš anksto nurodyti, kad bus analizuojami tik tam tikri (šiuo atveju Švedijos ir Lenkijos) duomenys. Kaip tą padaryti, jau buvo nurodyta 2.4 lentelėje , todėl čia nekartosime.	Nėra , nes Stata pakete pasirinkti duomenų aibes, su kuriomis bus dirbama, galima atliekant kiekvieną konkrečią analizę.
Sintaksės komandos	
<pre>USE ALL. COMPUTE filter_\$=(salis = "Švedija" salis = "Lenkija"). VARIABLE LABEL filter_\$ 'salis = "Švedija" salis = "Lenkija" (FILTER)'. VALUE LABELS filter_\$ 0 'Not Selected' 1 'Selected'. FORMAT filter_\$ (f1.0). FILTER BY filter_\$. EXECUTE .</pre>	Nėra.
Analizės veiksmai	
Meniu punktai: Graphs → Bar...	Meniu punktai: Graphics → Bar Chartt

²⁸ Išsamiau apie kiekybinių kintamųjų pogrupių reikšmių charakteristikų palyginimą bus kalbama penktajame skyriuje. Čia pateikiami tik bendrieji kiekybinių kintamųjų pogrupių reikšmių charakteristikų kryžminio grafinio vaizdavimo principai.

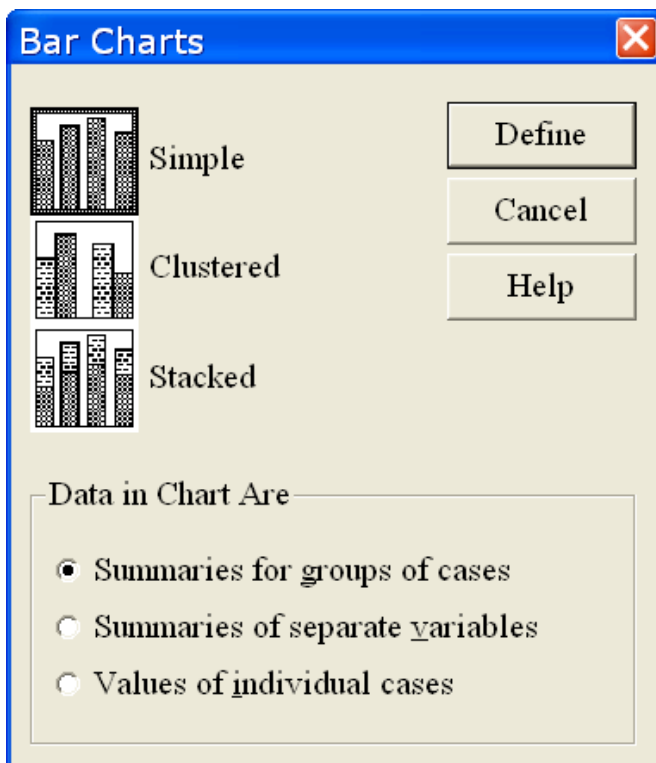
²⁹ Kaip minėta, yra ir kitų grafinio vaizdavimo variantų – pavyzdžiui, braižyti taškinės arba linijinės diagramas, – tačiau braižymo algoritmai yra tie patys, taigi pavyzdžiuose pateiksime tik populiariausią – stulpelinių diagramų – atvejį.

³⁰ Taip pat galima braižyti įvairių kitų charakteristikų diagramas, tačiau vėlgi – algoritmas yra tas pats, todėl čia patiekiamas tik populiariausias variantas – kryžminė stulpelinė vidurkių diagrama.

Sprendimas su SPSS 15.0

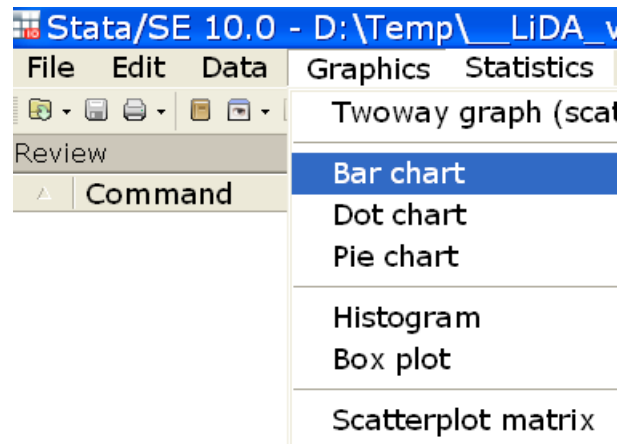


Atsidariusiame lange „**Bar Charts**“ nurodome, kad būtų vaizduojamas paprastas grafikas su kintamojo reikšmių charakteristikomis pagal nustatytas grupes, pažymėdami „**Simple**“ ir „**Summaries for group of cases**“, ir spaudžiame „**Define**“.

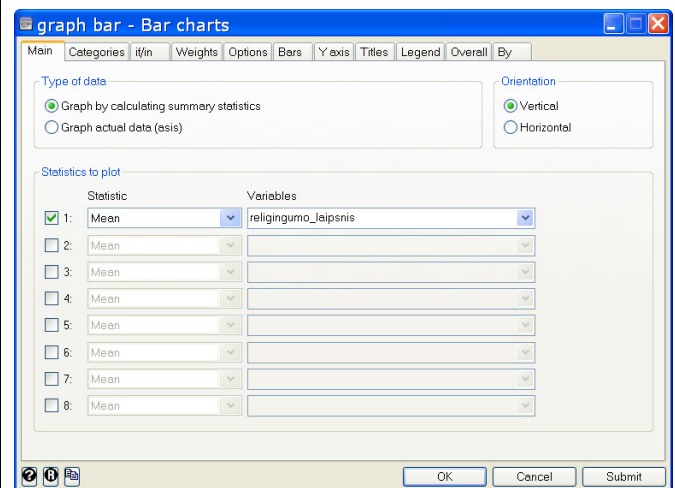


Atsidariusiame lange „**Define Simple Bar: Summaries for Groups of Cases**“ į dešinėje esantį langelį „**Variable**“ įkeliam kintamąjį **relingumo_laipsnis** (matome, kad programa braižys vidurkių grafiką, tačiau šią parinktį galima pakeisti), o į langelį „**Category Axis**“ – **salis** ir spaudžiame „**OK**“.

Sprendimas su Stata SE 10.0

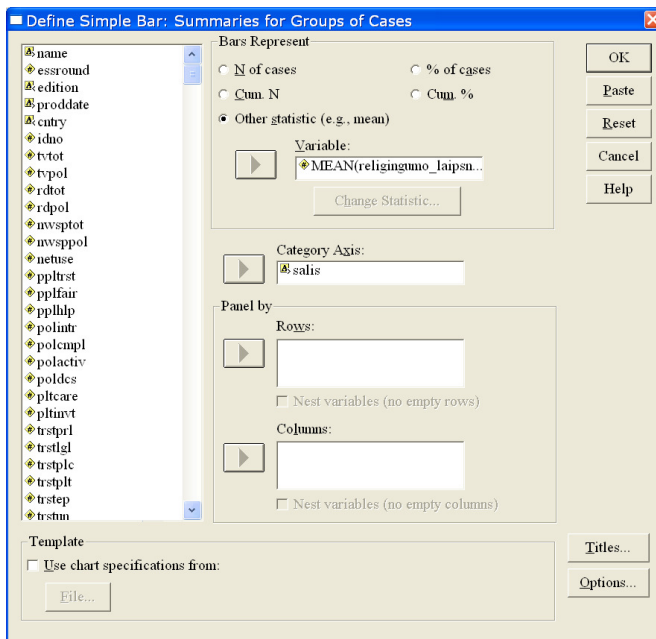


Atsidariusiame lange „**graph bar – Bar chart**“ parenkame analizės kintamąjį **relingumo_laipsnis** ties „**Variables**“.

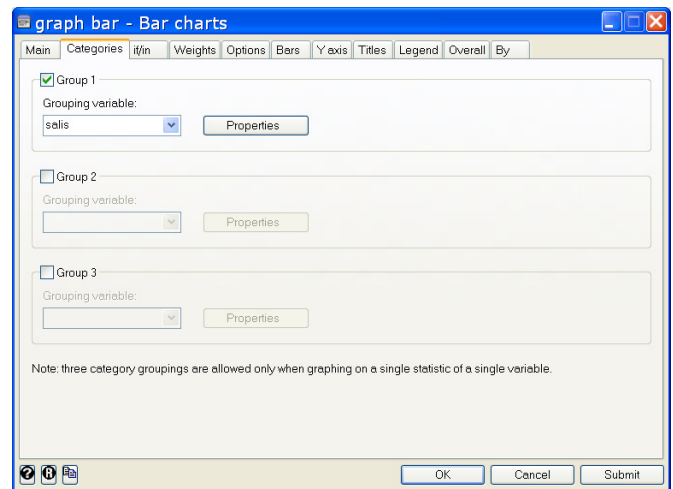


Atidare kortelę „**Categories**“ nurodome, kad duomenys bus grupuojami, pažymėdami varnelę ties „**Group 1**“, ir parenkame duomenų grupavimo kintamąjį **salis** ties „**Grouping variable**“.

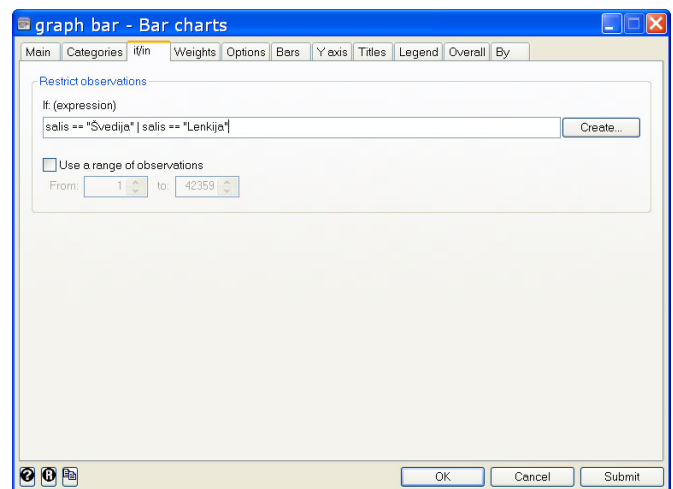
Sprendimas su SPSS 15.0



Sprendimas su Stata SE 10.0



Atidarekite kortelę „*if/in*“ nurodome sąlygą ties „*If: (expression)*“: `salis == "Švedija" | salis == "Lenkija"` ir spaudžiame „OK“.

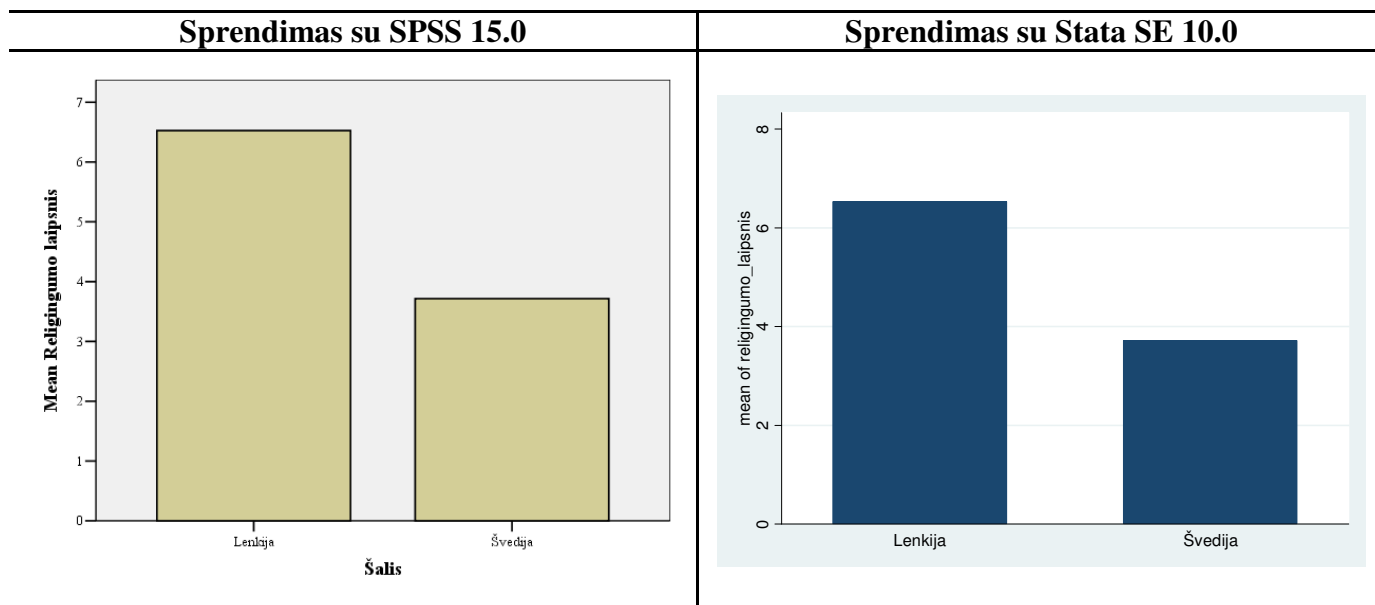


Sintaksės komandos

```
GRAPH /BAR(SIMPLE) =  
MEAN(religiumo_laipsnis) BY salis .
```

```
graph bar (mean) religiumo_laipsnis if  
salis == "Švedija" | salis == "Lenkija",  
over(salis)
```

Gauname tokius rezultatus **rezultatų išvesties lange**:



Taigi matome, kad kaip ir tikėtasi, Švedijos gyventojai yra gerokai mažiau religingi nei Lenkijos³¹: Lenkijos gyventojų vidutinis religingumo laipsnis yra virš 6,5 punktų, o Švedijos gyventojų – tik kiek virš 3,5 punktų.

³¹ Tiesa, statistiškai reikšmingas skirtumas ir to skirtumo dydis gali būti nustatytas tik atliekant statistinius testus. Apie tai kalbama penktajame skyriuje. Čia pateikiamas tik preliminarus rezultatų aptarimas.

3. PASIKLIAUTINIEJI INTERVALAI

Jei aprašomoji statistika yra susijusi su gana paprastais uždaviniais – duomenų (populiacijos arba imties požymių) aprašymu ir sisteminiu, tai **išvadų statistika** kelia sau gerokai didesnius uždavinius – remiantis imtimi padaryti išvadą apie visą populiaciją, iš kurios yra atrinkta toji imtis. Ši statistikos dalis buvo sukurta ir išstobulinta dėl to, kad praktiškai labai sunku nagrinėti išstisus masinius reiškinius, todėl paprastai iš jų atrenkamos imtys (tam tikru tikimybinio būdu³²) ir pagal jas **daromos išvados apie tų masinių reiškinių požymių apibendrintas charakteristikas** (vadinamuosius *parametrus*)³³.

Išvados apie masinių reiškinių (populiacijos) požymių apibendrintas charakteristikas daromos **laikantis intervalinių įverčių sudarymo principų**, t. y. pagal imties duomenis apskaičiuojamas taškinis įvertis ir aplink jį (**remiantis tikimybių teorijos ir matematinės statistikos metodais**) sudaromas intervalas, kurio ribos atitinka norimo tikslumo ir pasiklovimo lygmens kriterijus. Tame **intervale (su tam tikra garantija, paprastai, 95 proc.) yra tikroji ieškomos populiacijos požymio apibendrintos charakteristikos reikšmė**, nors neišnagrinėję visos populiacijos konkrečios tikrosios reikšmės mes niekada ir nesužinosime. Tiesa, tai netrukdo daryti masinių reiškinių požymių apibendrintų charakteristikų praktinių palyginimų, ieškoti koreliacijų tarp jų, ar jų priklausomumo viena nuo kitos ir pan., nes intervaliniai įverčiai turi savo ribas. Tam tikra prasme, intervaliniai įverčiai ir jų ribos yra išvadų statistikos pagrindas. Taigi prieš pereinat prie pagrindinės išvadų statistikos dalies – statistinių hipotezių tikrinimo – aptarimo, paprastai, statistikoje supažindinama su intervalinių įverčių sudarymo principais. Masinių reiškinių požymių apibendrintos charakteristikos, kurioms intervaliniai įverčiai yra sudarinėjami dažniausiai, yra **populiacijos požymio vidurkis ir proporcija**. Taigi šiame skyriuje parodysime, kaip sudaryti šių parametru intervalinius įverčius su SPSS ir Stata.

Pateikdami pasikliautiniųjų intervalų sudarymo pavyzdžius su SPSS ir Stata programiniais paketais naudosime Europos socialinio tyrimo (ESS) antrosios bangos (2004 m.) duomenis (SPSS laikmena – „*ESS2_e03_LT.sav*“, Stata laikmena – „*ESS2_e03_LT.dta*“). Šio tyrimo ir jame tyrinėjamų reiškinių bei kintamųjų aprašas yra pateiktas pirmojoje studijų paketo dalyje.

3.1. Populiacijos vidurkio pasikliautinis intervalas

Pirmiausiai aptarsime tai, kaip sudaryti **kiekybinio populiacijos požymio vidurkio pasikliautinąjį intervalą**. Kadangi vidurkis skaičiuojamas tik kiekybiniais kintamiesiems, tai tiriamas masinio reiškinių požymis turi būti išmatuotas³⁴ intervalų arba santykių skalėje. Taigi mūsų **tikslas** – sužinoti, kokia yra vidutinė mus dominančio kiekybinio masinio reiškinių požymio reikšmių skirstinio reikšmė. Kaip jau minėta, kadangi netirsime viso masinio reiškinių, o tik iš jo atrinktos imties duomenis, tai **tikrosios reikšmės nesužinosime**. Tačiau statistikos mokslo pagalba **su tam tikra garantija (paprastai 95 proc.) galime nurodyti intervalą**, kuriame ši reikšmė yra.

³² Čia ir toliau daroma prielaida (paprastumo dėlei), kad visi į pavyzdinę duomenų bazę įtraukti duomenys atrinkti vadovaujantis atsitiktinės tikimybinės atrankos logika, nors taip iš tiesų nėra.

³³ Tiesa, neparametrinėje statistikoje išvados daromos ne apie masinių reiškinių apibendrintas charakteristikas (*parametrus*), o apie jų reikšmių skirstinių pobūdį.

³⁴ Tiksliau, „turi būti iš principo matuojamas“, t.y. tas reiškinys turi būti suvokiamas kaip kiekybinis savo esme, todėl galimas išmatuoti kiekybinėje skalėje.

Kad iliustruotume tai, kaip sudaryti kiekybinio populiacijos požymio vidurkio pasikliautinąjį intervalą, panagrinėsime **europiečių³⁵ pasitenkinimą savo šalių ekonomika**. Buvo klausta, kiek respondentai įvairiose Europos šalyse yra patenkinti tuo, kaip veikia šalies ekonomika (atsakymų variantai: nuo 0 – visiškai nepatenkintas,-a iki 10 – visiškai patenkintas,-a). Kaip matome, klausimas matuoja pasitenkinimą ekonomika **tvarkos skalėje**, tačiau pasitenkinimo rangų yra santykinai didelis kiekis (11), todėl tokį kintamąjį **galime laikyti kiekybiniu³⁶**.

Bendroji kiekybinio masinio reiškinių požymio **vidutinės reikšmės intervalinio įverčio sudarymo logika** yra tokia³⁷:

Pasirenkame **pasiklivimo lygmenį $1 - \alpha$ ³⁸**, t. y. nusistatome, kiek norime būti garantuoti savo išvados tikrumu (**paprastai 95 proc.**);

Suskaičiuojame **taškinį** ieškomos kiekybinio masinio reiškinių požymio reikšmių vidurkio **įvertį** pagal atrinktos imties duomenis;

Parentame tinkamą statistiką, kurios pagrindu bus nustatytos intervalinio įverčio ribos, pagal tiriamo masinio reiškinių savybes (reikšmių skirstinio pobūdį³⁹), nežinomų to masinio reiškinių parametrų skaičių ir imties dydį;

Radę reikiamus (paprastai $\alpha/2$ ir $1-\alpha/2$) pasirinktos statistikos teorinio skirstinio kvantilius, **suskaičiuojame intervalinį įvertį**.

Pavyzdys, kaip randamas intervalinis vidurkio įvertis su SPSS ir Stata, yra pateiktas 3.1 lentelėje.

3.1 lentelė. Normaliojo kiekybinio masinio reiškinių požymio reikšmių vidurkio pasikliautinąjo intervalo radimas.

Sprendimas su SPSS 15.0	Sprendimas su Stata SE 10.0
Parengiamieji veiksmai	
Nėra	
Sintaksės komandos	
Nėra	
Analizės veiksmai	
Meniu punktai: Analyze → Descriptive Statistics → Explore...	Meniu punktai: Statistics → Summaries, tables, and tests → Summary and descriptive statistics → Confidence intervals

³⁵ Į ESS antrąją bangą įtrauktų šalių sąrašas pateiktas pirmoje studijų paketo dalyje (arba žr. http://www.europeansocialsurvey.org/index.php?option=com_content&task=view&id=49&Itemid=83). Taigi terminas „europiečiai“ šiuo atveju apima ne visas Europos šalis, o tik tas, kuriose tyrimas buvo vykdytas 2004 m.

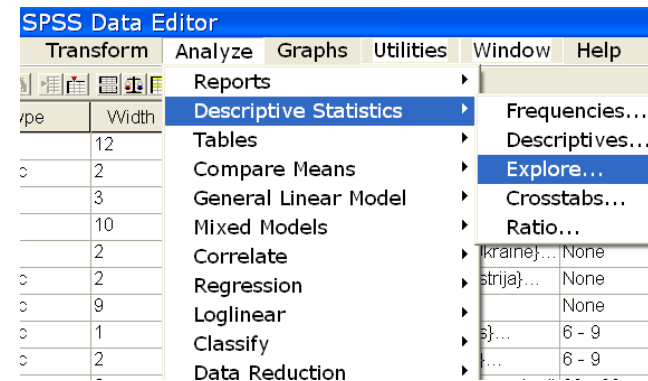
³⁶ Be to, daroma prielaida, kad šio kintamojo skirstinys suderinamas su normaliuoju. Kaip tikrinti skirstinių suderinamumo hipotezes kalbama 4.2. poskyryje, todėl čia kol kas ši prielaida netikrinama.

³⁷ Išsamiau apie tai kalbama antroje šio studijų paketo dalyje, 3 skyriuje.

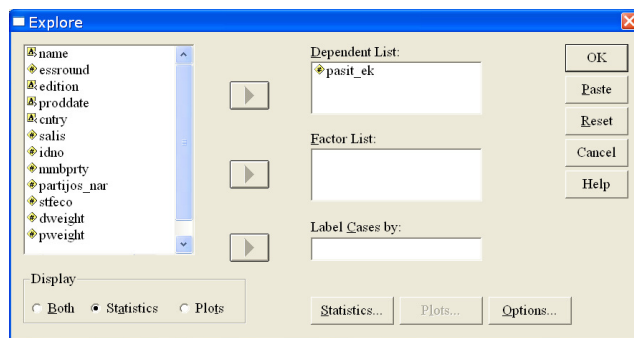
³⁸ α yra reikšmingumo lygmuo. Čia vertėtų priminti ir tai, kad pasiklivimo lygmuo, imties dydis ir intervalo tikslumas yra susiję dydžiai: kuo didesnis pasiklivimo lygmuo, tuo mažesnis intervalo tikslumas (intervalas platesnis); kuo didesnė imtis, tuo didesnis intervalo tikslumas (intervalas siauresnis); kuo didesnis intervalo tikslumas (intervalas siauresnis), tuo mažesnis pasiklivimo lygmuo.

³⁹ Tiesa, čia bus pristatomas tik bazinis atvejis, kai tiriamo masinio reiškinių požymio reikšmių skirstinio forma yra suderinama su normaliojo skirstinio forma bei turimi du nežinomi parametrai: vidurkis ir dispersija. Taigi tinkama statistika tokiu atveju laikoma *Stjudento t*.

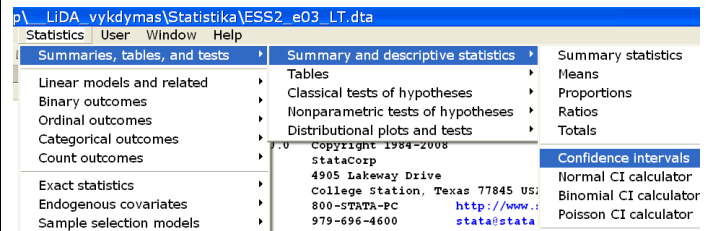
Sprendimas su SPSS 15.0



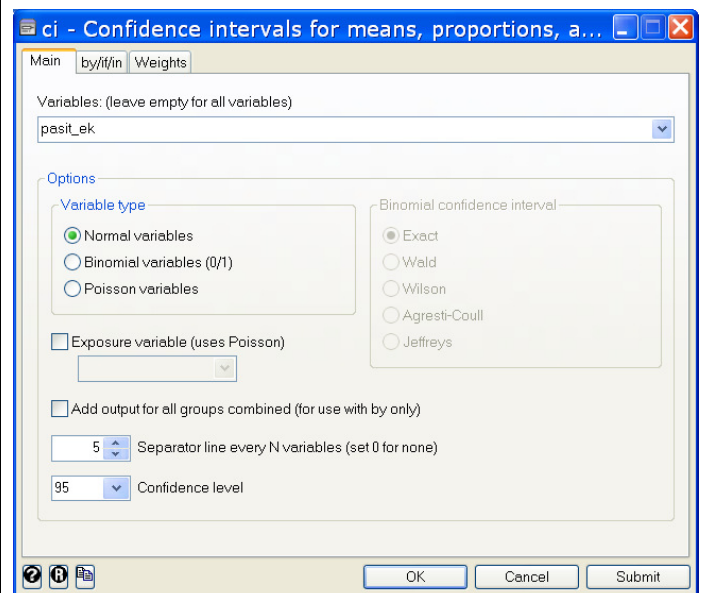
Atsidariusiame lange „*Explore*“ į dešinę pusėje esantį langelį „*Dependent List*“ įkeliam kintamąjį **pasit_ek**, pasirenkame rodyti tik skaitinę informaciją pažymėdami „*Statistics*“ ties „*Display*“ ir spaudžiame „*OK*“.



Sprendimas su Stata SE 10.0



Atsidariusiame lange „*ci – Confidence intervals for means, proportions, and counts*“ parenkame analizės kintamąjį **pasit_ek** ties „*Variables:...*“ ir spaudžiame „*OK*“.



Sintaksės komandos

```
EXAMINE VARIABLES=pasit_ek /PLOT NONE
/STATISTICS DESCRIPTIVES /CINTERVAL 95
/MISSING LISTWISE /NOTOTAL.
```

```
ci pasit_ek
```

Gauname tokius rezultatus **rezultatų išvesties lange**:

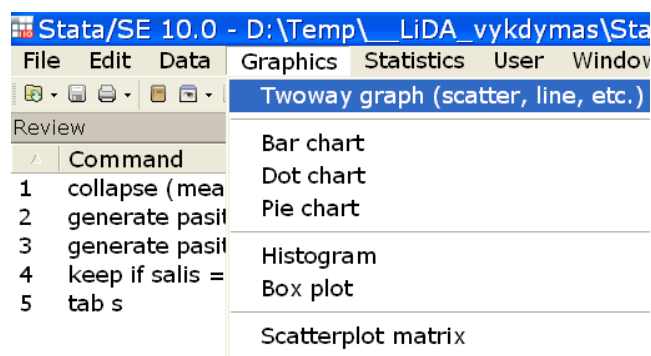
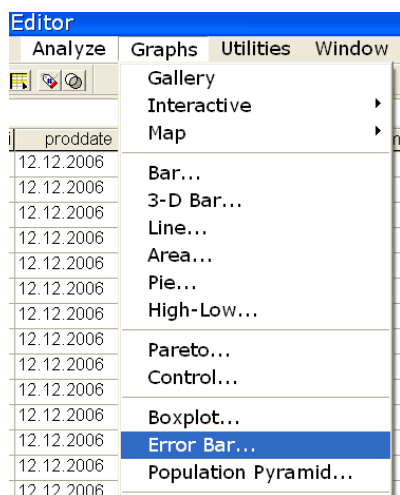
Sprendimas su SPSS 15.0				Sprendimas su Stata SE 10.0					
Descriptives									
Pasitenkinimas tuo, kaip veikia šalies ekonomika	Mean		Statistic	Std. Error					
	95% Confidence Interval for Mean	Lower Bound	4,65	,012					
		Upper Bound	4,63						
			4,68						
	5% Trimmed Mean		4,66						
	Median		5,00						
	Variance		6,107						
	Std. Deviation		2,471						
	Minimum		0						
	Maximum		10						
	Range		10						
	Interquartile Range		4						
	Skewness		-,078	,011					
	Kurtosis		-,698	,023					

Variable	Obs	Mean	Std. Err.	[95% Conf. Interval]	
pasit_ek	46077	4.654448	.0115124	4.631884	4.677012

Sprendimas su SPSS 15.0	Sprendimas su Stata SE 10.0
EXECUTE .	pasit_ek_semean keep if salis == 14 salis == 19

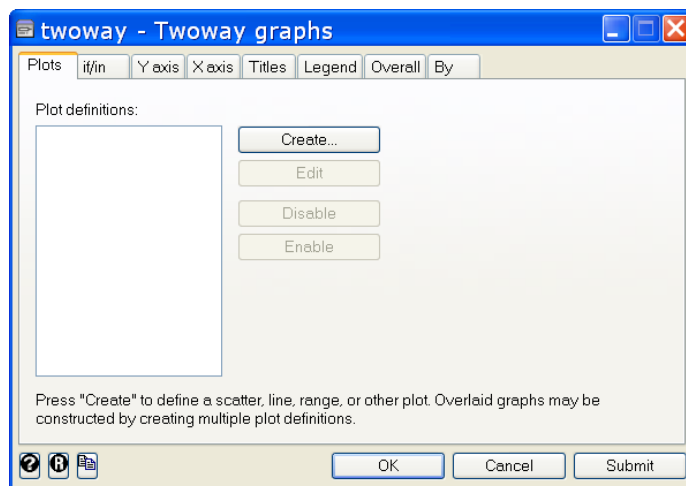
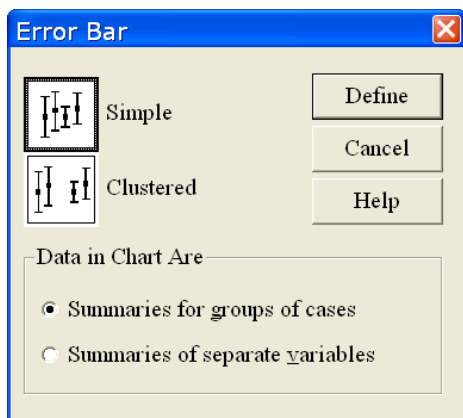
Analizės veiksmas

Meniu punktai: Analyze → Graphs → Error Bar...	Meniu punktai: Graphics → Twoway graph (scatter, line, etc.)
---	---



Atsidariusiame lange „**Error Bar**“ nurodome, kad būtų vaizduojamas paprastas grafikas su kintamojo reikšmių charakteristikomis pagal nustatytas grupes, pažymėdami „**Simple**“ ir „**Summaries for group of cases**“, ir spaudžiame „**Define**“.

Atsidariusiame lange „**twoway – Twoway graphs**“ spaudžiame mygtuką „**Create...**“, kad sukurti pirmąją grafiko dalį.

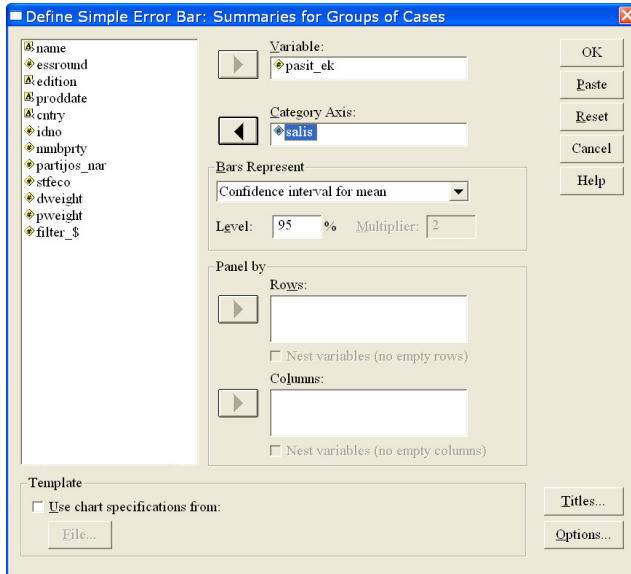


Atsidariusiame lange „**Define Simple Error Bar: Summaries for Groups of Cases**“ į dešinę esantį langelį „**Variable**“ įkeliam kintamąjį **pasit_ek** (matome ties „**Bars Represent**“, kad programa braižys vidurkio pasikliautinųjų intervalų grafiką, tačiau šią parinktį galima

Atsidarius langui „**Plot 1**“ sukuriame pirmąją grafiko dalį – vidurkių taškinį grafiką – pasirinkdami „**Scatter**“ ties „**Basic plots**“, įkeldami **pasit_ek_mean** kintamąjį ties „**Y variable**“ ir **salis** – ties „**X variable**“ bei parinkdami mažą apskritimą kaip taško grafike simbolį (spaudžiame mygtuką

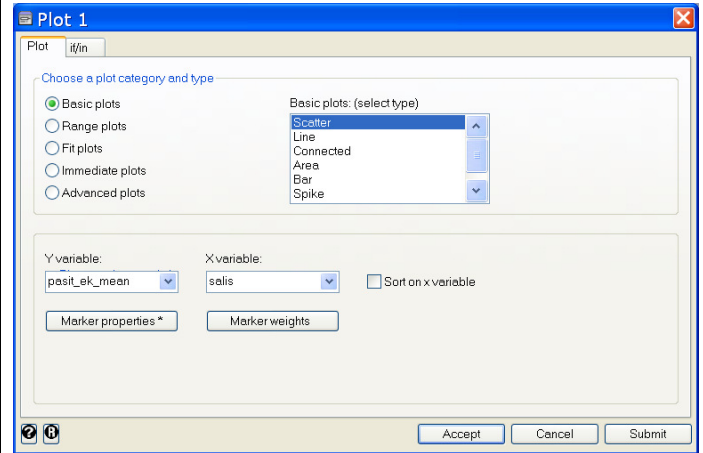
Sprendimas su SPSS 15.0

pakeisti), o į langelį „**Category Axis**“ – **salis** ir spaudžiame „**OK**“.

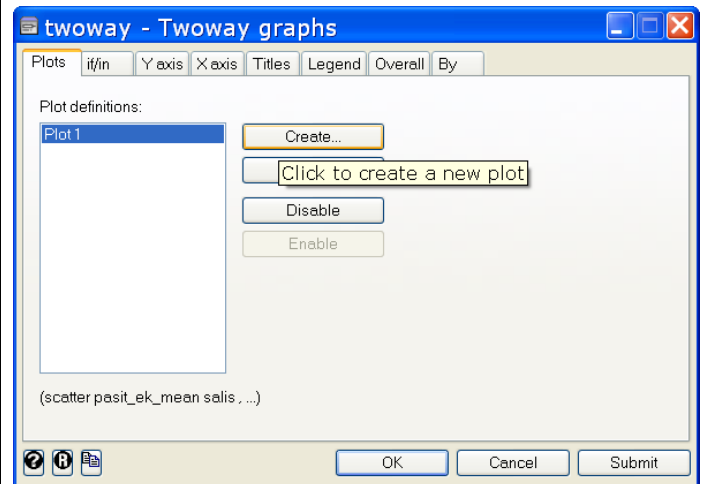


Sprendimas su Stata SE 10.0

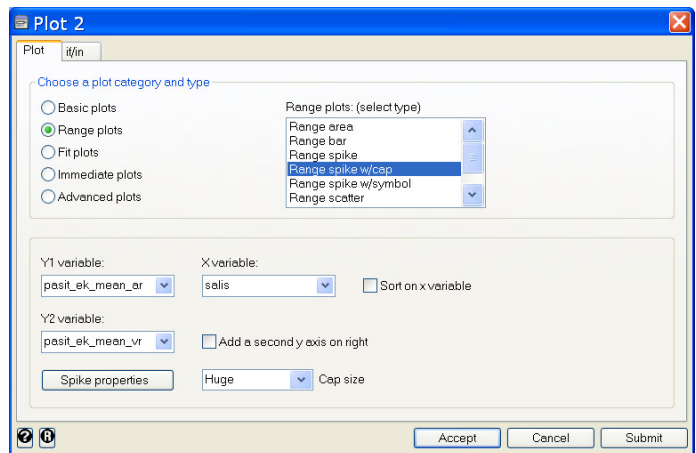
„**Marker properties**“ ir pasirenkame „**Small circle**“ ties „**Symbol**“ bei paspaudžiame „**Accept**“). Po to, spaudžiame „**Accept**“.



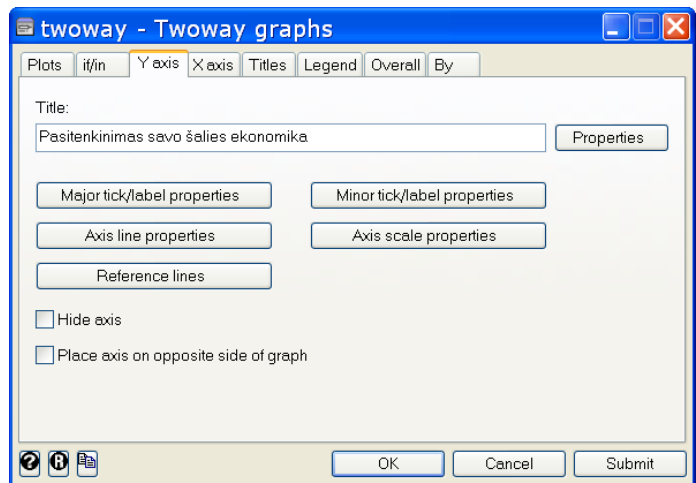
Grįžę į „**twoway – Twoway graphs**“ langą dar kartą spaudžiame mygtuką „**Create...**“, kad sukurti antrąją grafiko dalį.



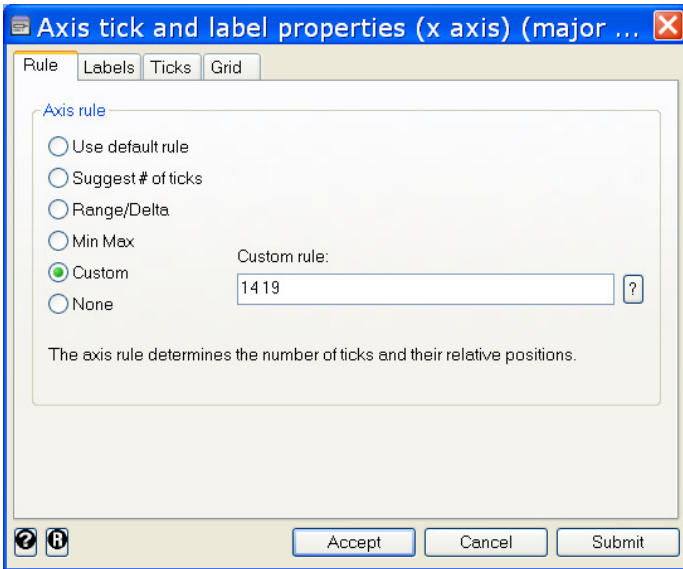
Atsidarius langui „**Plot 2**“ sukuriame antrąją grafiko dalį – vidurkio pasikliautinojo intervalo ribų grafiką – pasirinkdami „**Range plots**“ ir „**Range spike w/ caps**“ ties „**Basic plots**“, įkeldami **pasit_ek_mean_ar** kintamąjį ties „**Y1 variable**“, **pasit_ek_mean_vr** – ties „**Y2 variable**“ ir **salis** – ties „**X variable**“ bei parinkdami didelius ribų žymeklius „**Huge**“ ties „**Cap size**“. Po to, spaudžiame „**Accept**“.



Grįžę į „*twoway – Twoway graphs*“ langą, nurodome Y ašies pavadinimą kortelėje „*Y axis*“ ties „*Title*“ įrašydami Pasitenkinimas savo šalies ekonomika).



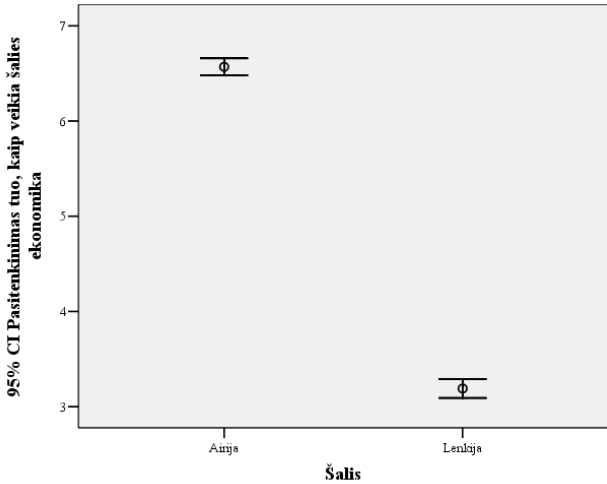
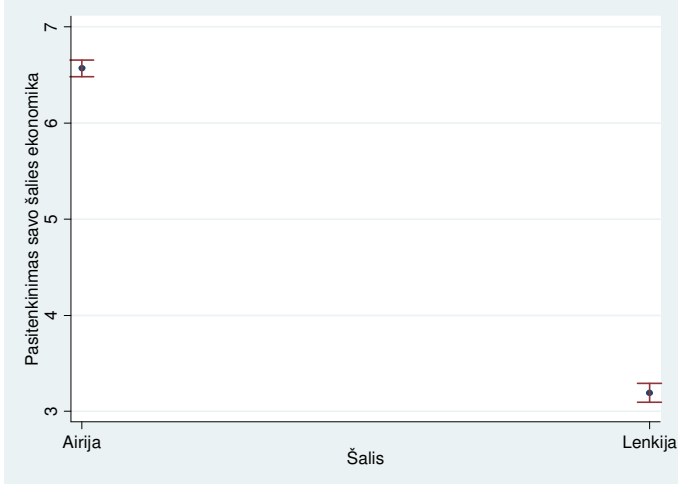
Po to, kortelėje „*X axis*“ nurodome, kad X ašyje būtų pateikiami tik reikiamų šalių pavadinimai. Tam spaudžiame mygtuką „*Major tick/label properties*“ ir atsidariusiame lange „*Axis tick and label properties (x axis) ...*“ pasirenkame „*Custom*“ ties „*Axis rule*“, o atsiradusiame laukelyje „*Custom rule*“ įvedame šalių reikšmes 14 ir 19. Po to, šio lango kortelėje „*Labels*“ pažymime varnelę ties „*Use value labels*“ ir spaudžiame „*Accept*“, o grįžę į pirminį langą „*twoway – Twoway graphs*“ kortelėje „*Legend*“ nurodome nerodyti paaiškinimų pasirinkdami „*Hide legend*“ ties „*Legend behavior*“ ir spaudžiame „*OK*“.

Sprendimas su SPSS 15.0	Sprendimas su Stata SE 10.0
	

Sintaksės komandos

<pre>GRAPH /ERRORBAR(CI 95) = pasit_ek BY salis .</pre>	<pre>twoway (scatter pasit_ek_mean salis, msymbol(smcircle)) (rcap pasit_ek_mean_ar pasit_ek_mean_vr salis, msize(huge)), yttitle(Pasitenkinimas savo šalies ekonomika) xlabel(14 19, valuelabel) legend(off)</pre>
---	---

Gauname tokius rezultatus **rezultatų išvesties lange**:

	
---	--

Taigi kaip ir tikėtasi, neturtingos šalies – Lenkijos – gyventojai yra gerokai mažiau patenkinti ekonomikos veikimu savo šalyje, nei turtingos šalies – Airijos. Verta pastebėti ir tai, kad intervaliniai vidurkio įverčiai yra gana siauri, tai lemia santykinai dideli imčių dydžiai (Airijoje apklausti – 2286 gyventojai, o Lenkijoje – 1716) ir santykinai nedidelis pasiklovimo lygmuo – 95 proc. Pakeitę pasiklovimo lygmenį į 99 proc. gautume šiek tiek platesnius intervalus.

3.2. Populiacijos proporcijos (tikimybės) pasikliautinis intervalas

Toliau aptarsime tai, kaip sudaryti **tam tikros kokybinės populiacijos reikšmės proporcijos pasikliautinąjį intervalą**. Kadangi proporcija paprastai skaičiuojama kokybinių kintamųjų reikšmių atveju, tai mūsų tiriamas masinio reiškinių požymis turi būti išmatuotas⁴¹ pavadinimų skalėje⁴². Taigi šiuo atveju mūsų **tikslas** – sužinoti, kokia yra mus dominančio tam tikros kokybinio masinio reiškinių požymio reikšmės proporcija populiacijoje. Kaip jau minėta, kadangi netirsime viso masinio reiškinių, o tik iš jo atrinktos imties duomenis, tai **tikrosios reikšmės nesužinosime**, tačiau statistikos mokslo pagalba **su tam tikra garantija (paprastai 95 proc.) nurodysime intervalą**, kuriame ši reikšmė yra.

Kad iliustruotume tai, kaip sudaryti dvireikšmio kintamojo proporcijos pasikliautinąjį intervalą, panagrinėsime **europiečių politinį anagažuotumą**. Buvo klausta, ar respondentai įvairiose Europos šalyse yra partijos nariai ar ne (atsakymų variantai: 0 – ne, 1 – taip). Kaip matome, atsakymai į klausimą yra išmatuoti **dvireikšmėje pavadinimų skalėje**. Bandysime nustatyti, **kokia proporcija europiečių priklauso partijoms** (kokia jų dalis atsakė į klausimą pasirinkdami atsakymą 1 – taip).

Bendroji tam tikros kokybinio masinio reiškinių požymio reikšmės **proporcijos intervalinio įverčio sudarymo logika** yra tokia⁴³:

Pasirenkame **pasiklovimo lygmenį $1 - \alpha$** ⁴⁴, t. y. nusistatome, kiek norime būti garantuoti savo išvados tikrumu (**paprastai 95 proc.**);

Suskaičiuojame **taškinį** ieškomos tam tikros kokybinio masinio reiškinių požymio reikšmės proporcijos **įvertį** pagal atrinktos imties duomenis;

Parenkame tinkamą statistiką⁴⁵, kurios pagrindu bus nustatytos intervalinio įverčio ribos, pagal tiriamo masinio reiškinių savybes (reikšmių skirstinio pobūdį), nežinomų to masinio reiškinių parametrų skaičių ir imties dydį;

Radę reikiamus (paprastai $\alpha/2$ ir $1-\alpha/2$) pasirinktos statistikos teorinio skirstinio kvantilius, **suskaičiuojame intervalinį įvertį**.

Pavyzdys, kaip randamas intervalinis proporcijos įvertis su SPSS ir Stata, yra pateiktas 3.3 lentelėje.

3.3 lentelė. Dvireikšmio masinio reiškinių požymio vienos reikšmės proporcijos pasikliautinąjo intervalo radimas.

Sprendimas su SPSS 15.0	Sprendimas su Stata SE 10.0
Parengiamieji veiksmai	
Nėra	

⁴¹ Tiksliau, „turi būti iš principo matuojamas“, t. y. tas reiškinys turi būti suvokiamas kaip kokybinis savo esme, todėl galimas išmatuoti vardų skalėje.

⁴² Be to, čia nagrinėjamas tik bazinis dvireikšmių kintamųjų atvejis.

⁴³ Išsamiau apie tai kalbama antroje šio studijų paketo dalyje, 3 skyriuje.

⁴⁴ α yra reikšmingumo lygmuo. Čia vėlgi vertėtų priminti ir tai, kad pasiklovimo lygmuo, imties dydis ir intervalo tikslumas yra susiję dydžiai: kuo didesnis pasiklovimo lygmuo, tuo mažesnis intervalo tikslumas (intervalas platesnis); kuo didesnė imtis, tuo didesnis intervalo tikslumas (intervalas siauresnis); kuo didesnis intervalo tikslumas (intervalas siauresnis), tuo mažesnis pasiklovimo lygmuo.

⁴⁵ Skaičiuojant dvireikšmio kintamojo proporcijos intervalinius įverčius, kai imtis yra pakankamai didelė, naudojama Z statistika.

Sintaksės komandos

Nėra

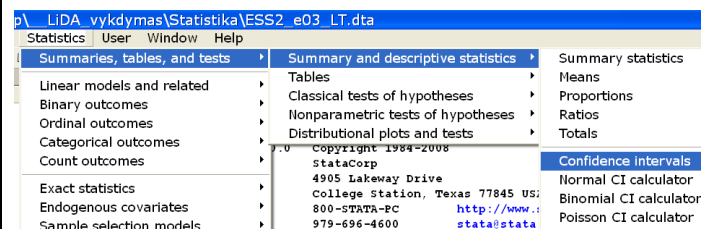
Analizės veiksmas

SPSS programa tiesiogiai neskaiciuoja proporcijų pasikliautinių intervalų, tačiau juos galima apskaičiuoti patiemis su SPSS pagalba. Pirmiausia reikia apskaičiuoti reikšmių proporcijas, dažnius ir tinkamų reikšmių dažnį.

Meniu punktai: Analyze → Descriptive Statistics → Frequencies...

Meniu punktai: Statistics → Summaries, tables, and tests → Summary and descriptive statistics → Confidence intervals

Partijos narys,ė					
		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	Ne	44814	94,3	94,7	94,7
	Taip	2502	5,26327	5,28785	100,0
	Total	47316	99,5	100,0	
Missing	Atsisakė nurodyti	46	,1		
	Nežinojo	114	,2		
	Neatsakė	61	,1		
	Total	221	,5		
Total		47537	100,0		



Po to, gautas reikšmes reikia įstatyti į formulę:

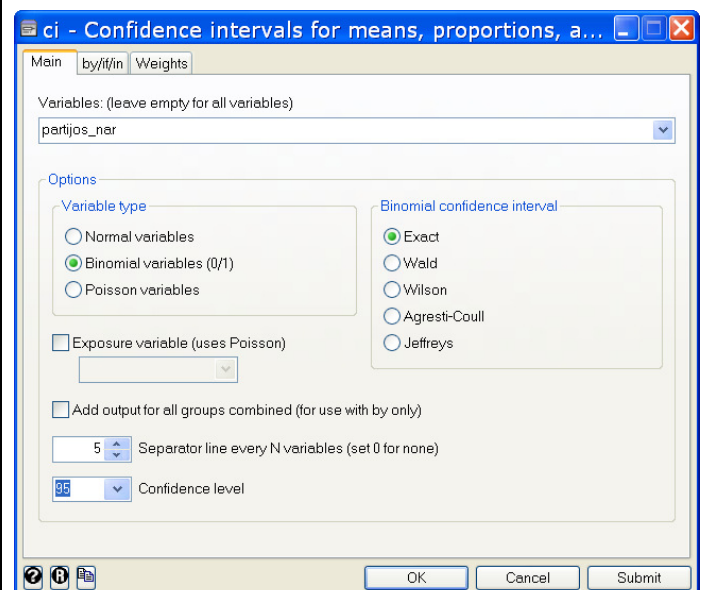
$$PI_{1-\alpha}(p) = \left(\hat{p} - z_{1-\frac{\alpha}{2}} \sqrt{\frac{\hat{p} \cdot (1-\hat{p})}{n}} ; \hat{p} + z_{\frac{\alpha}{2}} \sqrt{\frac{\hat{p} \cdot (1-\hat{p})}{n}} \right)$$

(Pasikliovimo lygmuo: 0,95)

Atsidariusiame lange „*ci – Confidence intervals for means, proportions, and counts*“ parenkame analizės kintamąjį **partijos_nar** ties „*Variables:...*“, nurodome, kad tai dvireikšmis kintamasis pasirinkdami „*Binomial variables (0/1)*“ ties „*Variable type*“, ir spaudžiame „*OK*“.

$$PI_{0,95}(p) = \left(\frac{2502}{47316} - 1,959 \sqrt{\frac{\frac{2502}{47316} * (1 - \frac{2502}{47316})}{47316}} ; \frac{2502}{47316} + 1,959 \sqrt{\frac{\frac{2502}{47316} * (1 - \frac{2502}{47316})}{47316}} \right)$$

$$PI_{0,95}(p) = (0,05288 - 1,959 \sqrt{\frac{0,05288 * (1 - 0,05288)}{47316}} ; 0,05288 + 1,959 \sqrt{\frac{0,05288 * (1 - 0,05288)}{47316}})$$



Sprendimas su SPSS 15.0	Sprendimas su Stata SE 10.0
$PI_{0,95}(p) = (0,05288 - 1,959\sqrt{\frac{0,05008}{47316}};$ $0,05288 + 1,959\sqrt{\frac{0,05008}{47316}})$ $PI_{0,95}(p) = (0,05288 - 0,00202; 0,05288 + 0,00202)$	

Sintaksės komandos

Nėra.	ci partijos_nar, binomial												
Gauname tokį rezultatą	Gauname tokius rezultatus rezultatų išvesties lange:												
$PI_{0,95}(p) = (0,050866; 0,054895)^{46}$	<table><tr><th>Variable</th><th>Obs</th><th>Mean</th><th>Std. Err.</th><th colspan="2">— Binomial Exact — [95% Conf. Interval]</th></tr><tr><td>partijos_nar</td><td>47316</td><td>.0528785</td><td>.0010288</td><td>.050879</td><td>.0549331</td></tr></table>	Variable	Obs	Mean	Std. Err.	— Binomial Exact — [95% Conf. Interval]		partijos_nar	47316	.0528785	.0010288	.050879	.0549331
Variable	Obs	Mean	Std. Err.	— Binomial Exact — [95% Conf. Interval]									
partijos_nar	47316	.0528785	.0010288	.050879	.0549331								

Statistinė išvada

Su 95 proc. garantija galime teigti, kad nuo 5,1 iki 5,5 proc. tirtų europiečių priklauso kokiai nors partijai.

Taigi atlikę statistinį tyrimą (atrinkę tinkamą dvireikšmiu masinio reiškimo požymiui tirti imtį, ir suskaičiavę vienos to masinio reiškimo požymio reikšmės proporcijos intervalinį įvertį) su 95 proc. garantija galime teigti, kad tik šiek tiek daugiau nei kas dvidešimtas europietis priklauso partijai, taigi politinis angažuotumas Europoje nėra didelis. Žinoma, tokia **apibendrinta išvada mažai pasako apie skirtingų šalių** gyventojų politinį angažuotumą. Empiriniai tyrimai rodo, kad Skandinavijos šalyse politinis angažuotumas yra žymiai didesnis nei Rytų ir Vidurio Europos šalyse, todėl įdomu būtų **palyginti⁴⁷ Rytų Europos šalies Estijos ir Skandinavijos atstovės Norvegijos** gyventojų priklausymo partijoms proporcijas. Kaip jau ne kartą buvo minėta, visada informatyviau palyginamuosius duomenis atvaizduoti grafiškai, taigi nuo šios taisyklės nenukrypsime ir šiuo atveju. Pavyzdys, kaip tai padaryti su Stata⁴⁸, yra pateiktas 3.4 lentelėje.

3.4 lentelė. Dvireikšmio masinio reiškimo požymio vienos reikšmės proporcijos tam tikrose grupėse pasikliautiniųjų intervalų radimas.

Sprendimas su Stata SE 10.0
Parengiamieji veiksmai
Prieš sudarant pasikliautiniųjų intervalų grafikus tam tikroms konkrečioms grupėms Stata pakete reikia transformuoti originalią duomenų laikmeną į proporcijų ir jų pasikliautiniųjų intervalų laikmeną. Tą

⁴⁶ Rezultatų skirtumai neturėtų stebinti, nes skaičiuojant „rankomis“ dažnai tenka apvalinti skaičius, todėl nuo tam tikro tikslumo lygmens rezultatai ima skirtis. Tačiau kaip matome, tai labai menkas skirtumas (prasidedantis tik nuo penkto skaičiaus po kablelio!).

⁴⁷ Tikrasis dvireikšmio kintamojo proporcijų pogrupiuose palyginimo algoritmas šioje medžiagoje nėra patiekiamas. Todėl būtina atkreipti dėmesį į pastabą, kad čia aptariami tik preliminarūs rezultatai ir grafinio vaizdavimo būdai.

⁴⁸ Kadangi SPSS programinis paketas neskaiciuoja dvireikšmių kintamųjų proporcijų pasikliautiniųjų intervalų, o kaip tai padaryti „rankiniu būdu“ jau buvo pademonstruota, tai to nebekartosime.

Sprendimas su Stata SE 10.0

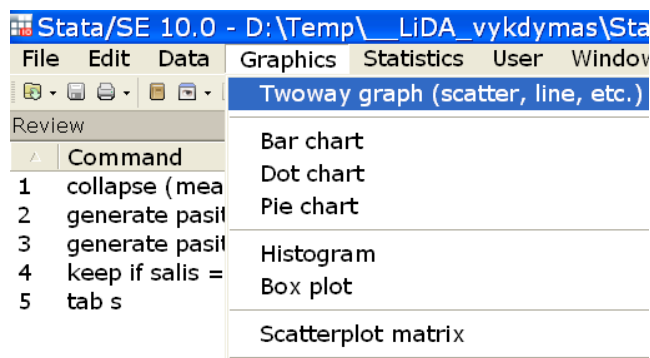
galima padaryti per meniu, tačiau paprasčiau – sintaksės komandų pagalba.

Sintaksės komandos

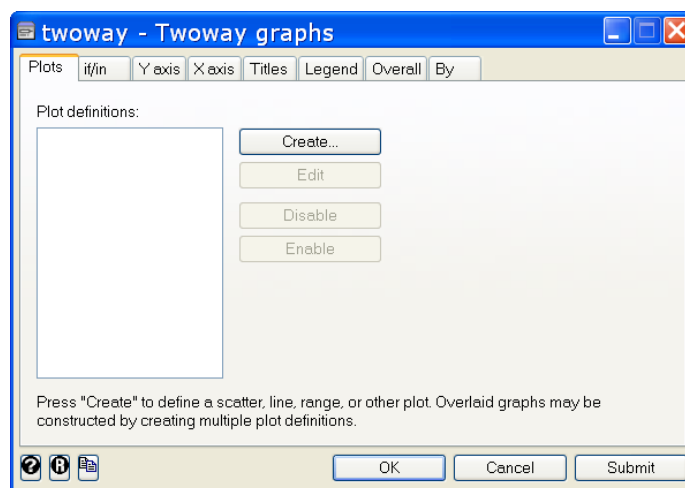
```
collapse (mean) partijos_nar_prop = partijos_nar (sebinomial) partijos_nar_seb =  
partijos_nar, by(salis)  
generate partijos_nar_prop_ar = partijos_nar_prop - invnormal(0.025) *  
partijos_nar_seb  
generate partijos_nar_prop_vr = partijos_nar_prop + invnormal(0.025) *  
partijos_nar_seb  
keep if salis == 7 | salis == 18
```

Analizės veiksmas

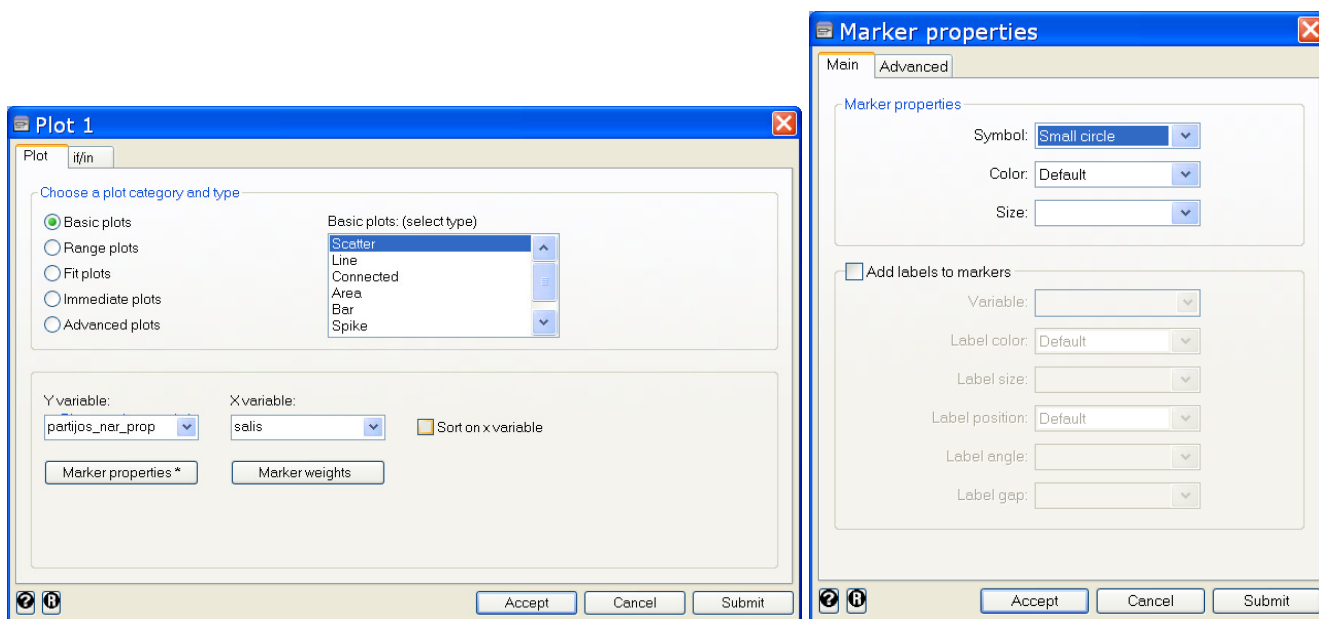
Meniu punktai: Graphics → Twoway graph (scatter, line, etc.)



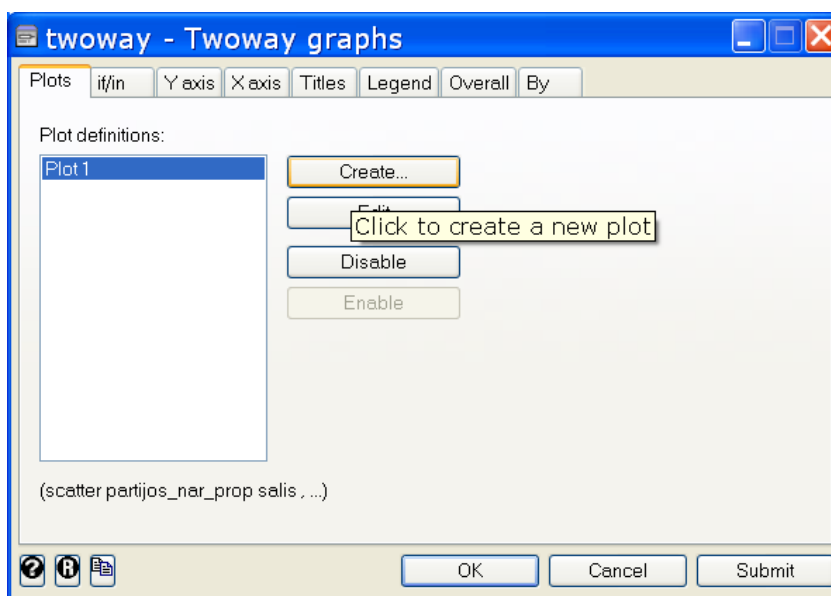
Atsidariusiame lange „*twoway – Twoway graphs*“ spaudžiame mygtuką „*Create...*“, kad sukurti pirmąją grafiko dalį.



Atsidarius langui „*Plot 1*“ sukuriame pirmąją grafiko dalį – proporcijų taškinį grafiką – pasirinkdami „*Scatter*“ ties „*Basic plots*“, įkeldami **partijos_nar_prop** kintamąjį ties „*Y variable*“ ir **salis** – ties „*X variable*“ bei parinkdami mažą apskritimą kaip taško grafike simbolį (spaudžiame mygtuką „*Marker properties*“ ir pasirenkame „*Small circle*“ ties „*Symbol*“ bei paspaudžiame „*Accept*“). Po to, spaudžiame „*Accept*“.

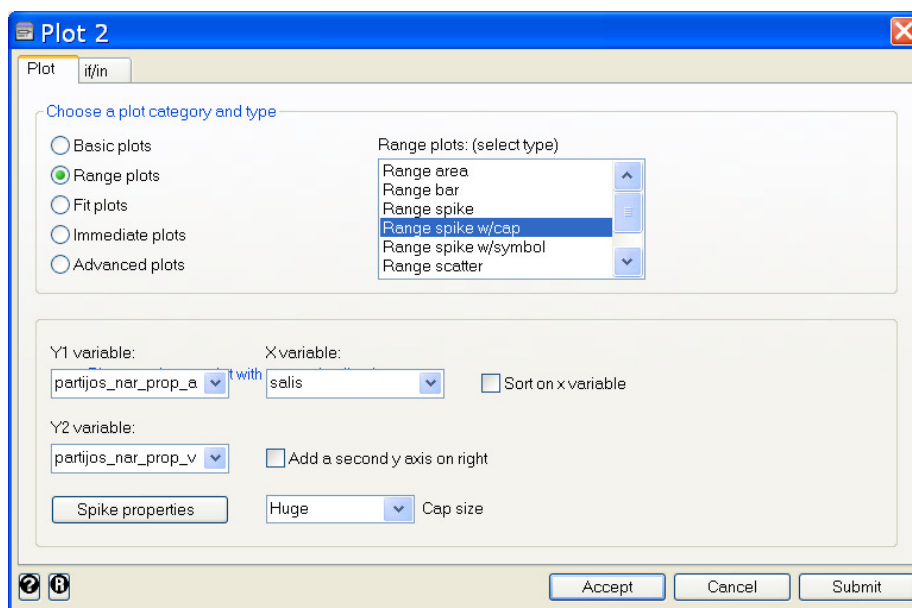


Grįžę į „*twoway – Twoway graphs*“ langą dar kartą spaudžiame mygtuką „*Create...*“, kad sukurti antrąją grafiko dalį.

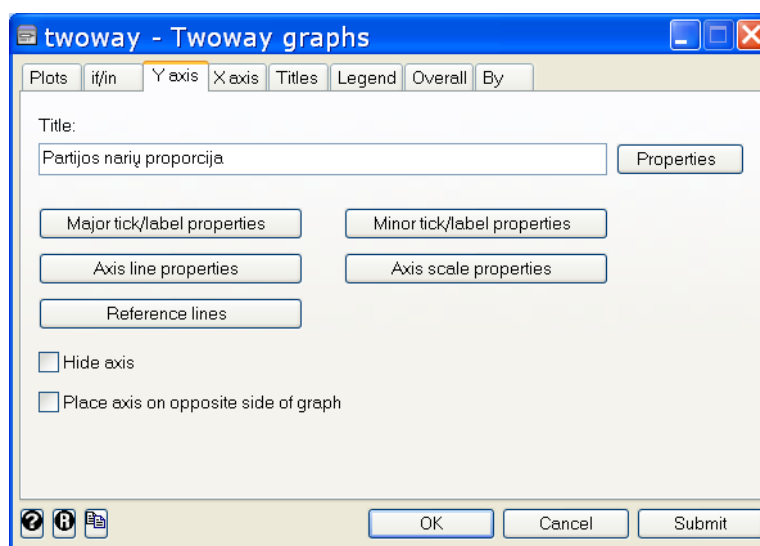


Atsidarius langui „*Plot 2*“ sukuriame antrąją grafiko dalį – proporcijos pasikliautinojo intervalo ribų grafiką – pasirinkdami „*Range plots*“ ir „*Range spike w/ caps*“ ties „*Basic plots*“, įkeldami **partijos_nar_prop_ar** kintamąjį ties „*Y1 variable*“, **partijos_nar_prop_vr** – ties „*Y2 variable*“ ir **salis** – ties „*X variable*“ bei parinkdami didelius ribų žymeklius „*Huge*“ ties „*Cap size*“. Po to, spaudžiame „*Accept*“.

Sprendimas su Stata SE 10.0

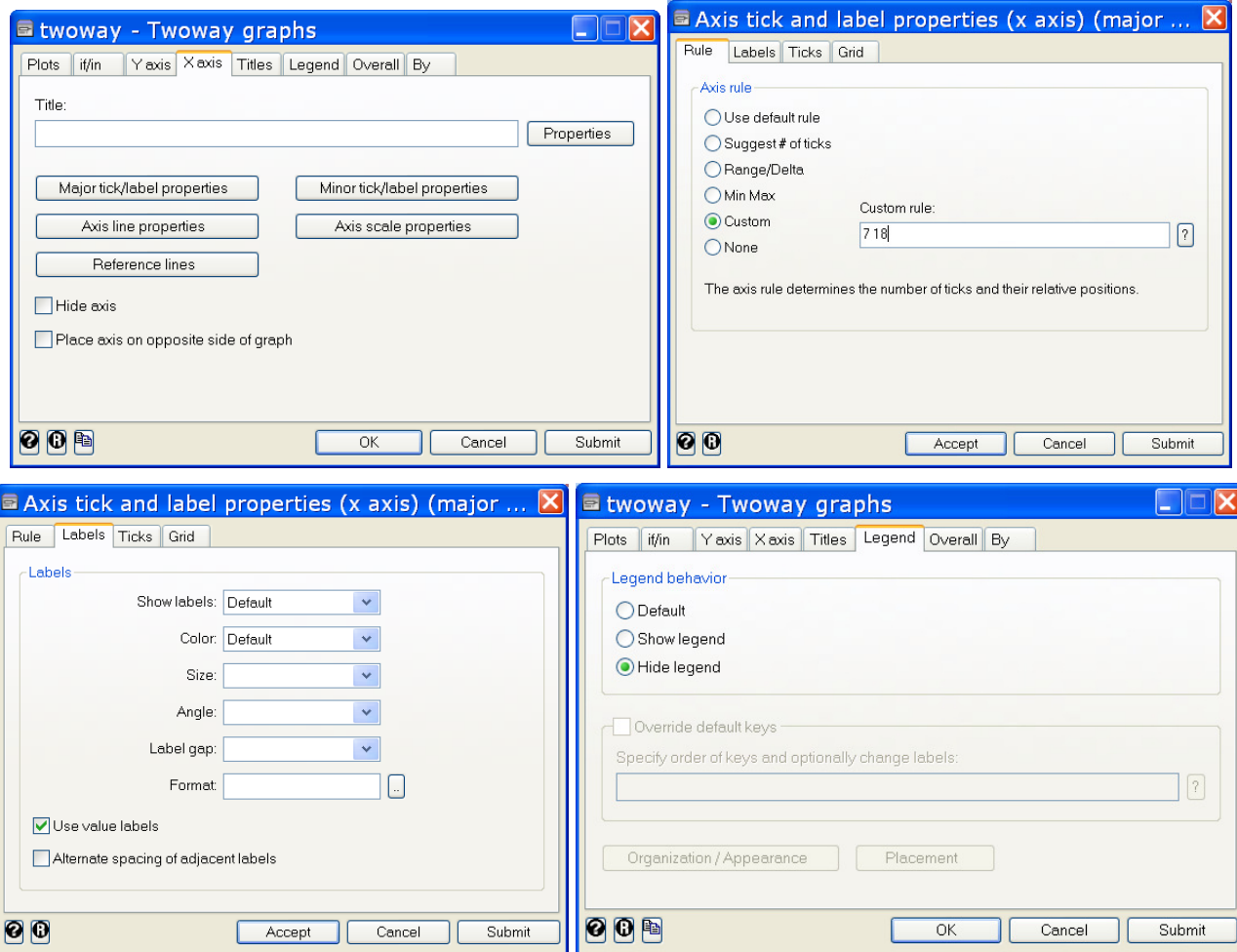


Grįžę į „*twoway – Twoway graphs*“ langą, nurodome Y ašies pavadinimą kortelėje „*Y axis*“ ties „*Title*“ įrašydami Pasitenkinimas savo šalies ekonomika).



Po to, kortelėje „*X axis*“ nurodome, kad X ašyje būtų pateikiami tik reikiamų šalių pavadinimai. Tam spaudžiame mygtuką „*Major tick/label properties*“ ir atsidariusiame lange „*Axis tick and label properties (x axis) ...*“ pasirenkame „*Custom*“ ties „*Axis rule*“, o atsiradusiame laukelyje „*Custom rule*“ įvedame šalių reikšmes 7 ir 18. Po to, šio lango kortelėje „*Labels*“ pažymime varnelę ties „*Use value labels*“ ir spaudžiame „*Accept*“, o grįžę į pirminį langą „*twoway – Twoway graphs*“ kortelėje „*Legend*“ nurodome nerodyti paaiškinimų pasirinkdami „*Hide legend*“ ties „*Legend behavior*“ ir spaudžiame „*OK*“.

Sprendimas su Stata SE 10.0



Sintaksės komandos

```
twoway (scatter partijos_nar_prop salis, msymbol(smcircle)) (rcap  
partijos_nar_prop_ar partijos_nar_prop_vr salis, msize(huge)), ytitle(Partijos narių  
proporcija) xlabel(7 18, valuelabel) legend(off)
```

Gauname tokius rezultatus **rezultatų išvesties lange**:



Taigi kaip ir tikėtasi, gerokai didesnė dalis Skandinavijos atstovės – Norvegijos – gyventojų priklauso savo šalies partijoms, nei Rytų Europos šalies – Estijos. Tą parodo nemažas atotrūkis (beveik 3 proc.) tarp pasikliautinių intervalų „uodegų“.

4. HIPOTEZIŲ TIKRINIMAS

Šiame skyriuje **tęsiamas išvadų statistikos** aptarimas ir pereinama prie statistinių hipotezių tikrinimo pavyzdžių pristatymo. Kaip minėta, išvadų statistika buvo sukurta ir išstbulinta dėl to, kad praktiškai labai sunku nagrinėti ištikus masinius reiškinius, todėl paprastai iš jų atrenkamos imtys ir pagal jas daromos išvados apie tų masinių reiškinių požymių apibendrintas charakteristikas. Jei populiacijos požymių apibendrintų charakteristikų intervalinių įverčių radimo logika labiau tinka nagrinėjant atskiras populiacijos požymių charakteristikas, **tai hipotezių tikrinimo logika yra skirta palyginti skirtingų populiacijų požymių apibendrintas charakteristikas.**

Tradiciškai sąvoka **hipotezė** yra suvokiama kaip teoriškai (moksliškai ir pan.) pagrįstas spėjimas apie tam tikro reiškinio būvį ar jo sąsajas su kitais reiškiniais. Statistikoje ši sąvoka turi konkretesnę ir siauresnę prasmę. Taigi **statistinė hipotezė** yra bet kokia prielaida apie masinio reiškinio požymio reikšmių pasiskirstymo pobūdį, jo apibendrintas charakteristikas ir tų apibendrintų charakteristikų palyginimą su kitų masinių reiškinių požymių apibendrintomis charakteristikomis. Paprasčiau tariant, tai **prielaidos apie tiriamų masinių reiškinių požymių parametrus arba jų skirstinių pobūdį ir tų parametrų arba skirstinių lygybę kitų masinių reiškinių požymių parametrams arba skirstiniams.** Dažniausiai formuluojamos tokios statistinės hipotezės:

populiacijos požymių **vidurkio ir standartinio nuokrypio** lygybės tam tikrai reikšmei (pavyzdžiui, Lietuvos gyventojų amžiaus vidurkis yra 45 metai);

skirtingų populiacijų tų pačių požymių **vidurkių ir dispersijų** lygybės (pavyzdžiui, Lietuvos moterų ir vyrų amžiaus vidurkis vienodas);

tos pačios populiacijos tų pačių požymių **vidurkių ir dispersijų** lygybės prieš ir po eksperimento (pavyzdžiui, studentų politologijos testo rezultatai yra aukštesni po semestro studijų);

vieno populiacijos požymio parametro priklausomybės nuo kito tos pačios populiacijos požymio parametro arba **koreliacijos koeficiento** lygybės tam tikrai reikšmei (pavyzdžiui, studentų pažymių lygis priklauso nuo mokymuisi skirtų valandų skaičiaus);

skirtingų populiacijų tų pačių požymių reikšmių **skirstinių** lygybės (pavyzdžiui, Lietuvos moterų ir vyrų amžiaus reikšmių skirstiniai skiriasi).

Statistinių hipotezių tikrinimo logika reikalauja **pirminę (teorinę) hipotezę suskaidyti į dvi statistines – nulinę ir alternatyviąją.** Taip daroma dėl to, kad būtų galima suformuluoti konkrečią išvadą, t. y. jei remiantis statistinio tyrimo metu surinktais duomenimis atmetama viena hipotezė, kaip teisinga priimama su ja konkuruojančioji. Tiesa, čia reikia prisiminti, kad kadangi statistinės hipotezės yra formuluojamos apie populiacijos požymių charakteristikas, o tikrinamos pagal imties duomenis, tai jos priimamos arba atmetamos tik su tam tikra tikimybe, o ne absoliučiai. Taigi visada lieka tikimybė, kad padaryta išvada bus klaidinga⁴⁹.

Paprastai **tikrinamoji hipotezė yra nulinė (žymima H_0).** Joje teigiama, kad nėra esminių skirtumų tarp lyginamų objektų (vienas rodiklis yra lygus (=) kitam). Konkuruojanti su nuline hipoteze yra **alternatyvioji (žymima H_A).** Joje teigiama, kad skirtumas tarp lyginamų objektų yra (vienas rodiklis nelygus ($\neq$) kitam, arba mažesnis (<), arba didesnis (>) už jį). Jei remiantis statistinio tyrimo metu surinktais duomenimis bus atmesta nulinė hipotezė, kaip teisinga bus priimama alternatyvioji. Taigi statistinių hipotezių tikrinimo **bendrasis algoritmas** yra toks:

⁴⁹ Apie tai, kokias klaidas galima padaryti tikrinant statistines hipotezes, išsamiai buvo išdėstyta antroje studijų paketo dalyje, 4.2 poskyryje.

Teorinė hipotezė apie populiacijos (ar kelių populiacijų) požymių charakteristikas ir jų lyginimą **performuluojama į statistinę nulinę⁵⁰ ir alternatyviąją hipotezę.**

Pasirenkamas **reikšmingumo lygmuo α** , t. y. galimybė padaryti pirmos rūšies klaidą (atmesti H_0 , kai ji teisinga), kuris paprastai esti **5 % (0,05) arba 1 % (0,01).**

Parenkamas **statistinis kriterijus, tinkamas nulinei hipotezei** apie nagrinėjamą masinio reiškinių požymio charakteristiką tikrinti (pavyzdžiui, jei tikriname populiacijos požymio reikšmių vidurkio skirtumą skirtingose populiacijos grupėse, tai statistinis kriterijus yra *Stjudento t statistika*)⁵¹.

Pagal pasirinktą reikšmingumo lygmenį α , statistinio kriterijaus tipą ir alternatyviąją hipotezę randamos to statistinio kriterijaus (t. y. nulinei hipotezei tikrinti naudojamos statistikos teorinio skirstinio) **kritinė(s) reikšmė(s)**⁵² (kuri(os) dar kartais priklauso ir nuo kitų parametrų, pavyzdžiui, imties didumo, laisvės laipsnių skaičiaus) **ir nulinės hipotezės priėmimo (pasikliautinoji) bei atmetimo (kritinė) sritys.**

Surenkami duomenys apie tikimybiškai atrinktą imtį iš tiriamos populiacijos.

Apskaičiuojama empirinė pasirinkto statistinio kriterijaus (t. y. empirinės skaičiuojamos statistikos) reikšmė pagal imties duomenis ir žiūrima, ar ji(os) patenka į nulinės hipotezės priėmimo ar atmetimo sritį.

Jei empirinė pasirinkto statistinio kriterijaus reikšmė patenka į nulinės hipotezės priėmimo sritį, tai nulinė hipotezė **priimama**, jei patenka į atmetimo sritį – **atmetama (su tam tikru pasiklovimo lygmeniu (arba garantija), kuris lygus $1 - \alpha$).**

Bendrasis statistinių hipotezių algoritmas gana sudėtingas, todėl praktiškai jį **lengviau suprasti vadovaujantis pasikliautinųjų intervalų** sudarymo algoritmu⁵³. Paprastai tariant, atliekant statistinį tyrimą, pirmiausia, ieškomai populiacijos požymio charakteristikai apskaičiuojamas teorinis pasikliautinis intervalas, koks jis turėtų būti pagal būsimos imties duomenis ir reikšmingumo lygmenį (paprastai 5 % (0,05) arba 1 % (0,01)). Tada atrenkama imtis, suskaičiuojama ieškomo požymio charakteristika pagal imtį ir žiūrima, ar ji patenka į teorinį pasikliautinąjį intervalą. Jei patenka, tai priimama vienokia išvada (su tam tikra garantija, paprastai 95 % arba 99 %), jei nepatenka – kitokia (su tam tikra garantija, paprastai 95 % arba 99 %). Išvados pobūdis priklauso nuo pirminio suformuluoto uždavinio. Pavyzdžiui, jei tikrinam, ar žmonių pasitikėjimo Prezidentu vidurkis yra didesnis už vidutinę matavimo skalės (nuo 0 iki 10) reikšmę (5), tai sudarę teorinį vidurkio (lygaus penkiems) pasikliautinąjį intervalą pagal būsimos imties dydį ir pasirinktą reikšmingumo lygmenį, žiūrim, ar atrinkus imtį ir apskaičiavus jos pagrindu pasitikėjimo Prezidentu vidurkį, šis patenka į teorinį pasikliautinąjį intervalą arba yra mažesnis, ar jo reikšmė yra už jo ribų ir yra didesnė.

Statistinės analizės programinėje įrangoje statistinių hipotezių tikrinimo algoritmas šiek tiek skiriasi nuo bendrojo. Kad supaprastinti hipotezių tikrinimo procedūrą buvo įvestas **p-reikšmės metodas**. Jis, galima sakyti, bendrąjį statistinių hipotezių tikrinimo algoritmą apverčia aukštyn kojomis. Pirmiausiai, pagal imties duomenis yra suskaičiuojamas ieškoma populiacijos charakteristiką atitinkantis statistinis kriterijus (statistika) ir jo empirinis reikšmingumo lygmuo (vadinamoji p-reikšmė, kuri reiškia tikimybės pagal turimus imties duomenis padaryti pirmos rūšies

⁵⁰ Formaliai žiūrint, ji ir bus iš tikrųjų „tikrinama“.

⁵¹ Čia reikia nepamiršti, kad svarbus ir pats masinio reiškinių pobūdis (kokybinis ar kiekybinis, kokioje matavimo skalėje išmatuotas) bei jo reikšmių skirstinio pobūdis (binominis, normalusis, Puasono ar dar kitoks). Nuo to, koks yra kintamojo reikšmių skirstinys, priklauso ir taikomas statistinis kriterijus.

⁵² Jei alternatyvioji hipotezė yra vienpusė (< arba >), tai kritinė reikšmė esti viena, o jei dvipusė ($\neq$) – kritinės reikšmės yra dvi.

⁵³ Kaip minėta, tam tikra prasme, intervaliniai įverčiai yra išvadų statistikos pagrindas.

klaidą, t. y. atmesti H_0 , kai ji teisinga, didumą). Tada empirinis reikšmingumo lygmuo (p-reikšmė) yra lyginamas su nusistatytuoju reikšmingumo lygmeniu (α -reikšmė, paprastai 0,05 arba 0,01). Sprendimas priimamas taip:

jei $p\text{-reikšmė} \geq \alpha\text{-reikšmė}$, tai priimama (neatmetama) nulinė hipotezė (H_0);
jei $p\text{-reikšmė} < \alpha\text{-reikšmė}$, tai H_0 atmetama ir priimama alternatyvioji hipotezė (H_A).

Pastarasis algoritmas yra universalus daugelyje statistinės analizės programinės įrangos paketų. Tereikia žinoti, kokia nulinė ir alternatyvioji hipotezė yra tikrinama konkrečioje programoje konkrečios statistinės analizės atveju, ir pagal rezultatus galėsime nesunkiai padaryti pagrįstą išvadą. Taigi pateikiant konkrečius hipotezių tikrinimo pavyzdžius visada bus supažindinama su informacija apie tai, kokios būtent formos nulinė ir alternatyvioji hipotezės yra tikrinamos SPSS ir Stata programinių paketų atveju.

Pateikdami statistinių hipotezių tikrinimo pavyzdžius naudosime Europos socialinio tyrimo (ESS) trečiosios bangos (2006 m.) duomenis (SPSS laikmena – „*ESS3_e03.1_LT.sav*“, Stata laikmena – „*ESS3_e03.1_LT.dta*“). Šio tyrimo ir jame tyrinėjamų reiškinių bei kintamųjų aprašas yra pateiktas pirmojoje studijų paketo dalyje. Dažniausiai formuluojamos statistinės hipotezės jau buvo išvardytos, o šiame skyriuje pateiksime pavyzdžius, kaip tikrinti hipotezes apie populiacijos požymio reikšmių vidurkio lygybę tam tikrai reikšmei, populiacijos požymio reikšmių proporcijų lygybę tam tikroms reikšmėms ir skirstinių suderinamumo hipotezes.

4.1. Hipotezių apie populiacijos vidurkį tikrinimas

Pirmiausiai aptarsime tai, kaip patikrinti hipotezę, ar tam tikro kiekybinio masinio reiškinio požymio reikšmių vidurkis yra lygus hipotetiniam, t. y. tyrinėtojo numatomam pagal teoriją ar ankstesnius empirinius rezultatus, vidurkiui. Kadangi vidurkis skaičiuojamas tik kiekybiniais kintamiesiems, tai mūsų tiriamas masinio reiškinio požymis turi būti matuojamas intervalų arba santykių skalėje. Mūsų **tikslas** – sužinoti, ar mus dominančio kiekybinio masinio reiškinio požymio reikšmių vidutinė reikšmė skiriasi nuo hipotetinės. Kadangi netirsime viso masinio reiškinio, o tik iš jo atrinktos imties duomenis, tai **visiškai užtikrintai pasakyti, ar tos reikšmės skiriasi, ar – ne, negalėsime**. Tačiau statistikos mokslo pagalba **pasitelkdami hipotezių tikrinimo algoritmą būsime tam tikru mastu garantuoti (paprastai 95 proc.)** apie to skirtumo buvimą ar nebuvimą.

Kad iliustruotume tai, kaip tikrinamos hipotezės apie populiacijos požymio reikšmių vidurkio lygybę hipotetiniam vidurkiui, panagrinėsime **europiečių⁵⁴ pasitenkinimą savo gyvenimo lygiu**. Buvo klausta, kiek respondentai įvairiose Europos šalyse yra patenkinti savo gyvenimo lygiu (atsakymų variantai: nuo 0 – visiškai nepatenkintas, -a iki 10 – visiškai patenkintas, -a). Kaip matome, klausimas matuoja pasitenkinimą gyvenimo lygiu **tvarkos skalėje**, tačiau pasitenkinimo rangų yra santykinai didelis kiekis (11), todėl tokį kintamąjį **galime laikyti kiekybiniu**. Bandysime patikrinti prielaidą, kad europiečių pasitenkinimas yra didesnis nei vidurinė matavimo skalės reikšmė, t. y. didesnis nei 5 balai.

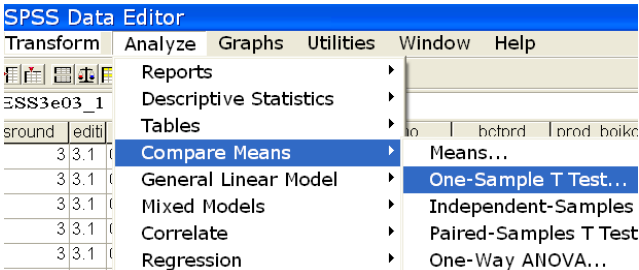
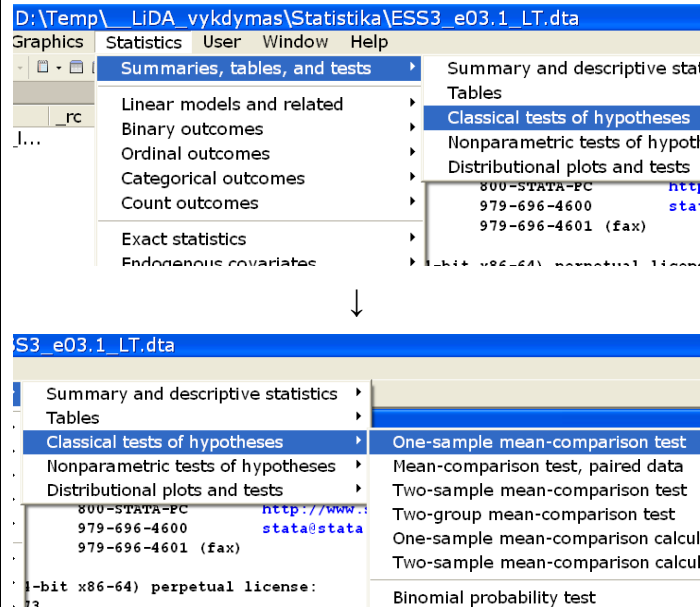
Šios hipotezės tikrinimui naudojamas *Stjudento t* statistinis kriterijus (statistika), o pats testas vadinamas *vienos imties t testu (one sample t test)*. Primename, kad sprendimas, ar europiečių pasitenkinimo savo gyvenimo lygiu vidurkis lygus hipotetiniam priimamas taip:

jei $p\text{-reikšmė} \geq \alpha\text{-reikšmė}$, tai priimama (neatmetama) nulinė hipotezė (H_0);
jei $p\text{-reikšmė} < \alpha\text{-reikšmė}$, tai H_0 atmetama ir priimama alternatyvioji hipotezė (H_A).

⁵⁴ Į ESS trečiąją bangą įtrauktų šalių sąrašas pateiktas pirmoje studijų paketo dalyje (arba žr. http://www.europeansocialsurvey.org/index.php?option=com_content&task=view&id=50&Itemid=84). Taigi terminas „europiečiai“ šiuo atveju apima ne visas Europos šalis, o tik tas, kuriose tyrimas buvo vykdytas 2006 m.

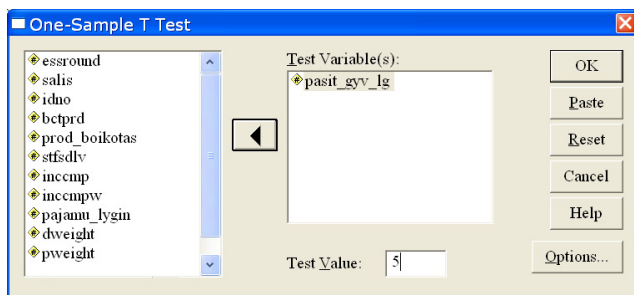
Pavyzdys, kaip tikrinamos hipotezės apie kiekybinio populiacijos požymio reikšmių vidurkio lygybę hipotetiniam vidurkiui su SPSS ir Stata, yra pateiktas 4.1 lentelėje.

4.1 lentelė. Hipotezės apie kiekybinio masinio reiškinio požymio reikšmių vidurkio lygybę hipotetiniam vidurkiui tikrinimas: vienos imties t testas (reikšmingumo lygmuo $\alpha = 0,05$).

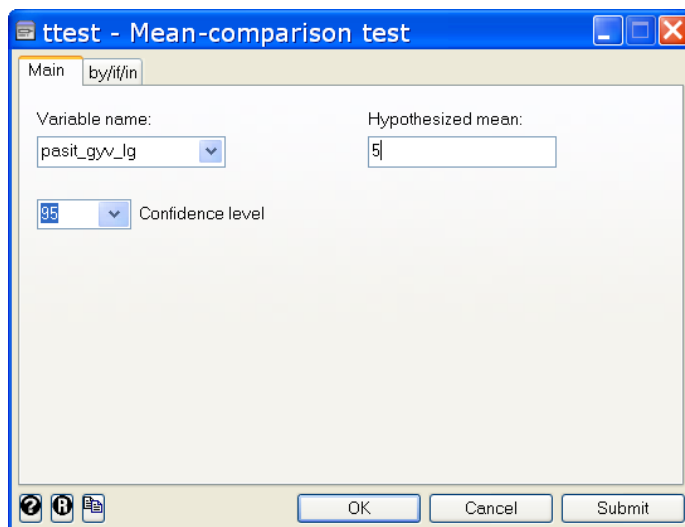
Sprendimas su SPSS 15.0	Sprendimas su Stata SE 10.0
Parengiamieji veiksmai	
Nėra	
Sintaksės komandos	
Nėra	
Tikrinamos hipotezės	
H_0 : populiacijos požymio reikšmių vidurkis – hipotetinis vidurkis = 0 H_A : populiacijos požymio reikšmių vidurkis – hipotetinis vidurkis $\neq 0$	H_0 : populiacijos požymio reikšmių vidurkis = 5 H_{A1} : populiacijos požymio reikšmių vidurkis < 5 H_{A2} : populiacijos požymio reikšmių vidurkis $\neq$ ⁵⁵ 5 H_{A3} : populiacijos požymio reikšmių vidurkis > 5
Analizės veiksmai	
Meniu punktai: Analyze → Descriptive Statistics → Explore...	Meniu punktai: Statistics → Classical tests of hypotheses → One-sample mean-comparison test
	
Atsidariusiame lange „ <i>One-Sample T Test</i> “ į dešinėje pusėje esantį langelį „ <i>Test Variable(s)</i> “ įkeliam kintamąjį pasit_gyv_lg , nurodome hipotetinio vidurkio reikšmę ties „ <i>Test Value</i> “ įvesdami skaičių 5 ir spaudžiame „OK“.	Atsidariusiame lange „ <i>ttest – Mean-comparison test</i> “ parenkame analizės kintamąjį pasit_gyv_lg ties „ <i>Variable name</i> “, nurodome hipotetinio vidurkio reikšmę ties „ <i>Hypothesized mean</i> “ įvesdami skaičių 5 ir spaudžiame „OK“.

⁵⁵ Stata pakete nelygybės ženklas žymimas „!+“.

Sprendimas su SPSS 15.0



Sprendimas su Stata SE 10.0



Sintaksės komandos

T-TEST /TESTVAL = 5 /MISSING = ANALYSIS
/VARIABLES = pasit_gyv_lg /CRITERIA =
CI(.95) .

ttest pasit_gyv_lg == 5

Gauname tokius rezultatus rezultatų išvesties lange:

One-Sample Statistics				
	N	Mean	Std. Deviation	Std. Error Mean
Pasitenkinimas savo gyvenimo lygiu	42754	6,46	2,400	,012

One-Sample Test						
	Test Value = 5					
	t	df	Sig. (2-tailed)	Mean Difference	95% Confidence Interval of the Difference	
Pasitenkinimas savo gyvenimo lygiu	126,209	42753	,000	1,465	Lower 1,44	Upper 1,49

One-sample t test						
Variable	Obs	Mean	Std. Err.	Std. Dev.	[95% Conf. Interval]	
pasit_g	42754	6.464682	.0116052	2.39961	6.441935	6.487428
mean = mean(pasit_gyv_lg) t = 126.2092						
Ho: mean = 5 degrees of freedom = 42753						
Ha: mean < 5 Ha: mean != 5 Ha: mean > 5						
Pr(T < t) = 1.0000 Pr(T > t) = 0.0000 Pr(T > t) = 0.0000						

Statistinė išvada

Kaip matome, SPSS paketu galima patikrinti tik vieną alternatyviają hipotezę, kad empirinis ir hipotetinis vidurkis nelygūs vienas kitam, taigi tam kad patikrinti, ar empirinis vidurkis yra didesnis už hipotetinį, reikės atlikti papildomų skaičiavimų. Pagal SPSS analizės rezultatus galime tik pasakyti, kad su 95 proc. garantija empirinis vidurkis (6,46) statistiškai reikšmingai skiriasi nuo 5 ($t = 126,209$, $df = 42753$, p (Sig. (2-tailed)) = $0,000 < \alpha = 0,05$).

Kad patikrinti hipotezę, ar empirinis vidurkis

Kaip matome, Stata pakete yra tikrinamos visos alternatyviosios hipotezės, todėl iš karto galime pasakyti, kad su 95 proc. garantija empirinis vidurkis (6,464682) yra statistiškai reikšmingai didesnis už 5 ($t = 126,2092$, $df = 42753$, p ($Pr(T > t)$) = $0,000 < \alpha = 0,05$).

⁵⁶ t^* šiuo atveju reiškia statistikos t reikšmę, kuri pateikiama SPSS rezultatų išvesties lange.

⁵⁷ p^* šiuo atveju reiškia p -reikšmę, kuri pateikiama SPSS rezultatų išvesties lange.

Sprendimas su SPSS 15.0	Sprendimas su Stata SE 10.0
didesnis už hipotetinį, reikia perskaičiuoti p-reikšmę ir ją lyginti su nusistatytu reikšmingumo lygmeniu α: - jeigu alternatyvioji hipotezė teigia, kad empirinis vidurkis $<$ hipotetinį ir $t^* \leq 0$, tai p-reikšmė = $p^*/2$; - jeigu alternatyvioji hipotezė teigia, kad empirinis vidurkis $<$ hipotetinį ir $t^* > 0$, tai p-reikšmė = $1-p^*/2$; - jeigu alternatyvioji hipotezė teigia, kad empirinis vidurkis $>$ hipotetinį ir $t^* > 0$, tai p-reikšmė = $p^*/2$; - jeigu alternatyvioji hipotezė teigia, kad empirinis vidurkis $>$ hipotetinį ir $t^* \leq 0$, tai p-reikšmė = $1-p^*/2$. Mūsų atveju yra c situacija (norime išsiaiškinti, ar empirinis europiečių pasitenkinimo savo gyvenimo lygiu vidurkis (6,46) yra statistiškai reikšmingai didesnis nei hipotetinis (5), ir apskaičiuotoji t^* reikšmė = 126,209 $>$ 0), taigi turime tikrinti, ar $p^*/2 \geq \alpha$ ar $p^*/2 < \alpha$. Jei $p^*/2 \geq \alpha$ H_0 priimama, jei $p^*/2 < \alpha$ H_0 atmetama. Taigi galime suformuluoti tokią galutinę išvadą: su 95 proc. garantija galime teigti, kad empirinis vidurkis (6,46) yra statistiškai reikšmingai didesnis už 5 ($t = 126,209$, $df = 42753$, p^* (Sig. (2-tailed)) / 2 = 0,000 $<$ $\alpha = 0,05$).	

Taigi atlikę statistinį tyrimą ir patikrinę suformuluotą hipotezę vadovaudamiesi statistinių hipotezių tikrinimo logika su 95 proc. garantija galime teigti, kad vidutinis europiečių pasitenkinimas savo gyvenimo lygiu yra didesnis nei vidurinė intervalo reikšmė 5. Tiesa, ši išvada padaryta kiek skubotai, nes statistinių kriterijų (statistikų) taikymas susijęs su tam tikrų sąlygų tenkinimu. Nagrinėtu atveju buvo taikytas vienas iš populiariausių statistinių kriterijų (statistikų) – **Stjudento t**. Vienas iš pagrindinių šio **kriterijaus taikymo sąlygų** reikalauja, kad populiacijos, iš kurios atrinkta imtis, požymio reikšmių skirstinys būtų suderinamas su normaliuoju skirstiniu. Būtent šią prielaidą ir patikrinsime pateikdami suderinamumo hipotezių tikrinimo pavyzdžius sekančiame poskyryje.

4.2. Suderinamumo hipotezių tikrinimas

Apskritai kalbant, suderinamumo hipotezės tikrinamos, kai lyginami skirstiniai: keli empiriniai ar empirinis su teoriniu. Jos naudojamos, kai populiacijos požymio reikšmių skirstiniai nežinomi, todėl dar vadinamos neparametrinėmis (statistinės hipotezės keliamos apie skirstinio formą, o ne apie skirstinio apibendrintas charakteristikas – parametrus). Egzistuoja daugybė būdų ir statistinių kriterijų (statistikų), kaip patikrinti skirtingas suderinamumo hipotezes, nes sugalvota daug skirtingų matų, kaip apskaičiuoti atstumą tarp skirstinių (teorinio ir empirinio, ar kelių

empirinių). Populiariausias iš jų yra χ^2 . Jis gali būti naudojamas ir skirstinio suderinamumui su normaliuoju skirstiniu tikrinimui, tačiau dažniau tokiu atveju taikomi kiti kriterijai⁵⁸. Vienas iš tokių yra Kolmogorovo ir Smirnovo Z ⁵⁹. Čia nesigilinsime į jo naudojimo teorinį algoritmą, užteks pasakyti, kad nustatyti atstumą tarp empirinio (populiacijos) ir teorinio skirstinio rusų matematikai Kolmogorovas ir Smirnovas pasiūlė didžiausio skirtumo formulę⁶⁰.

Kaip minėta, kad iliustruotume tai, kaip tikrinamos empirinio kiekybinio kintamojo skirstinio suderinamumo su normaliuoju skirstiniu hipotezės, patikrinsime, ar **europiečių pasitenkinimo savo gyvenimo lygiu reikšmių skirstinys yra suderinamas su normaliuoju**⁶¹. Primename, kad sprendimas, kaip ir kitais atvejais, priimamas taip:

jei $p\text{-reikšmė} \geq \alpha\text{-reikšmė}$, tai priimama (neatmetama) nulinė hipotezė (H_0);
jei $p\text{-reikšmė} < \alpha\text{-reikšmė}$, tai H_0 atmetama ir priimama alternatyvioji hipotezė (H_A).

Pavyzdys, kaip tikrinamos empirinio kiekybinio kintamojo skirstinio suderinamumo su normaliuoju skirstiniu hipotezės su SPSS ir Stata programiniai paketais, yra pateiktas 4.2 lentelėje.

4.2 lentelė. Hipotezės apie kiekybinio masinio reiškinio požymio reikšmių skirstinio suderinamumą su normaliuoju skirstiniu tikrinimas: Kolmogorovo-Smirnovas testas (reikšmingumo lygmuo $\alpha = 0,05$).

Sprendimas su SPSS 15.0		Sprendimas su Stata SE 10.0	
Parengiamieji veiksmai			
Nėra.		Stata programiniame pakete reikės nurodyti lyginamojo teorinio skirstinio funkciją (mūsų atveju – normaliojo), o tam reikia į Stata „atmintį“ įrašyti keletą parametrų – empirinio skirstinio vidurkį ir standartinį nuokrypį – kad būtų galima užrašyti funkcijos formulę. Tą galima padaryti per meniu, tačiau daug praprasciau įvedant sintaksės komandą.	
Sintaksės komandos			
Nėra.		summarize pasit_gyv_lg	
Tikrinamos hipotezės			
H ₀ : populiacijos požymio reikšmių skirstinys suderinamas su normaliuoju			

⁵⁸ Apskritai, suderinamumo hipotezės rekomenduotina tikrinti keliais būdais, nes remiantis keliais skirtingais kriterijais (kurie rodo tą pačią išvadą) yra mažesnė tikimybė padaryti klaidingą statistinę išvadą. Čia reikia prisiminti ir tai, kad neparametriniai kriterijai yra mažiau galingi nei parametriniai (t.y. naudojant juos tikimybė nepadaryti antros rūšies klaidos – neatmesti klaidingos H_0 – mažesnė). Taigi daugelio neparametrinių kriterijų taikymas tikrinant tą pačią suderinamumo hipotezę iš dalis kompensuoja jų galingumo trūkumą.

⁵⁹ Kitaip šis kriterijus dar vadinamas Kolmogorovo-Smirnovas D, bet tai nekeičia jo esmės. Reikia pažymėti ir tai, kad šis kriterijus yra bendro pobūdžio, t. y. skirtas lyginti du bet kokius skirstinius, taigi nėra specialiai skirtas empirinio skirstinio suderinamumo su normaliuoju skirstiniu hipotezei tikrinti ir kartais nesugeba „kokybiškai“ fiksuoti to suderinamumo. Todėl egzistuoja keletas jo korekčių skirtų specifiskai suderinamumo su normaliuoju skirstiniu hipotezėms tikrinti. SPSS atveju yra skaičiuojami Lilliefors D ir Shapiro-Wilk W, o Stata atveju – Asimetrijos ir eksceso, Shapiro-Wilk W ir Shapiro-Francia W' testai.

⁶⁰ Apie tai plačiau kalbama studijų paketo antroje dalyje, 4.8 poskyryje ir, pavyzdžiui, <http://www.physics.csbsju.edu/stats/KS-test.html> (žiūrėta 2008.05.30).

⁶¹ Taip kartu patikrinsime ir tai, ar galėjome taikyti *Studento t* kriterijų (statistiką) praeitame poskyryje tikrindami hipotezę apie europiečių pasitenkinimo savo gyvenimo lygiu reikšmių vidurkio lygybę vidurinei matavimo skalės reikšmei 5.

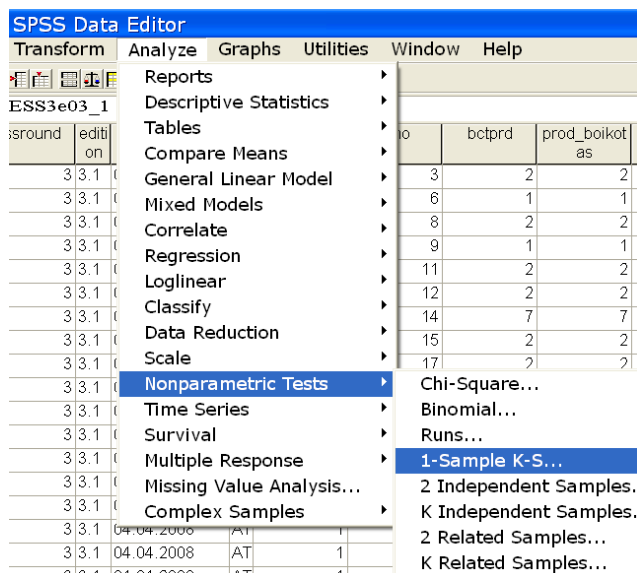
Sprendimas su SPSS 15.0

Sprendimas su Stata SE 10.0

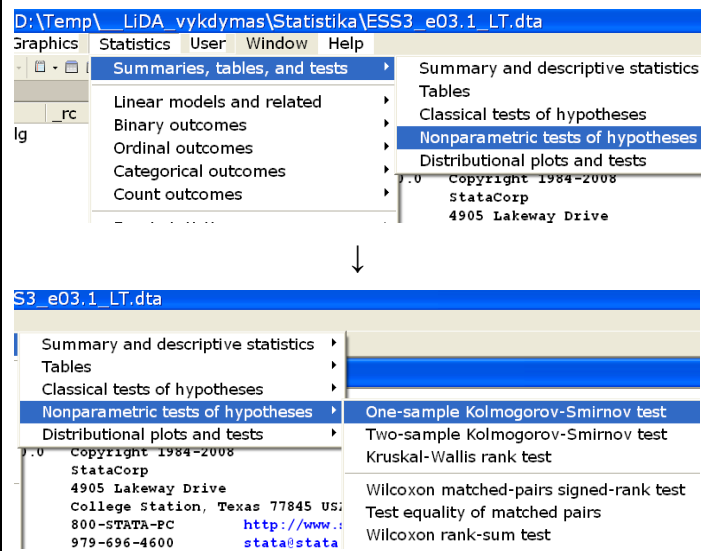
H_A : populiacijos požymio reikšmių skirstinys nesuderinamas su normaliuoju

Analizės veiksmai

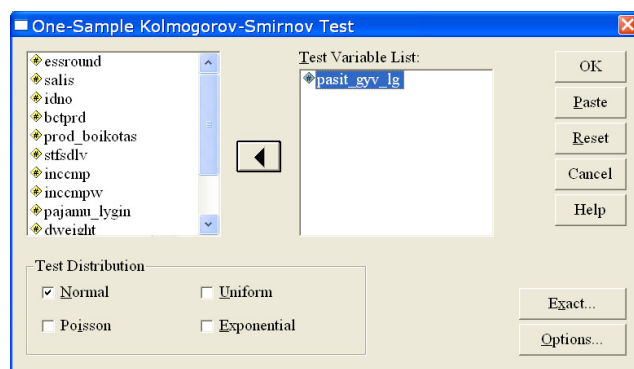
Meniu punktai: Analyze → Nonparametric Tests → 1-Sample K-S...



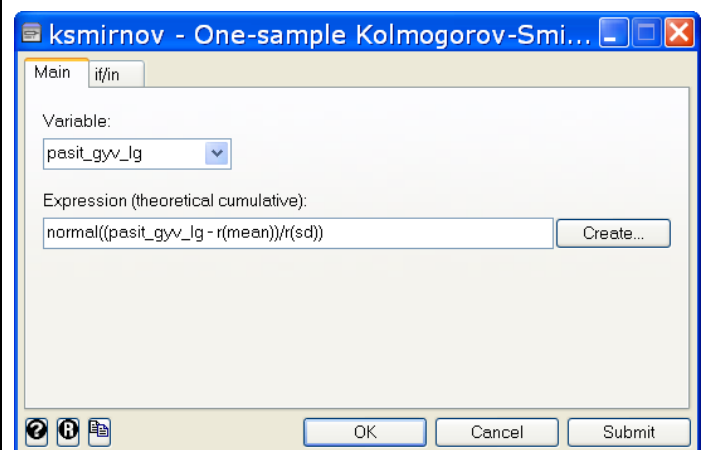
Meniu punktai: Statistics → Classical tests of hypotheses → One-sample mean-comparison test



Atsidariusiame lange „**One-Sample Kolmogorov-Smirnov Test**“ į dešinėje pusėje esantį langelį „**Test Variable List**“ įkeliam kintamąjį **pasit_gyv_lg**, nurodome, kad empirinis skirstinys būtų lyginamas su normaliuoju ties „**Test Distribution**“ pažymėdami „**Normal**“, ir spaudžiame „**OK**“.



Atsidariusiame lange „**ksmirnov – One-sample Kolmogorov-Smirnov test**“ parenkame analizės kintamąjį **pasit_gyv_lg** ties „**Variable**“, nurodome lyginamojo teorinio skirstinio funkciją ties „**Expression (theoretical cumulative)**“ įvesdami formulę $\text{normal}((\text{pasit_gyv_lg} - r(\text{mean})) / r(\text{sd}))$ ir spaudžiame „**OK**“.



Sintaksės komandos

`NPARTESTS /K-S(NORMAL)= pasit_gyv_lg /MISSING ANALYSIS .`

`ksmirnov pasit_gyv_lg = normal((pasit_gyv_lg - r(mean))/r(sd))`

Sprendimas su SPSS 15.0

Sprendimas su Stata SE 10.0

Gauname tokius rezultatus **rezultatų išvesties lange**:

One-Sample Kolmogorov-Smirnov Test

		Pasitenkinimas savo gyvenimo lygiu
N		42754
Normal Parameters ^{a,b}	Mean	6,46
	Std. Deviation	2,400
Most Extreme Differences	Absolute	,175
	Positive	,083
	Negative	-,175
Kolmogorov-Smirnov Z		36,222
Asymp. Sig. (2-tailed)		,000

a. Test distribution is Normal.

b. Calculated from data.

One-sample Kolmogorov-Smirnov test against theoretical distribution
normal((pasit_gyv_lg - r(mean))/r(sd))

Smaller group	D	P-value	Corrected
pasit_gyv_lg:	0.0830	0.000	
Cumulative:	-0.1752	0.000	
Combined K-S:	0.1752	0.000	0.000

Note: ties exist in dataset;
there are 11 unique values out of 42754 observations.

Statistinė išvada

Su 95 proc. garantija galime teigti, kad europiečių pasitenkinimo savo gyvenimo lygiu reikšmių skirstinys nėra suderinamas su normaliuoju (H_0 atmesta, nes $p = 0,000 < \alpha = 0,05$).

Taigi atlikę statistinį tyrimą ir patikrinę suformuluotą hipotezę vadovaudamiesi statistinių hipotezių tikrinimo logika su 95 proc. garantija galime teigti, kad europiečių pasitenkinimo savo gyvenimo lygiu reikšmių skirstinys nėra suderinamas su normaliuoju, ir **todėl praeitame skyriuje atlikta analizė bei padaryta išvada yra nekorektiška**.

4.3. Hipotezių apie populiacijos proporciją tikrinimas

Iki šiol kalbėjome apie statistinių hipotezių tikrinimą susijusį su kiekybinių masinių reiškinių požymių charakteristikų palyginimu su tam tikromis hipotetinėmis reikšmėmis. Analogiška analizė galima ir susidūrus su kokybiniais masinių reiškinių požymiais. Tokiais atvejais tikrinama, ar tam tikros kokybinio požymio reikšmės proporcija lygi hipotetinei, t. y. tyrinėtojo numatoma pagal teoriją ar ankstesnius empirinius rezultatus, reikšmei. Kadangi proporcija paprastai skaičiuojama kokybinių kintamųjų reikšmių atveju, tai mūsų tiriamas masinio reiškinio požymis turi būti matuojamas pavadinimų skalėje. Taigi šiuo atveju mūsų **tikslas** – sužinoti, ar mus dominančio kokybinio masinio reiškinio požymio tam tikros reikšmės proporcija skiriasi nuo hipotetinės. Kadangi netirsime viso masinio reiškinio, o tik iš jo atrinktos imties duomenis, tai **visiškai užtikrintai pasakyti, ar tos proporcijos skiriasi, ar – ne, negalėsime**. Tačiau statistikos mokslo pagalba ir **pasitelkdami hipotezių tikrinimo algoritmą būsime tam tikru mastu garantuoti (paprastai 95 proc.)** apie to skirtumo buvimą ar nebuvimą.

Čia reikia iš karto pasakyti, kad hipotezių apie kokybinio masinio reiškinio požymio tam tikros reikšmės proporciją tikrinimo algoritmai skiriasi priklausomai nuo to kiek reikšmių tas kokybinis požymis turi: dvi ar daugiau nei dvi. Pirmiausiai, panagrinėsime atvejį, kai kokybinis požymis yra dvireikšmis. Kad iliustruotume tai, kaip tikrinamos hipotezės apie dvireikšmio masinio reiškinio požymio vienos reikšmės proporciją, ištirsime **europiečių polinkį boikotuoti (dėl ekologinių ar politinių priežasčių) tam tikrų produktų pirkimą**. Buvo klausta, ar respondentai įvairiose Europos šalyse boikotavo tam tikrų produktų pirkimą per praėjusius 12 mėn. (atsakymų variantai: 1 – taip, 0 – ne). Kaip matome, atsakymai į klausimą yra išmatuoti **dvireikšmėje pavadinimų skalėje**. Kadangi apibendrintos visos Europos mastu išvados nėra labai informatyvios, tai visada įdomiau palyginti atskirų šalių padėtį. Empiriniai tyrimai rodo, kad politinė nekonvencinė

elgsena (boikotai, streikai, peticijos ir pan.) yra tradiciškai labiau išreikšta Vakarų Europos šalyse, ypač Skandinavijoje, negu Rytų Europoje. Taigi yra pagrindo manyti, kad Skandinavijos šalyse tokia politinė nekonvencinė elgsena būdinga ne mažiau kaip ketvirtadaliui gyventojų, tuo tarpu Rytų Europos šalyse – mažiau nei kas dvidešimtas pilietis imasi nekonvencinių politinių veiksmų. Todėl patikrinsime dvi hipotezės:

Daugiau nei 25 proc. Suomijos gyventojų boikotavo kokius nors produktus vienu metu laikotarpyje (tokia jų dalis atsakė į nagrinėjamą klausimą pasirinkdami atsakymą 1 – taip).

Mažiau nei 5 proc. Estijos gyventojų boikotavo kokius nors produktus vienu metu laikotarpyje (tokia jų dalis atsakė į nagrinėjamą klausimą pasirinkdami atsakymą 1 – taip).

Hipotezės, kai kokybinis požymis turi dvi reikšmes, tikrinimui naudojamas Z statistinis kriterijus (statistika)⁶², o pats testas vadinamas *binominiu testu* (**binomial test**). Primename, kad sprendimas, kaip ir kitais atvejais, priimamas taip:

jei $p\text{-reikšmė} \geq \alpha\text{-reikšmė}$, tai priimama (neatmetama) nulinė hipotezė (H_0);
jei $p\text{-reikšmė} < \alpha\text{-reikšmė}$, tai H_0 atmetama ir priimama alternatyvioji hipotezė (H_A).

Pavyzdys, kaip tikrinamos hipotezės apie dvireikšmio masinio reiškimo požymio vienos reikšmės proporcijos lygybę hipotetinei proporcijai su SPSS ir Stata, yra pateiktas 4.3 lentelėje.

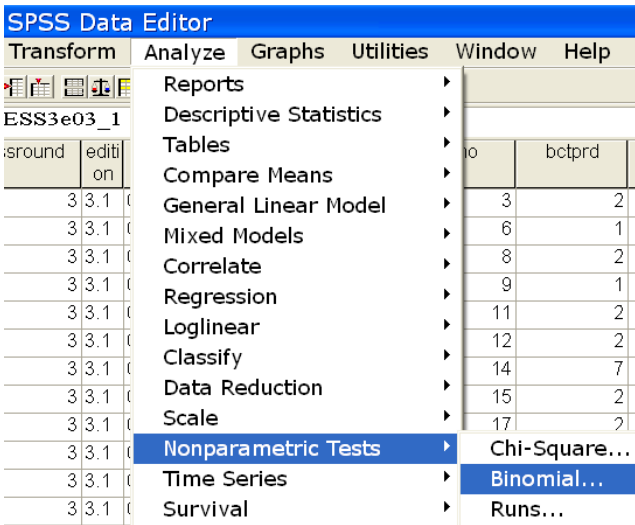
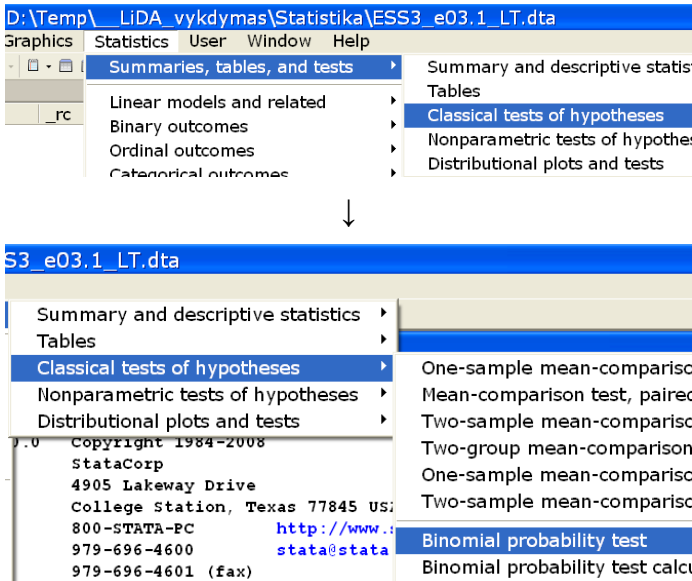
4.3 lentelė. Hipotezės apie dvireikšmio masinio reiškimo požymio vienos reikšmės proporcijos lygybę hipotetinei proporcijai tikrinimas: binominis testas (reikšmingumo lygmuo $\alpha = 0,05$).

Sprendimas su SPSS 15.0	Sprendimas su Stata SE 10.0
Hipotezė apie Suomijos gyventojus	
Parengiamieji veiksmai	
Prieš tikrinant hipotezę apie dvireikšmio masinio reiškimo požymio vienos reikšmės proporcijos lygybę hipotetinei proporcijai tam tikroje grupėje SPSS pakete reikia iš anksto nurodyti, kad bus analizuojami tik tam tikri (šiuo atveju Suomijos) duomenys. Kaip tą padaryti, jau buvo nurodyta 2.4 lentelėje , todėl čia nekartosime.	Nėra.
Sintaksės komandos	
USE ALL. COMPUTE filter_\$=(salis = 10). VARIABLE LABEL filter_\$ 'salis = 10 (FILTER)'. VALUE LABELS filter_\$ 0 'Not Selected' 1 'Selected'. FORMAT filter_\$ (f1.0). FILTER BY filter_\$. EXECUTE .	Nėra.
Tikrinamos hipotezės	
H_0 : populiacijos požymio reikšmės „1“ proporcija = 0,25	H_0 : populiacijos požymio reikšmės „1“ proporcija = 0,25

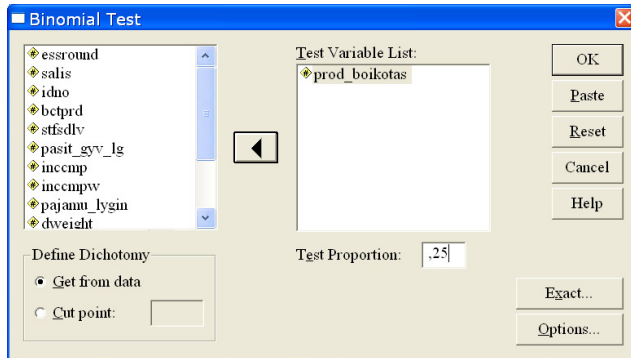
⁶² Tiesa, jis naudojamas tik tuo atveju, kai imtis yra pakankamai didelė (didesnė už 30).

Sprendimas su SPSS 15.0	Sprendimas su Stata SE 10.0
H_A : populiacijos požymio reikšmės „1“ proporcija $> 0,25$	H_{A1} : populiacijos požymio reikšmės „1“ proporcija $< 0,25$ H_{A2} : populiacijos požymio reikšmės „1“ proporcija $\neq 0,25$ H_{A3} : populiacijos požymio reikšmės „1“ proporcija $> 0,25$

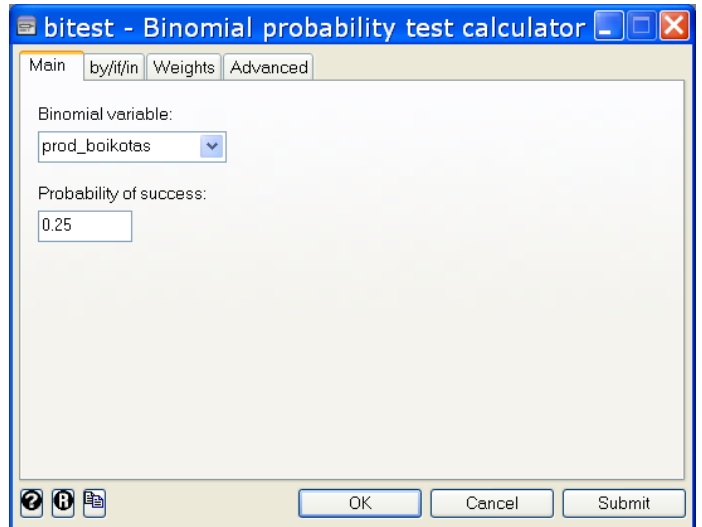
Analizės veiksmiai

Meniu punktai: Analyze → Nonparametric Tests → Binomial... 	Meniu punktai: Statistics → Classical tests of hypotheses → Binomial probability test 
Atsidariusiame lange „Binomial Test“ į dešinėje pusėje esantį langelį „Test Variable List“ įkeliam kintamąjį prod_boikotas, nurodome hipotetinės proporcijos reikšmę ties „Test Proportion“ įvesdami skaičių , 25 ir spaudžiame „OK“.	Atsidariusiame lange „bitest – Binomial test probability calculator“ parenkame analizės kintamąjį prod_boikotas ties „Binomial variable“ ir nurodome hipotetinės proporcijos reikšmę ties „Probability of success“ įvesdami skaičių 0.25.

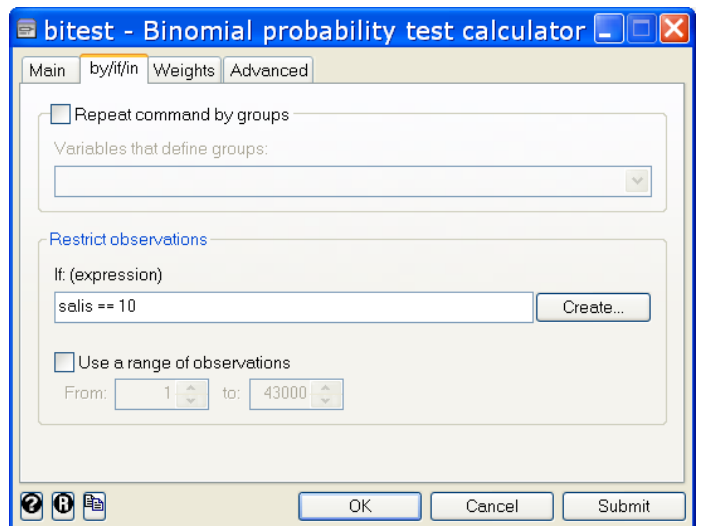
Sprendimas su SPSS 15.0



Sprendimas su Stata SE 10.0



Atidare kortelę „by/if/in“ nurodome sąlygą ties „If: (expression)“: `salis == 10` ir spaudžiame „OK“.



Sintaksės komandos

```
NPAR TEST /BINOMIAL (.25) =  
prod_boikotas /MISSING ANALYSIS .
```

```
bitest prod_boikotas == 0.25 if salis ==  
10
```

Gauname tokius rezultatus **rezultatų išvesties lange**:

Binomial Test					
	Category	N	Observed Prop.	Test Prop.	Asymp. Sig. (1-tailed)
Boikotavo tam tikrų produktų pirkimą per pastaruosius 12 mėn.	Group 1 Ne	1358	.72	.25	.000 ^a
	Group 2 Taip	535	.28		
	Total	1893	1.00		

a. Based on Z Approximation.

Variable	N	Observed k	Expected k	Assumed p	Observed p
prod_boik-as	1893	535	473.25	0.25000	0.28262
Pr(k >= 535) = 0.000660 (one-sided test)					
Pr(k <= 535) = 0.999449 (one-sided test)					
Pr(k <= 412 or k >= 535) = 0.001199 (two-sided test)					

Statistinė išvada

Kaip matome, SPSS paketu galima patikrinti tik

Kaip matome, Stata pakete yra tikrinamos visos

Sprendimas su SPSS 15.0	Sprendimas su Stata SE 10.0
vieną alternatyviąją hipotezę, kad empirinė proporcija didesnė už hipotetinę. Tačiau šiuo atveju, mums reikalinga alternatyvioji hipotezė. Todėl su 95 proc. garantija galime teigti, kad empirinė proporcija (0,28) yra statistiškai reikšmingai didesnė už 0,25 (p (Asymp. Sig. (1-tailed)) = 0,000 < α = 0,05).	alternatyviosios hipotezės, todėl iš karto galime pasakyti, kad su 95 proc. garantija empirinė proporcija (0,28262) yra statistiškai reikšmingai didesnė už 0,25 (p (Pr(k >= 535)) = 0,000660 < α = 0,05).

Hipotezė apie Estijos gyventojus

Parengiamieji veiksmai

Prieš tikrinant hipotezę apie dvireikšmio masinio reiškimo požymio vienos reikšmės proporcijos lygybę hipotetinei proporcijai tam tikroje grupėje SPSS pakete reikia iš anksto nurodyti, kad bus analizuojami tik tam tikri (šiuo atveju Suomijos) duomenys. Kaip tą padaryti, jau buvo nurodyta 2.4 lentelėje , todėl čia nekartosime.	Nėra.
---	--------------

Sintaksės komandos

<pre>USE ALL. COMPUTE filter_\$=(salis = 8). VARIABLE LABEL filter_\$ 'salis = 8 (FILTER)'. VALUE LABELS filter_\$ 0 'Not Selected' 1 'Selected'. FORMAT filter_\$ (f1.0). FILTER BY filter_\$. EXECUTE .</pre>	Nėra.
---	-------

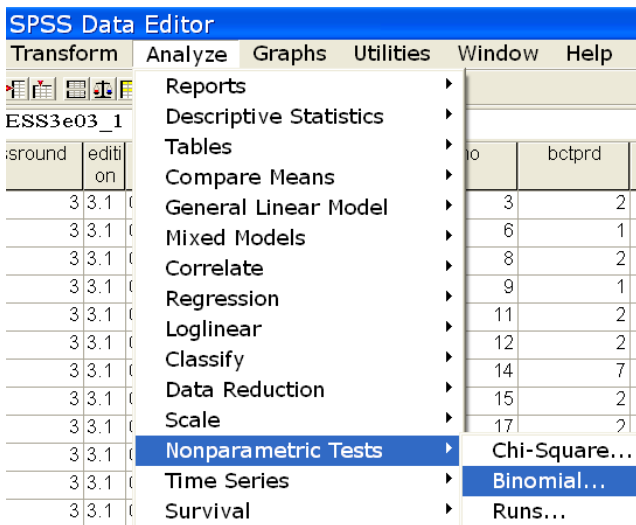
Tikrinamos hipotezės

H_0 : populiacijos požymio reikšmės „1“ proporcija = 0,05 H_A : populiacijos požymio reikšmės „1“ proporcija < 0,05	H_0 : populiacijos požymio reikšmės „1“ proporcija = 0,05 H_{A1} : populiacijos požymio reikšmės „1“ proporcija < 0,05 H_{A2} : populiacijos požymio reikšmės „1“ proporcija $\neq$ 0,05 H_{A3} : populiacijos požymio reikšmės „1“ proporcija > 0,05
--	--

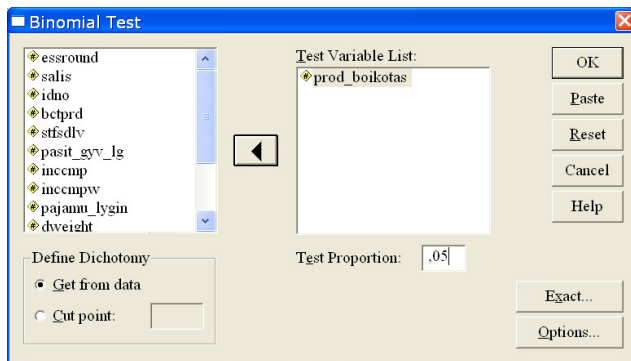
Analizės veiksmai

Meniu punktai: Analyze → Nonparametric Tests → Binomial...	Meniu punktai: Statistics → Classical tests of hypotheses → Binomial probability test
---	--

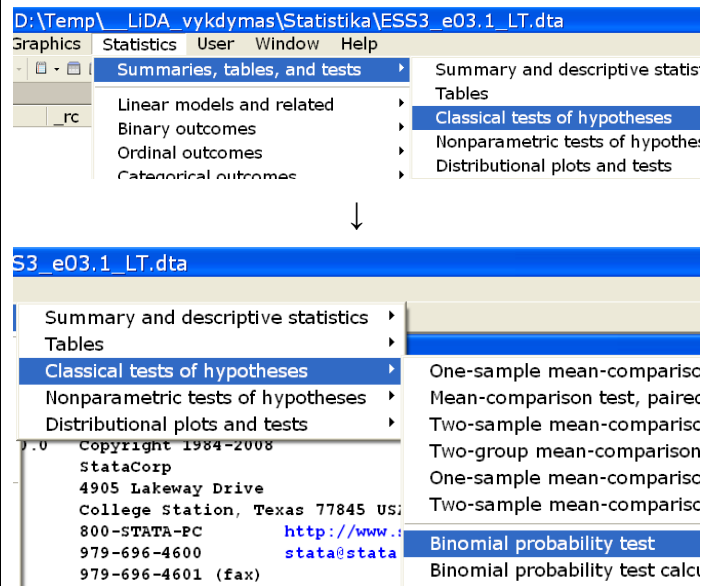
Sprendimas su SPSS 15.0



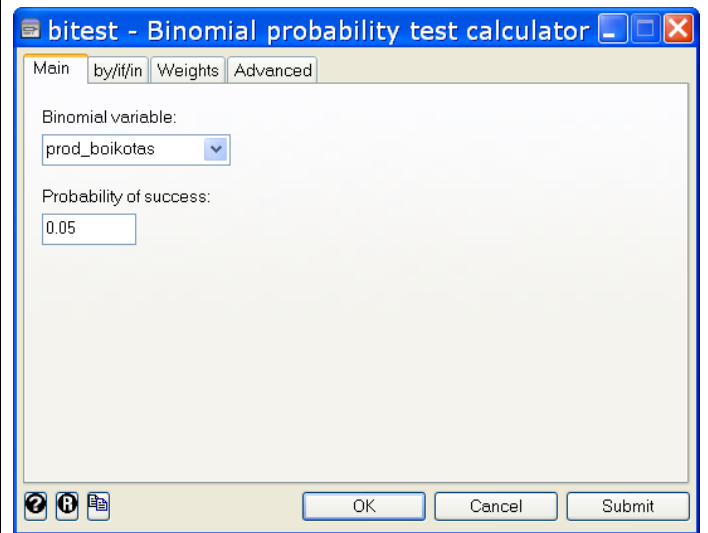
Atsidariusiame lange „**Binomial Test**“ į dešinę pusėje esantį langelį „**Test Variable List**“ įkeliam kintamąjį **prod_boikotas**, nurodome hipotetinės proporcijos reikšmę ties „**Test Proportion**“ įvesdami skaičių , 05 ir spaudžiame „**OK**“.



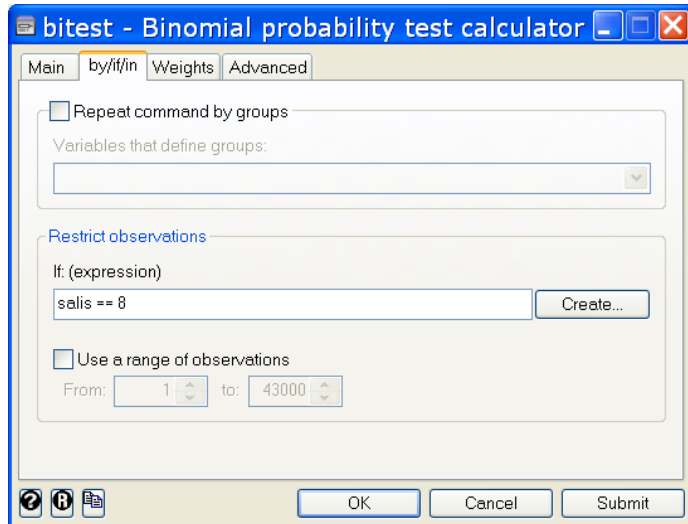
Sprendimas su Stata SE 10.0



Atsidariusiame lange „**bitest – Binomial test probability calculator**“ parenkame analizės kintamąjį **prod_boikotas** ties „**Binomial variable**“ ir nurodome hipotetinės proporcijos reikšmę ties „**Probability of success**“ įvesdami skaičių 0.05.



Atidarę kortelę „**by/if/in**“ nurodome sąlygą ties „**If (expression)**“: **salis == 8** ir spaudžiame „**OK**“.

Sprendimas su SPSS 15.0	Sprendimas su Stata SE 10.0
	

Sintaksės komandos

<pre>NPAR TEST /BINOMIAL (.05) = prod_boikotas /MISSING ANALYSIS .</pre>	<pre>bitest prod_boikotas == 0.05 if salis == 8</pre>
--	---

Gauname tokius rezultatus **rezultatų išvesties lange**:

Binomial Test						
		Category	N	Observed Prop.	Test Prop.	Asymp. Sig. (1-tailed)
Boikotavo tam tikrų produktų pirkimą per pastaruosius 12 mėn.	Group 1	Taip	73	,05	,05	,418 ^{a,b}
	Group 2	Ne	1435	,95		
	Total		1508	1,00		

a. Alternative hypothesis states that the proportion of cases in the first group < .05.

b. Based on Z Approximation.

Variable	N	Observed k	Expected k	Assumed p	observed p
prod_boik-as	1508	73	75.4	0.05000	0.04841
Pr (k >= 73) = 0.628122 (one-sided test)					
Pr (k <= 73) = 0.417786 (one-sided test)					
Pr (k <= 73 or k >= 77) = 0.859203 (two-sided test)					

Statistinė išvada

Kaip matome, SPSS paketu galima patikrinti tik vieną alternatyviąją hipotezę, kad empirinė proporcija mažesnė už hipotetinę. Tačiau šiuo atveju, tai kaip tik mums reikalinga alternatyvioji hipotezė. Todėl su 95 proc. garantija galime teigti, kad empirinė proporcija (0,05) statistiškai reikšmingai nesiskiria nuo hipotetinės 0,05 (p (Asymp. Sig. (1-tailed)) = 0,418 > α = 0,05).	Kaip matome, Stata pakete yra tikrinamos visos alternatyviosios hipotezės, todėl iš karto galime pasakyti, kad su 95 proc. garantija empirinė proporcija (0,04841) yra statistiškai reikšmingai nesiskiria nuo hipotetinės 0,05 (p (Pr(k <= 73)) = 0,417786 > α = 0,05).
--	---

Taigi atlikę statistinį tyrimą ir patikrinę suformuluotą hipotezę vadovaudamiesi statistinių hipotezių tikrinimo logika su 95 proc. garantija galime teigti, kad kaip ir tikėtasi, daugiau nei ketvirtadalis Suomijos gyventojų boikotavo kokius nors produktus vienu metų laikotarpyje. Estijoje tu tarpu produktų pirkimą boikotavo apie 5 proc. gyventojų ir tai iš dalies nepatvirtino pradinės prielaidos, kurioje buvo teigiama, kad ši proporcija bus mažesnė už 5 proc. (tiesa, ji ir ne didesnė).

Toliau panagrinėsime atvejį, kai kokybinis požymis turi daugiau nei dvi reikšmes. Kad iliustruotume tai, kaip tikrinamos hipotezės apie daigianario kokybinio masinio reiškinio požymio reikšmių skirstinio lygybę hipotetiniam skirstiniui⁶³, patyrinėsime **europiečių polinkį lyginti savo**

⁶³ Šiuo atveju tas skirstinys yra diskretusis, taigi tikrinamos empirinės reikšmių proporcijos su hipotetinėmis.

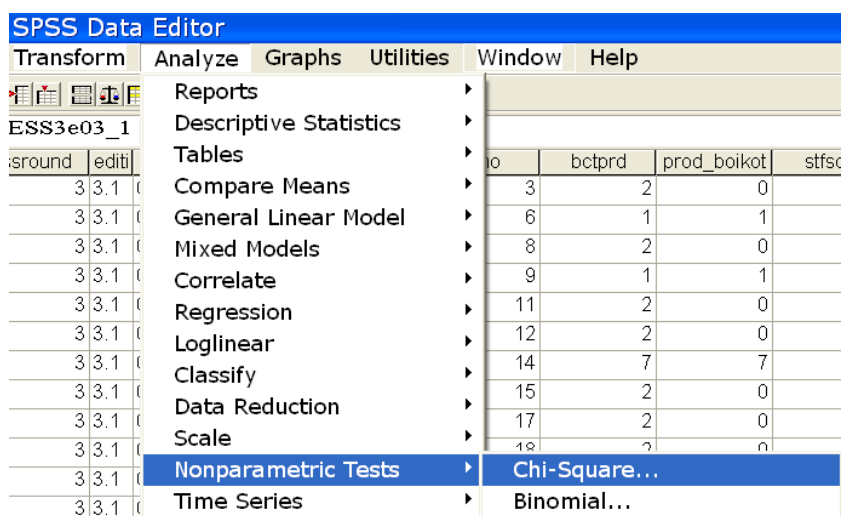
pajamas su kitais. Buvo klausta, su kuo respondentai įvairiose Europos šalyse labiausiai lygina savo pajamas (atsakymų variantai: 1 – kolegomis darbe, 2 – šeimos nariais, draugais ir kitais žmonėmis, 3 – nelygina su niekuo). Kaip matome, atsakymai į klausimą yra išmatuoti **tris reikšmes turinčioje pavadinimų skalėje**. Šiuo atveju pabandysime patikrinti hipotezę, kad europiečių lyginimosi subjektai yra **pasiskirstę proporcingai**, t. y. trečdalis jų lygina savo pajamas su kolegomis darbe, trečdalis – su šeimos nariais, draugais ir pan., o dar trečdalis – nelygina savo pajamų su niekuo. Hipotezės, kai kokybinis požymis turi daugiau nei dvi reikšmes, tikrinimui naudojamas χ^2 statistinis kriterijus (statistika), o pats testas vadinamas χ^2 testu (*chi-square test*). Primename, kad sprendimas, kaip ir kitais atvejais, priimamas taip:

jei $p\text{-reikšmė} \geq \alpha\text{-reikšmė}$, tai priimama (neatmetama) nulinė hipotezė (H_0);
jei $p\text{-reikšmė} < \alpha\text{-reikšmė}$, tai H_0 atmetama ir priimama alternatyvioji hipotezė (H_A).

Pavyzdys, kaip tikrinamos hipotezės apie dauginario kokybinio masinio reiškinio požymio reikšmių skirstinio lygybę hipotetiniui skirstiniui su SPSS⁶⁴, yra pateiktas 4.4 lentelėje.

4.4 lentelė. Hipotezės apie dauginario kokybinio masinio reiškinio požymio reikšmių skirstinio lygybę hipotetiniui skirstiniui tikrinimas: χ^2 testas (reikšmingumo lygmuo $\alpha = 0,05$).

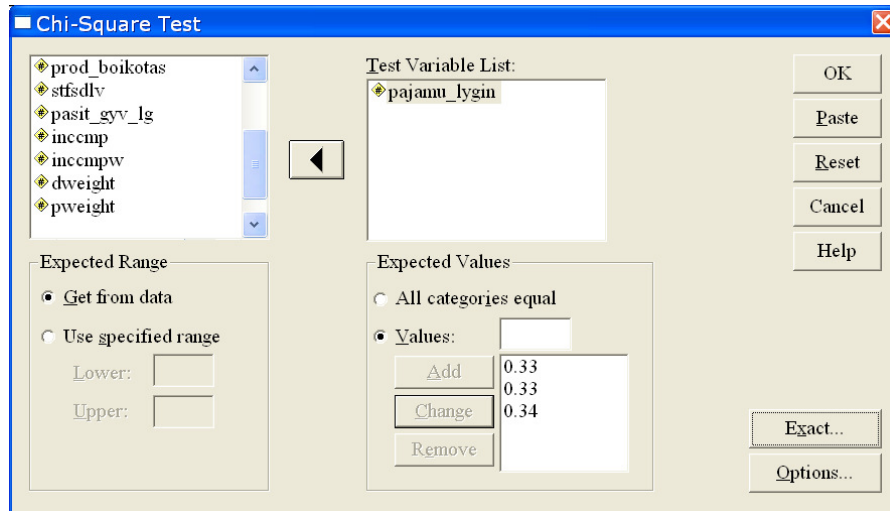
Sprendimas su SPSS 15.0	
Parengiamieji veiksmai	
Nėra.	
Sintaksės komandos	
Nėra.	
Tikrinamos hipotezės	
H_0 : populiacijos požymio reikšmių proporcijų skirstinys suderinamas su hipotetiniu H_A : populiacijos požymio reikšmių proporcijų skirstinys nesuderinamas su hipotetiniu	
Analizės veiksmai	
Meniu punktai: Analyze → Nonparametric Tests → Chi-Square...	



⁶⁴ Stata pakete atlikti tokią analizę galimybės nėra.

Sprendimas su SPSS 15.0

Atsidariusiame lange „*Chi-Square Test*“ į dešinę pusėje esantį langelį „*Test Variable List*“ įkeliam kintamąjį **pajamu_lyginimikotas**, nurodome hipotetinių proporcijų reikšmes ties „*Expected Values*“ įvesdami skaičius 0.33, 0.33 ir 0.34 ir spaudžiame „*OK*“.



Sintaksės komandos

```
NPAR TEST /CHISQUARE = pajamu_lygin /EXPECTED = 0.33 0.33 0.34 /MISSING ANALYSIS.
```

Gauname tokius rezultatus **rezultatų išvesties lange**:

Su kuo labiausiai lygina savo pajamas

	Observed N	Expected N	Residual
Kolegomis darbe	8109	7459,3	649,7
Šeimos nariais, draugais ir kitais žmonėmis	6279	7459,3	-1180,3
Nelygina su niekuo	8216	7685,4	530,6
Total	22604		

Test Statistics

	Su kuo labiausiai lygina savo pajamas
Chi-Square ^a	279,990
df	2
Asymp. Sig.	,000

a. 0 cells (.0%) have expected frequencies less than
5. The minimum expected cell frequency is 7459,3.

Statistinė išvada

Su 95 proc. garantija galime teigti, kad europiečių lyginimosi su kitais reikšmių skirstinys skiriasi nuo hipotetinio – lygių dalių (H_0 atmesta, nes $p = 0,000 < \alpha = 0,05$).

Taigi atlikę statistinį tyrimą ir patikrinę suformuluotą hipotezę vadovaudamiesi statistinių hipotezių tikrinimo logika su 95 proc. garantija galime teigti, kad europiečių pajamų lyginimosi subjektai **nėra pasiskirstę proporcingai**, t. y. vienos, kelių ar visų reikšmių proporcijų dydžiai skiriasi nuo $1/3$ ⁶⁵.

⁶⁵ Čia reikia pažymėti, kad šioje analizėje nenustatoma, kurių būtent reikšmių proporcijos skiriasi.

5. VIDURKIŲ PALYGINIMAS. DISPERSINĖ ANALIZĖ (ANOVA)

Šiame skyriuje pateikiami pavyzdžiai, kaip su SPSS ir Stata atlikti vidurkių palyginimo ir dispersinę (ANOVA) analizę.

5.1. Vidurkių palyginimo analizė

5.1.1. Vidurkių palyginimas su SPSS

Parametriniai kriterijai (dviejų grupių/imčių):

- a) dviejų nepriklausomų imčių T-kriterijus leidžia nustatyti, ar to pačio kintamojo vidurkių skirtumai sugrupuotuose duomenyse (t.y., tarp grupių) yra statistiškai reikšmingi. Bene pagrindinė problema tyrimuose yra tai, kad apklausomis grįstuose tyrimuose, ypač sociologiniuose ir politikos mokslų, duomenys labai retai yra intervalinės skalės ir normalaus skirstinio, todėl dažniausiai tenka taikyti neparametrinius kriterijų analogus. Ištirkime, remdamiesi 2004 m. Naujojo Baltijos Barometro tyrimo duomenimis, ar respondentų iš rajono centrų pasitikėjimas demokratinėmis bei tradicinio autoriteto valstybės institucijomis statistiškai reikšmingai skiriasi nuo didžiųjų miestų, neįtraukiant Vilniaus, pasitikėjimo lygio (žr. vienfaktorinės dispersinės analizės pavyzdį, kuriame įvertinami skirtumai tarp daugiau negu dviejų grupių):

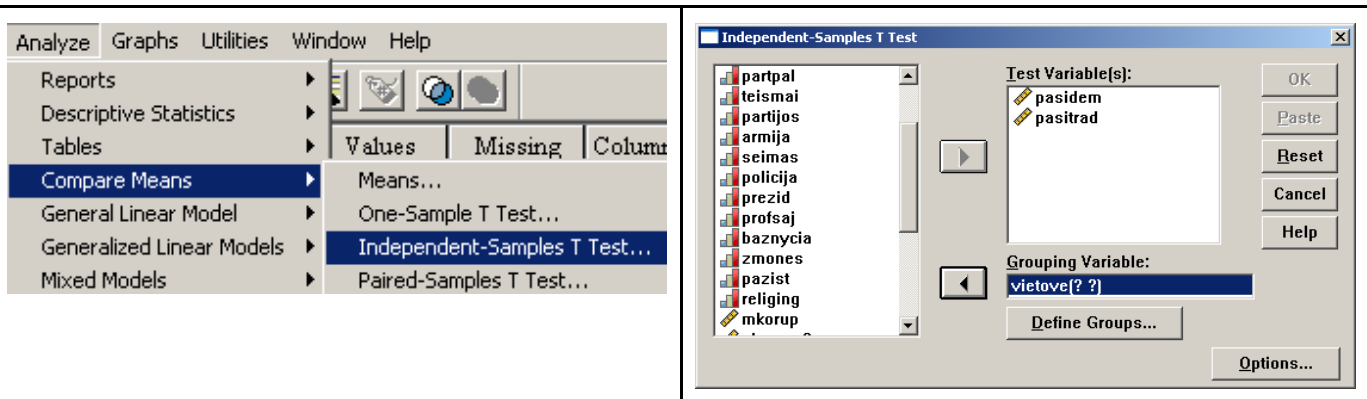
Sprendimas su SPSS 15.0

Parengiamieji veiksmai

Abu priklausomi kintamieji atitinka normalumo sąlygą (žr. dispersinės analizės pavyzdį).

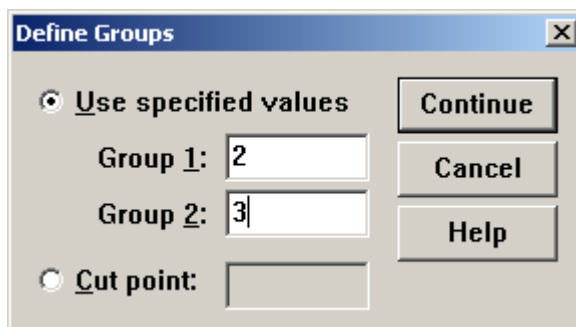
Analizės veiksmai

Meniu punktai: Analyze → Compare Means... → Independent-Samples T test...



Iš kairės pusės kintamųjų sąrašo į dešinę pusę perkeliame kintamuosius **pasidem** ir **pasitrad**. „**Grouping Variable**“: perkeliame iš kairės pusės kintamųjų sąrašo į dešinę pusę kintamąjį **vietove**, paspaudžiame „**Define Groups**“, atsiradusiame lange pasirenkame, kurias dvi grupes lyginsime: įrašome tų grupių kodus, pvz., didžiųjų miestų, išskyrus Vilnių, respondentų kodas yra „2“, o respondentų iš rajonų centrų kodas yra „3“ → Continue → OK.

Sprendimas su SPSS 15.0



Sintaksės komandos

T-TEST

GROUPS = vietove(2 3)
 /MISSING = ANALYSIS
 /VARIABLES = pasidem pasitrad
 /CRITERIA = CI(.95) .

Gauname tokius rezultatu :

Group Statistics

	Town size	N	Mean	Std. Deviation	Std. Error Mean
Pasitikėjimas demokratinės valstybės institucijomis	Kaunas, Klaipeda, Siauliai, Panevezys	270	-,1339558	1,04087786	,06334581
	Rajono centras	265	,1030562	,90217421	,05542012
Pasitikėjimas tradicinio autoriteto institucijomis	Kaunas, Klaipeda, Siauliai, Panevezys	270	,0139743	1,04933947	,06386077
	Rajono centras	265	-,1172733	,97691705	,06001154

Independent Samples Test

		Levene's Test for Equality of Variances		t-test for Equality of Means						
		F	Sig.	t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	95% Confidence Interval of the Difference	
									Lower	Upper
Pasitikėjimas demokratinės valstybės institucijomis	Equal variances assumed	3,596	,058	-2,812	533	,005	-,23701199	,08427900	-,40257175	-,07145223
	Equal variances not assumed			-2,816	524,993	,005	-,23701199	,08416698	-,40235743	-,07166654
Pasitikėjimas tradicinio autoriteto institucijomis	Equal variances assumed	,558	,455	1,497	533	,135	,13124756	,08769183	-,04101643	,30351155
	Equal variances not assumed			1,498	531,523	,135	,13124756	,08763323	-,04090240	,30339752

Remiantis antroje lentelėje pateiktu Levene kriterijaus empiriniu reikšmingumu, abiejuose grupėse dispersijos yra vienos ($p_1=0,058>0,05$; $p_2=0,455>0,05$), todėl remiamės t-kriterijaus reikšmingumais, pateiktas eilutėse su prierašu „Dispersija grupėse vienoda“ („Equal variances assumed“).

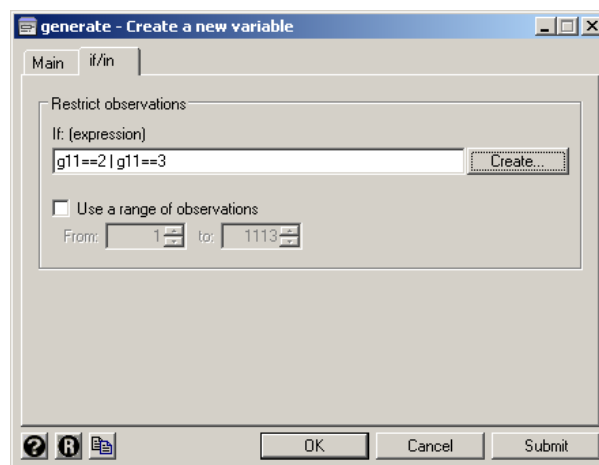
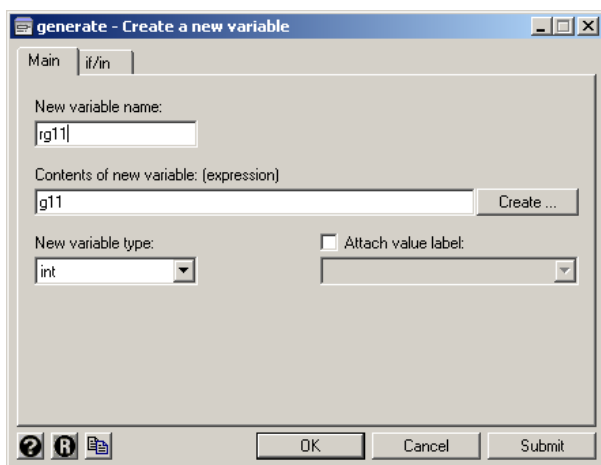
Gavome, kad pasitikėjimo lygis demokratinėmis valstybės institucijomis statistiškai reikšmingai skiriasi tarp rajonų ir didžiųjų miestų (neįtraukus Vilniaus) gyventojų ($p=0,005<0,01$), tuo tarpu pasitikėjimo lygis tradicinio autoriteto institucijomis tarp šių dviejų grupių nesiskiria ($p=0,135>0,05$). Pirmoji lentelė parodo, kad rajonų centrų gyventojai linkę daugiau pasitikėti demokratinėmis valstybės institucijomis (vidurkis 0,103) negu didžiųjų miestų (neįtraukus Vilniaus) gyventojai (vidurkis -0,134).

5.1.2. Vidurkių palyginimas su Stata

Sprendimas su Stata SE 10.0

Parengiamieji veiksmai

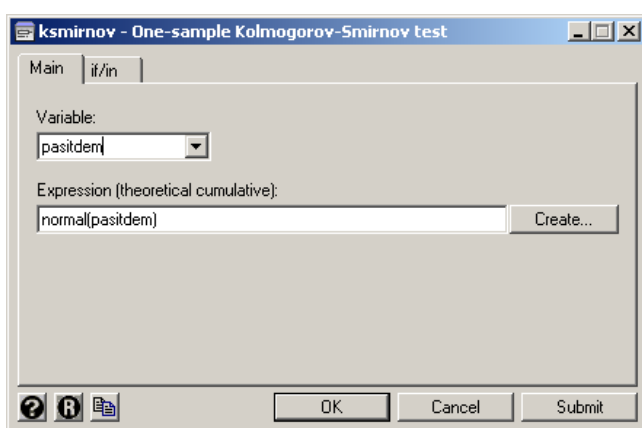
Iš 5 vietovių grupių kintamojo *g11* sukuriame 2 grupių kintamąjį *rg11*:



Sintaksės komandos

generate int rg11 = g11 if g11==2 | g11==3

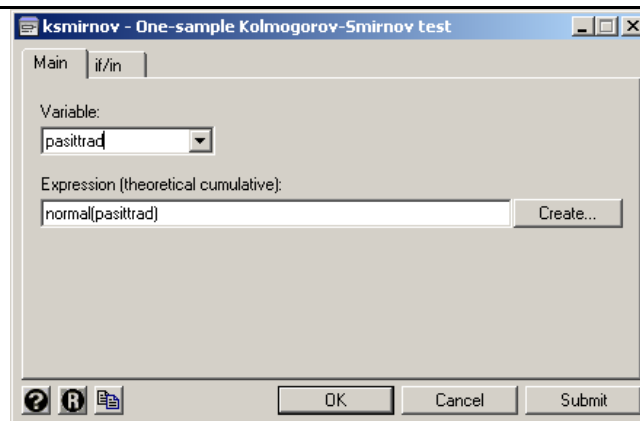
Patikriname, ar priklausomi kintamieji **pasidem** (pasitikėjimas demokratinėmis valstybės institucijomis) ir **pasitrad** (pasitikėjimas tradicinio autoriteto valstybės institucijomis) turi normalųjį skirstinį:



One-sample Kolmogorov-Smirnov test against theoretical distribution normal(F2)

Smaller group	D	P-value	Corrected
F2:	0.0295	0.187	
Cumulative:	-0.0169	0.575	
Combined K-S:	0.0295	0.371	0.355

Note: ties exist in dataset;
there are 934 unique values out of 967 observations.



One-sample Kolmogorov-Smirnov test against theoretical distribution normal(pasitrad)

Smaller group	D	P-value	Corrected
pasitrad:	0.0238	0.284	
Cumulative:	-0.0394	0.032	
Combined K-S:	0.0394	0.064	0.059

Note: ties exist in dataset;
there are 1078 unique values out of 1113 observations.

Gauname, kad tiek **pasidem**, tiek **pasitrad** kintamojo skirstinys yra suderinamas su normaliuoju

Sprendimas su Stata SE 10.0

skirstiniu.

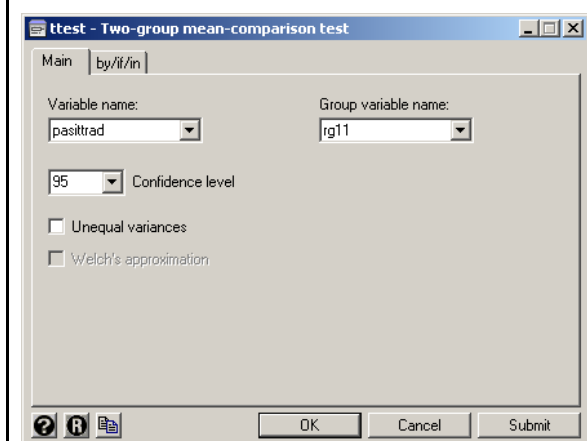
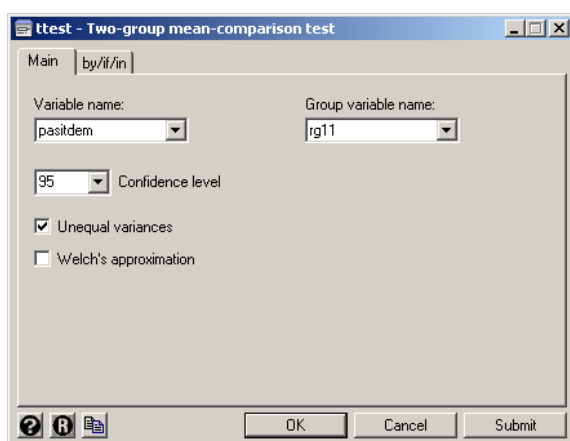
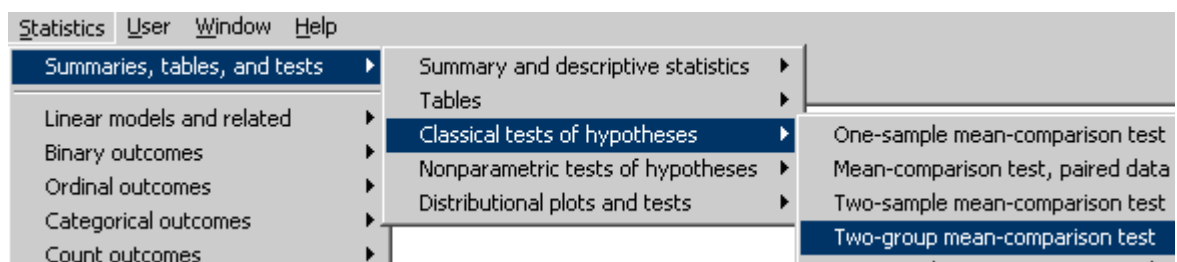
Patikrinkime, ar priklausomų kintamųjų dispersija nepriklausomo kintamojo grupėse yra vienoda (įrašome STATA komandų eilutėje)

Sintaksės komandos

sdtest pasidem, by(rg11)							sdtest pasitrad, by(rg11)						
Variance ratio test							Variance ratio test						
Group	obs	Mean	Std. Err.	Std. Dev.	[95% Conf. Interval]		Group	obs	Mean	Std. Err.	Std. Dev.	[95% Conf. Interval]	
2	270	-.1339558	.0633458	1.040878	-.2586724	-.0092392	2	270	.0139743	.0638608	1.049339	-.1117562	.1397047
3	265	.1030562	.0554201	.9021742	-.0060655	.2121779	3	265	-.1172733	.0600115	.9769171	-.2354355	.0008888
combined	535	-.0165573	.0424094	.9809314	-.099867	.0667523	combined	535	-.0510362	.0438949	1.015291	-.137264	.0351916
ratio = sd(2) / sd(3) f = 1.3311							ratio = sd(2) / sd(3) f = 1.1538						
Ho: ratio = 1 degrees of freedom = 269, 264							Ho: ratio = 1 degrees of freedom = 269, 264						
Ha: ratio < 1 Pr(F < f) = 0.9900							Ha: ratio < 1 Pr(F < f) = 0.8780						
Ha: ratio != 1 2*Pr(F > f) = 0.0200							Ha: ratio != 1 2*Pr(F > f) = 0.2439						
Ha: ratio > 1 Pr(F > f) = 0.0100							Ha: ratio > 1 Pr(F > f) = 0.1220						
Gavome, kad pasidem kintamojo dispersijos grupėse yra nelygios (p = 0,020)							Gavome, kad pasitrad kintamojo dispersijos grupėse yra lygios (0,244)						

Nepriklausomų imčių t-kriterijaus analizės veiksmas

Meniu punktai: Atliekame dviejų nepriklausomų imčių t-kriterijaus analizę:



Sintaksės komandos

ttest pasidem, by(rvietove) unequal	ttest pasitrad, by(rvietove)
-------------------------------------	------------------------------

Gauname tokius rezultatus:

Sprendimas su Stata SE 10.0

Two-sample t test with unequal variances						Two-sample t test with equal variances					
Group	obs	Mean	Std. Err.	Std. Dev.	[95% Conf. Interval]	Group	obs	Mean	Std. Err.	Std. Dev.	[95% Conf. Interval]
2	270	-.1339558	.0633458	1.040878	-.2586724	2	270	.0139743	.0638608	1.049339	-.1117562
3	265	-.1030562	.0554201	.9021742	-.0060655	3	265	-.1172733	.0600115	.9769171	-.2354355
combined	535	-.0165573	.0424094	.9809314	-.099867	combined	535	-.0510362	.0438949	1.015291	-.137264
diff		-.237012	.084167		-.4023574	diff		.1312476	.0876918		-.0410164
diff = mean(2) - mean(3)						diff = mean(2) - mean(3)					
Ho: diff = 0						Ho: diff = 0					
Satterthwaite's degrees of freedom = 524.993						degrees of freedom = 533					
t = -2.8160						t = 1.4967					
Ha: diff < 0						Ha: diff != 0					
Pr(T < t) = 0.0025						Pr(T < t) = 0.9325					
Ha: diff != 0						Pr(T > t) = 0.1351					
Pr(T > t) = 0.9975						Ha: diff > 0					
						Pr(T > t) = 0.0675					

Gavome, kad **pasidem** kintamojo vidurkiai **rvietove** kintamojo grupėse statistiškai reikšmingai skiriasi ($p = 0,005 < 0,01$), o **pasitrad** kintamojo vidurkiai **rvietove** kintamojo grupėse nesiskiria ($p = 0,135 > 0,05$).

Tad formuluojuame išvadą, kad pasitikėjimo demokratinės valstybės institucijomis lygmuo statistiškai reikšmingai didesnis tarp rajonų centrų gyventojų negu tarp didžiųjų miestų, išskyrus Vilnių, gyventojų (skirtumo 95% pasikliautinis intervalas yra $[-0.402; -0.072]$); pasitikėjimo tradicinio autoriteto institucijomis lygmuo vienodas tarp didžiųjų miestų, išskyrus Vilnių, ir rajonų centrų gyventojų.

5.2. Dispersinė analizė (ANOVA)

5.2.1. Dispersinė analizė su SPSS

Pateikdami pavyzdžius apie dispersinę analizę, aptarsime tik pilnai randomizuotą vienfaktorinės dispersinės analizės modelį. Panagrinėsime, kaip pasitikėjimas institucijomis skiriasi respondentų grupėse pagal vietas, kuriose buvo atliekamas tyrimas (Naujasis Baltijos Barometras, 2004 m.).

Sprendimas su SPSS 15.0

Parengiamieji veiksmai: prielaidų tikrinimas

Turi būti tenkinamos šios prielaidos: 1. kiekviena grupė yra nepriklausoma atsitiktinė imtis iš normalaus skirtingo populiacijos (mažiau griežtas reikalavimas yra skirstinių simetriškumas, kadangi ANOVA testas yra nejautrus nuokrypiams nuo reikalavimo nagrinėti normalaus skirtingo kintamąjį); 2. grupės turi turėti lygias dispersijas (tikrinama Levene variacijos homogeniškumo kriterijumi).

Visų pirma patikriname, ar k populiacijų (t.y., respondentų grupių pagal miestus) turi normalųjį skirstinį.

Meniu punktai: Analyze → Nonparametric tests → 1 Sample K-S: pasirenkame kintamuosius **pasidem** ir **pasitrad**, spaudžiame „OK“.

Sprendimas su SPSS 15.0

One-Sample Kolmogorov-Smirnov Test

		Pasitikėjimas demokratinės valstybės institucijomis	Pasitikėjimas tradicinio autoriteto institucijomis
N		1113	1113
Normal Parameters ^{a,b}	Mean	,0000000	,0000000
	Std. Deviation	1,00000000	1,00000000
Most Extreme Differences	Absolute	,027	,039
	Positive	,027	,024
	Negative	-,013	-,039
Kolmogorov-Smirnov Z		,917	1,313
Asymp. Sig. (2-tailed)		,369	,064

a. Test distribution is Normal.

b. Calculated from data.

Lentelėje paskutinė eilutė rodo rezultato reikšmingumo lygmenį $p > 0,05$, todėl kintamųjų skirstiniai yra normalieji.

Galime pakartoti šią analizę kiekvienoje k grupių, kurių vidurkius lyginsime, t.y., patikrinti, ar kiekviename mieste pasitikėjimo kintamieji yra normalaus skirstinio.

Meniu punktai: Data → Split File...: pažymime „Compare Groups“; kairėje pusėje pasirenkame kintamąjį **vietove**, perkeliame į dešinę pusę, spaudžiame „OK“. Tada pakartojame aukščiau atliktą Kolmogorovo-Smirnovo kriterijaus tikrinimo procedūrą.

Sprendimas su SPSS 15.0

One-Sample Kolmogorov-Smirnov Test

Vietovė			Pasitikėjimas demokratinės valstybės institucijomis	Pasitikėjimas tradicinio autoriteto institucijomis
Vilnius	N		210	210
	Normal Parameters ^{a,b}	Mean	-,0506240	-,0597996
		Std. Deviation	1,04624166	1,02215373
	Most Extreme Differences	Absolute	,047	,065
		Positive	,047	,027
		Negative	-,026	-,065
	Kolmogorov-Smirnov Z		,684	,945
	Asymp. Sig. (2-tailed)		,738	,334
Kaunas, Klaipėda, Siauliai, Panevėžys	N		270	270
	Normal Parameters ^{a,b}	Mean	-,1339558	,0139743
		Std. Deviation	1,04087786	1,04933947
	Most Extreme Differences	Absolute	,058	,065
		Positive	,058	,044
		Negative	-,040	-,065
	Kolmogorov-Smirnov Z		,954	1,070
	Asymp. Sig. (2-tailed)		,323	,202
Rajono centras	N		265	265
	Normal Parameters ^{a,b}	Mean	,1030562	-,1172733
		Std. Deviation	,90217421	,97691705
	Most Extreme Differences	Absolute	,037	,042
		Positive	,037	,020
		Negative	-,032	-,042
	Kolmogorov-Smirnov Z		,600	,688
	Asymp. Sig. (2-tailed)		,865	,731
Kiti miestai	N		47	47
	Normal Parameters ^{a,b}	Mean	,3816100	,1993648
		Std. Deviation	,90587085	1,06346865
	Most Extreme Differences	Absolute	,075	,091
		Positive	,075	,091
		Negative	-,063	-,075
	Kolmogorov-Smirnov Z		,514	,625
	Asymp. Sig. (2-tailed)		,954	,830
Kaimo vietovė	N		321	321
	Normal Parameters ^{a,b}	Mean	,0048397	,0949911
		Std. Deviation	1,00640241	,94248275
	Most Extreme Differences	Absolute	,047	,067
		Positive	,047	,035
		Negative	-,028	-,067
	Kolmogorov-Smirnov Z		,835	1,199
	Asymp. Sig. (2-tailed)		,488	,113

a. Test distribution is Normal.

b. Calculated from data.

Reikšmingumo lygmenys visose grupėse $p > 0,05$, t.y., nulinė hipotezė, kad skirstiniai grupėse yra normalieji, neatmetama.

Dispersijos grupėse lygybės tikrinimas atliekamas per Analyze → Descriptives → Explore: į priklausomų kintamųjų sąrašą („Dependent List“) įtraukiame kintamuosius **pasidem** ir **pasitrad**, į faktorių sąrašą („Factor List“) įtraukiame kintamąjį **vietove**. „Plots...“ kortelėje pažymime „Spread vs. Level with Levene Test“ - „Untransformed“ → Continue → OK.

Sprendimas su SPSS 15.0

Test of Homogeneity of Variance

		Levene Statistic	df1	df2	Sig.
Pasitikėjimas demokratinės valstybės institucijomis	Based on Mean	2,224	4	1108	,064
	Based on Median	2,238	4	1108	,063
	Based on Median and with adjusted df	2,238	4	1092,371	,063
	Based on trimmed mean	2,219	4	1108	,065
Pasitikėjimas tradicinio autoriteto institucijomis	Based on Mean	,673	4	1108	,611
	Based on Median	,590	4	1108	,670
	Based on Median and with adjusted df	,590	4	1084,210	,670
	Based on trimmed mean	,648	4	1108	,628

Kitas kelias – tikrinti dispersijos lygybę atliekant vienfaktorinę dispersinę analizę: Analize → Compare Means → One Way ANOVA: į priklausomų kintamųjų sąrašą įtraukiame kintamuosius **pasidem** ir **pasitrad**, į faktorių sąrašą („Factor“) įtraukiame kintamąjį **vietove** . „Options...“ kortelėje pažymime „Homogeneity of variance test“ → Continue → OK

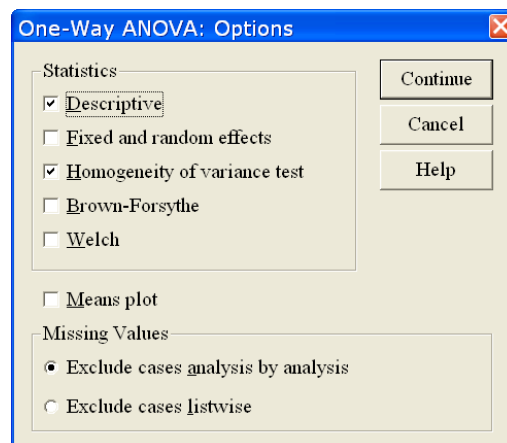
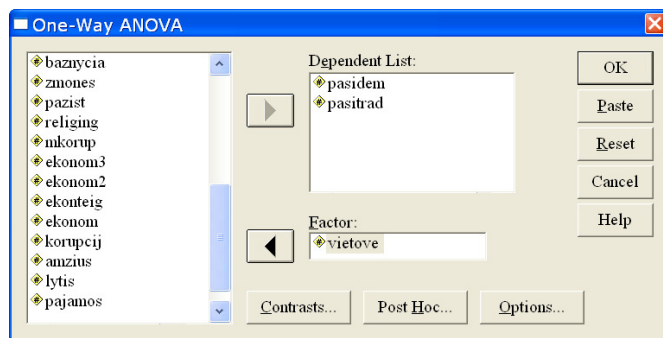
Test of Homogeneity of Variances

	Levene Statistic	df1	df2	Sig.
Pasitikėjimas demokratinės valstybės institucijomis	2,224	4	1108	,064
Pasitikėjimas tradicinio autoriteto institucijomis	,673	4	1108	,611

Levene statistika yra 2,224 ir 0,673 tiriamiems kintamiesiems, o reikšmingumai (Sig.) > 0,05, todėl prie reikšmingumo lygio $p = 0,05$, hipotezė apie variacijų homogeniškumą miestų grupėse yra priimama, t.y., variacijų homogeniškumo grupėse sąlyga yra tenkinama. Jei ši sąlyga nebūtų tenkinama, reiktų transformuoti duomenis, siekiant padidinti jų simetriškumą (pvz., ištraukiant kvadratinę šaknį iš kintamųjų, matuojančių pasitikėjimą institucijomis) arba naudoti neparametrinius kriterijus.

Vidurkių skirtumų analizės veiksmai

Meniu punktai: Analize → Compare Means → One-Way ANOVA: Dependent List: pasirenkame **pasidem**, **pasitrad** kintamuosius; „Factor“: kintamasis **vietove**. „Options“ kortelėje: „Statistics“ skirlyje pažymime „Descriptive“, „Homogeneity-of-variance“. „Post Hoc...“ kortelėje „Equal Variances Assumed“ skiltyje pažymime „Bonferoni“, „Tukey“.



Sprendimas su SPSS 15.0

Sintaksės komandos

```
ONEWAY
  pasidem pasitrad BY vietove
  /STATISTICS DESCRIPTIVES HOMOGENEITY
  /MISSING ANALYSIS
  /POSTHOC = TUKEY BONFERRONI ALPHA(.05).
```

Gauname tokius rezultatus:

ANOVA

		Sum of Squares	df	Mean Square	F	Sig.
Pasitikėjimas demokratinės valstybės institucijomis	Between Groups	15,050	4	3,762	3,800	,004
	Within Groups	1096,950	1108	,990		
	Total	1112,000	1112			
Pasitikėjimas tradicinio autoriteto institucijomis	Between Groups	9,213	4	2,303	2,314	,056
	Within Groups	1102,787	1108	,995		
	Total	1112,000	1112			

ANOVA lentelė parodo, kad pasitikėjimo institucijomis skirtumai pagal vietovės yra reikšmingi pasitikėjimo demokratinėmis institucijomis atveju ($p=0,004<0,05$), o pasitikėjimo tradicinio autoriteto institucijomis skirtumų tarp skirtingų vietovių nėra ($p=0,056>0,05$). Maži statistinio reikšmingumo (Sig.) įverčiai (mažiau negu 0,05) paskutiniame lentelės stulpelyje parodo, kad bent vienos grupės vidurkis statistikai reikšmingai skiriasi nuo kitų grupių vidurkių.

Norint nustatyti, tarp kurių konkrečiai grupių yra reikšmingi skirtumai, reikia naudoti post hoc daugkartinio lyginimo (Pairwise Multiple Comparisons) procedūrą. Populiariausi daugialypio palyginimo metodai yra Bonferoni (jautresnis už Tukey, jei palyginamų grupių skaičius mažas) ir Tukey (jautresnis už Bonferoni, jei palyginamų grupių skaičius didelis) kriterijai, su sąlyga, kad variacijos grupėse yra lygios (tikrinome Levene testu), o jeigu variacijos respondentų grupėse nelygios, tada reiktų naudoti Games-Howell, Tamhane's T2, Dunnett's T3, Dunnett's C metodus (kriterijus).

Sprendimas su SPSS 15.0

Multiple Comparisons

Dependent Variable: Pasitikėjimas demokratinės valstybės institucijomis

	(I) Vietovė	(J) Vietovė	Mean Difference (I-J)	Std. Error	Sig.	95% Confidence Interval	
						Lower Bound	Upper Bound
Tukey HSD	Vilnius	Kaunas, Klaipeda, Siauliai, Panevezys	,08333179	,09154881	,893	-,1668046	,3334682
		Rajono centras	-,15368020	,09192589	,452	-,4048469	,0974865
		Kiti miestai	-,43223406	,16055788	,056	-,8709222	,0064540
		Kaimo vietove	-,05546374	,08830977	,971	-,2967502	,1858227
	Kaunas, Klaipeda, Siauliai, Panevezys	Vilnius	-,08333179	,09154881	,893	-,3334682	,1668046
		Rajono centras	-,23701199*	,08603907	,047	-,4720943	-,0019297
		Kiti miestai	-,51556585*	,15726151	,009	-,9452474	-,0858843
		Kaimo vietove	-,13879554	,08216426	,441	-,3632908	,0856997
	Rajono centras	Vilnius	,15368020	,09192589	,452	-,0974865	,4048469
		Kaunas, Klaipeda, Siauliai, Panevezys	,23701199*	,08603907	,047	,0019297	,4720943
		Kiti miestai	-,27855386	,15748133	,392	-,7088360	,1517283
		Kaimo vietove	,09821645	,08258420	,758	-,1274262	,3238591
	Kiti miestai	Vilnius	,43223406	,16055788	,056	-,0064540	,8709222
		Kaunas, Klaipeda, Siauliai, Panevezys	,51556585*	,15726151	,009	,0858843	,9452474
		Rajono centras	,27855386	,15748133	,392	-,1517283	,7088360
		Kaimo vietove	,37677032	,15539824	,110	-,0478202	,8013609
	Kaimo vietove	Vilnius	,05546374	,08830977	,971	-,1858227	,2967502
		Kaunas, Klaipeda, Siauliai, Panevezys	,13879554	,08216426	,441	-,0856997	,3632908
		Rajono centras	-,09821645	,08258420	,758	-,3238591	,1274262
		Kiti miestai	-,37677032	,15539824	,110	-,8013609	,0478202
Bonferroni	Vilnius	Kaunas, Klaipeda, Siauliai, Panevezys	,08333179	,09154881	1,000	-,1741646	,3408282
		Rajono centras	-,15368020	,09192589	,948	-,4122372	,1048768
		Kiti miestai	-,43223406	,16055788	,072	-,8838301	,0193620
		Kaimo vietove	-,05546374	,08830977	1,000	-,3038498	,1929223
	Kaunas, Klaipeda, Siauliai, Panevezys	Vilnius	-,08333179	,09154881	1,000	-,3408282	,1741646
		Rajono centras	-,23701199	,08603907	,060	-,4790113	,0049874
		Kiti miestai	-,51556585*	,15726151	,011	-,9578903	-,0732414
		Kaimo vietove	-,13879554	,08216426	,915	-,3698963	,0923053
	Rajono centras	Vilnius	,15368020	,09192589	,948	-,1048768	,4122372
		Kaunas, Klaipeda, Siauliai, Panevezys	,23701199	,08603907	,060	-,0049874	,4790113
		Kiti miestai	-,27855386	,15748133	,772	-,7214966	,1643888
		Kaimo vietove	,09821645	,08258420	1,000	-,1340655	,3304984
	Kiti miestai	Vilnius	,43223406	,16055788	,072	-,0193620	,8838301
		Kaunas, Klaipeda, Siauliai, Panevezys	,51556585*	,15726151	,011	,0732414	,9578903
		Rajono centras	,27855386	,15748133	,772	-,1643888	,7214966
		Kaimo vietove	,37677032	,15539824	,155	-,0603134	,8138540
	Kaimo vietove	Vilnius	,05546374	,08830977	1,000	-,1929223	,3038498
		Kaunas, Klaipeda, Siauliai, Panevezys	,13879554	,08216426	,915	-,0923053	,3698963
		Rajono centras	-,09821645	,08258420	1,000	-,3304984	,1340655
		Kiti miestai	-,37677032	,15539824	,155	-,8138540	,0603134

*. The mean difference is significant at the .05 level.

Lentelėje „Daugkartiniai lyginimai“ („Multiple Comparisons“) kiekviena respondentų grupė lyginama su visomis kitomis. Svarbiausia šioje lentelėje yra atkreipti dėmesį į Sig. stulpelio reikšmes, nes jeigu jos yra mažesnės negu 0,05, tada nagrinėjamas skirtumas yra statistiškai reikšmingas. Skirtumo dydis nurodytas stulpelyje „Mean Difference (I-J)“ (Vidurkių skirtumas, iš I atėmus J). Nagrinėjant šiuos stulpelius, matome, kad didžiuosiuose miestuose (Kaune, Klaipėdoje, Šiauliuose, Panevėžyje) statistiškai reikšmingai mažiau pasitikima demokratinėmis institucijomis negu rajonų centruose ir kituose miestuose.

Pastebime, kad Bonferoni ir Tukey metodais palygintų vidurkių skirtumų reikšmingumas skiriasi: tarp

Sprendimas su SPSS 15.0

rajonų centrų ir didžiųjų miestų (išskyrus Vilnių) respondentų skirtumo nėra, remiantis Bonferoni kriterijumi, o remiantis Tukey HSD – yra.

Pasitikėjimas demokratinės valstybės institucijomis

Vietovė	N	Subset for alpha = .05	
		1	2
Tukey HSD ^{a,b} Kaunas, Klaipėda, Šiauliai, Panevėžys	270	-,1339558	
Vilnius	210	-,0506240	
Kaimo vietovė	321	,0048397	
Rajono centras	265	,1030562	,1030562
Kiti miestai	47		,3816100
Sig.		,283	,142

Means for groups in homogeneous subsets are displayed.

a. Uses Harmonic Mean Sample Size = 136,496.

b. The group sizes are unequal. The harmonic mean of the group sizes is used. Type I error levels are not guaranteed.

Homogeniškų grupių (respondentų grupių, kurių vidurkiai tarpusavyje nesiskiria) identifikavimas galimas pasirinkus Tukey ir kai kuriuos kitus aposteriorinius (Post Hoc) kriterijus (Bonferoni metodui homogeniškos grupės nėra identifikuojamos). Atskirų grupių palyginimo rezultatai yra tikslesni ir suteikia daugiau informacijos negu homogeniškų grupių identifikavimas, kuris pateikia kiek „grubesnę“ vaizdą, nes juo remiantis rajonų centrų respondentų vertinimai nesiskiria nuo Kauno, Klaipėdos, Šiaulių ir Panevėžio respondentų vertinimų (jie priskirti vienai grupei), tačiau daugkartinių lyginimų Tukey HSD kriterijus parodė, kad tarp jų skirtumas yra.

5.2.2. Dispersinė analizė su Stata

Sprendimas su Stata SE 10.0

Parengiamieji veiksmai: prielaidų tikrinimas

Tikriname, ar skirstiniai priklausomų kintamųjų **pasidem** ir **pasitrad** skirstiniai suderinami su normaliuoju (žr. 3 sk. apie Kolmogorovo-Smirnovo suderinamumo kriterijaus taikymą). Gauname, kad nulinė hipotezė apie **pasidem** ir **pasitrad** skirstinių suderinamumą su normaliuoju skirstiniu neatmetama ($p > 0,05$).

Patikriname, naudodami Levene kriterijų, ar priklausomų kintamųjų dispersija nepriklausomo kintamojo grupėse yra vienoda:

Sprendimas su Stata SE 10.0

Statistics User Window Help

Summaries, tables, and tests

Linear models and related

Binary outcomes

Ordinal outcomes

Categorical outcomes

Count outcomes

Exact statistics

Endogenous covariates

Sample selection models

Multilevel mixed-effects models

Generalized linear models

Nonparametric analysis

Time series

Multivariate time series

Longitudinal/panel data

Survival analysis

Epidemiology and related

Survey data analysis

Multivariate analysis

Summary and descriptive statistics

Tables

Classical tests of hypotheses

Nonparametric tests of hypotheses

Distributional plots and tests

One-sample mean-comparison test

Mean-comparison test, paired data

Two-sample mean-comparison test

Two-group mean-comparison test

One-sample mean-comparison calculator

Two-sample mean-comparison calculator

Binomial probability test

Binomial probability test calculator

One-sample proportion test

Two-sample proportion test

Two-group proportion test

One-sample proportion calculator

Two-sample proportion calculator

One-sample variance-comparison test

Two-sample variance-comparison test

Two-group variance-comparison test

One-sample variance-comparison calculator

Two-sample variance-comparison calculator

Robust equal variance test

	P-value	Corrected
0000	1.000	
9602	0.000	
9602	0.000	0.000

aset;
ue values out of 1053 obs

rmal(pasidem)

mirnov test against theor
em)

	P-value	Corrected
0275	0.186	
0130	0.687	
0275	0.369	0.355

robvar - Robust tests for equality of variance

Main by/if/in

Variable:
pasidem

Variable defining comparison groups:
vietove

OK Cancel Submit

robvar - Robust tests for equality of variance

Main by/if/in

Variable:
pasitrad

Variable defining comparison groups:
vietove

OK Cancel Submit

Sintaksės komandos

robvar pasidem, by(vietove)

robvar pasitrad, by(vietove)

Gauname tokius rezultatus:

Sprendimas su Stata SE 10.0

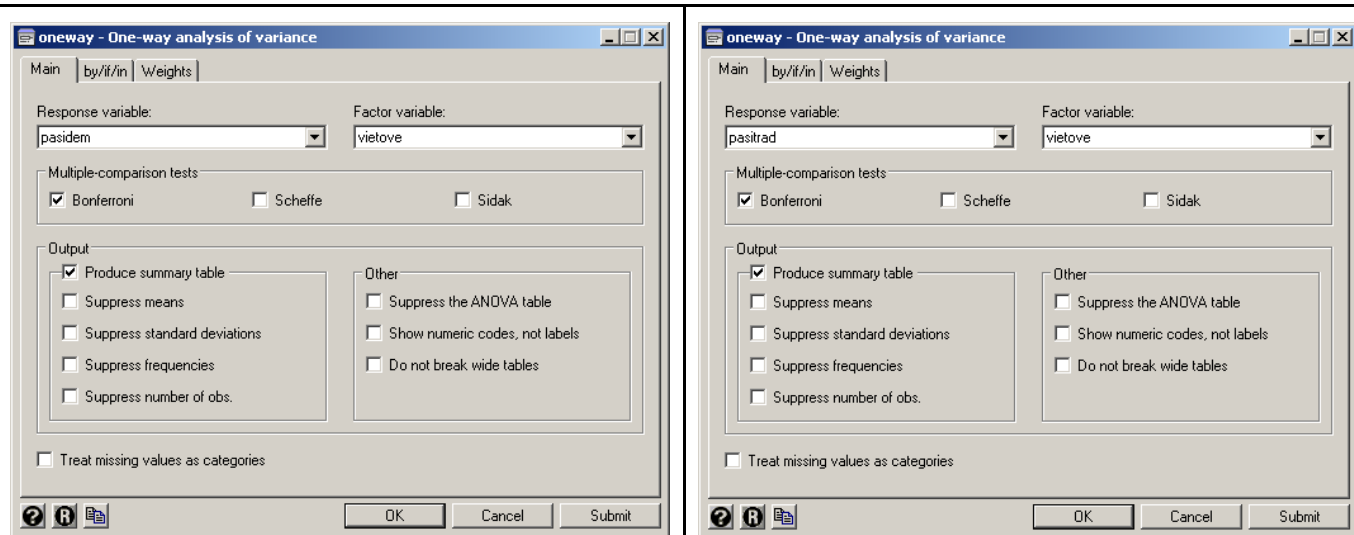
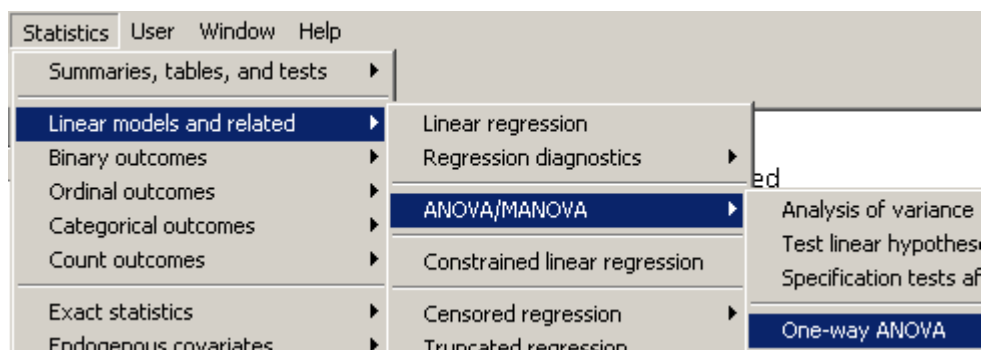
Summary of pasitikėjimas demokratinės valstybės institucijomis				Summary of pasitikėjimas tradicinio autoriteto institucijomis			
vietovė	Mean	Std. Dev.	Freq.	vietovė	Mean	Std. Dev.	Freq.
vilnius	-.05062402	1.0462417	210	vilnius	-.05979963	1.0221537	210
kaunas, k	-.13395581	1.0408779	270	kaunas, k	.01397425	1.0493395	270
raiono ce	.10305618	.90217421	265	raiono ce	-.1172733	.97691705	265
kiti mies	.38161004	.90587085	47	kiti mies	.1993648	1.0634686	47
kaimo vie	.00483973	1.0064024	321	kaimo vie	.09499113	.94248275	321
Total	-1.978e-10	1	1113	Total	-1.624e-10	1	1113
w0 = 2.2239764	df(4, 1108)	Pr > F = 0.0644751		w0 = 0.67320947	df(4, 1108)	Pr > F = 0.61061581	
w50 = 2.2382538	df(4, 1108)	Pr > F = 0.06299975		w50 = 0.59021785	df(4, 1108)	Pr > F = 0.66977931	
w10 = 2.2234285	df(4, 1108)	Pr > F = 0.06453236		w10 = 0.64567141	df(4, 1108)	Pr > F = 0.63001379	

W0 –Levene kriterijaus reikšmė, W50 – Levene kriterijaus, kurį skaičiuojant vidurkis pakeičiamas mediana, reikšmė, W10 - Levene kriterijaus, kurį skaičiuojant vidurkis pakeičiamas vidurkiu be 10% išskirčių (“10% trimmed mean”), reikšmė. Tiek **pasidem**, tiek ir **pasitrad** kintamojo atveju nulinė hipotezė apie dispersijų grupėse lygybę neatmetama ($p>0,05$) su 95% statistine garantija.

Dispersijų lygybę grupėse taip pat įvertinama dispersinės analizės metu, naudojant Bartelto testą (žr. žemiau)

Vienfaktorinės dispersinės analizės veiksmai

Meniu punktai: Statistics → Linear models and related → ANOVA/MANOVA → One-way ANOVA



Sprendimas su Stata SE 10.0

Sintaksės komandos

oneway pasidem vietove, bonferroni tabulate	oneway pasitrad vietove, bonferroni tabulate
--	---

Gauname tokius rezultatus:

Summary of pasitikėjimas demokratinės valstybės institucijomis				Summary of pasitikėjimas tradicinio autoriteto institucijomis			
vietovė	Mean	Std. Dev.	Freq.	vietovė	Mean	Std. Dev.	Freq.
vilnius	-.05062402	1.0462417	210	vilnius	-.05979963	1.0221537	210
kaunas, k	-.13395581	1.0408779	270	kaunas, k	.01397425	1.0493395	270
rajono ce	.10305618	.90217421	265	rajono ce	-.1172733	.97691705	265
kiti mies	.38161004	.90587085	47	kiti mies	.1993648	1.0634686	47
kaimo vie	.00483973	1.0064024	321	kaimo vie	.09499113	.94248275	321
Total	-1.978e-10	1	1113	Total	-1.624e-10	1	1113

Source	SS	df	MS	F	Prob > F
Between groups	15.049513	4	3.76237826	3.80	0.0045
Within groups	1096.95049	1108	.990027515		
Total	1112	1112	1		

Bartlett's test for equal variances: $\chi^2(4) = 7.7882$ Prob> $\chi^2 = 0.100$

Source	SS	df	MS	F	Prob > F
Between groups	9.21279827	4	2.30319957	2.31	0.0557
Within groups	1102.7872	1108	.995295311		
Total	1112	1112	1		

Bartlett's test for equal variances: $\chi^2(4) = 4.2737$ Prob> $\chi^2 = 0.370$

Comparison of pasitikėjimas demokratinės valstybės institucijomis by vietovė (Bonferroni)					Comparison of pasitikėjimas tradicinio autoriteto institucijomis by vietovė (Bonferroni)				
Row Mean- Col Mean	vilnius	kaunas,	rajono c	kiti mie	Row Mean- Col Mean	vilnius	kaunas,	rajono c	kiti mie
kaunas,	-.083332 1.000				.073774 1.000				
rajono c	.15368 0.948	.237012 0.060			-.057474 1.000	-.131248 1.000			
kiti mie	.432234 0.072	.515566 0.011	.278554 0.772		.259164 1.000	.185391 1.000	.316638 0.452		
kaimo vi	.055464 1.000	.138796 0.915	-.098216 1.000	-.37677 0.155	.154791 0.807	.081017 1.000	.212264 0.105	-.104374 1.000	

Pirmoji lentelė pateikia vidurkius grupėse, nuokrypius ir grupės dydį. Antrojoje pateikiamas nulinės vidurkių lygybės hipotezės tikrinimo rezultatas: gauname, kad **pasidem** kintamojo atveju nulinė hipotezė apie vienodus vidurkius atmetama ($p < 0,01$), o **pasitrad** kintamojo atveju nulinė hipotezė apie vienodus vidurkius neatmetama ($p > 0,05$). Taip pat, antrojoje lentelėje pateikiama dispersijos priklausomo kintamojo grupėse lygybės (nulinės) hipotezės tikrinimo rezultatai: **pasidem** ($p = 0,100 > 0,05$) ir **pasitrad** ($p = 0,370 > 0,05$) atveju nulinė hipotezė priimama su 95% statistine garantija.

Trečiojoje lentelėje pateikiamas visų grupių porų vidurkių skirtumai ir tų skirtumų statistinis reikšmingumas. Remiantis Bonferoni kriterijumi, vienintelis statistiškai reikšmingas ($p < 0,05$) skirtumas yra tarp Kauno (ir kitų didžiųjų miestų, išskyrus Vilnių) ir „kiti miestai“ grupių, demokratinės valstybės institucijomis daugiau pasitikima miestuose, priskirtuose „kitų miestų“ kategorijai.

5.3. Neparametriniai kriterijai

Neparametriniai kriterijai (dviem ir daugiau imtims/grupėms) labai praverčia tais atvejais, kada netenkinamos priklausomo kintamojo normalumo ir dispersijų lygybės grupėse sąlygos. Pavyzdžiui naudosime 2004 m. Naujojo Baltijos Barometro duomenis.

5.3.1. Mano-Vitnio-Vilkoksono rangų sumų kriterijus dviems nepriklausomoms imtims su SPSS

Mano-Vitnio-Vilkoksono rangų sumų kriterijus dviem nepriklausomoms imtims ($k=2$) yra neparametrinė vidurkių palyginimo (t-testo) analizės alternatyva, taikoma, kai priklausomas kintamasis netenkina normalumo sąlygos.

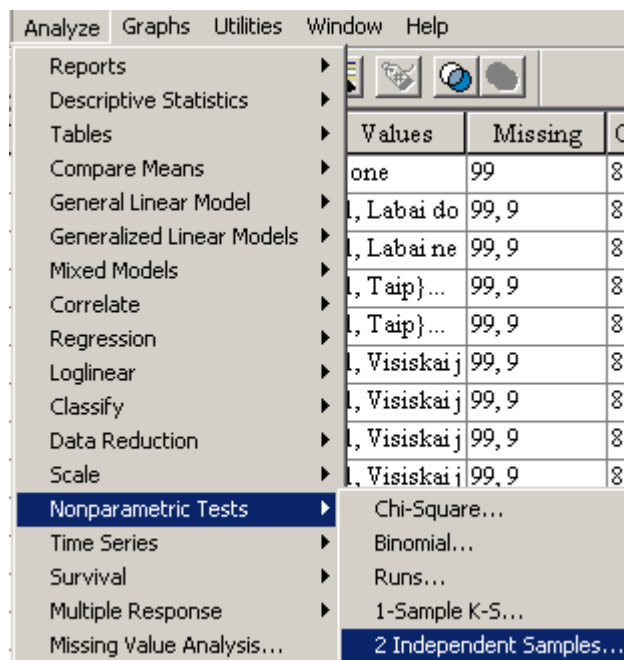
Sprendimas su SPSS 15.0

Parengiamieji veiksmai

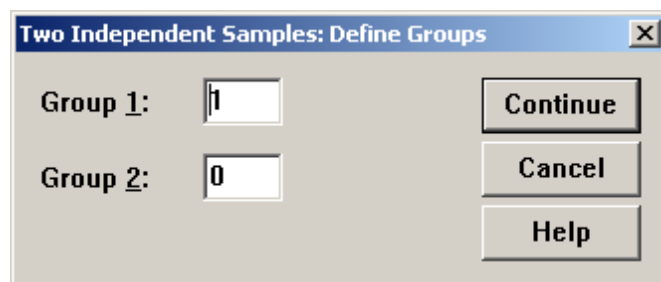
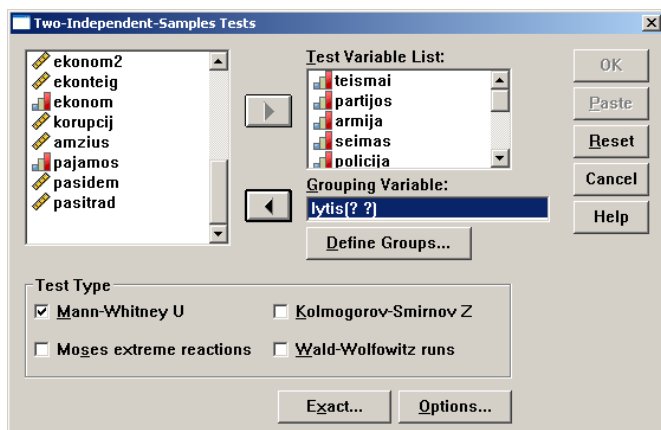
Nėra

Analizės veiksmai

Meniu punktai (Mano-Vitnio-Vilkoksono rangų sumų kriterijus dviems nepriklausomoms imtims): Analyze → Nonparametric tests → 2-Independent-Samples... → Perkeliame iš kairės pusės pasitikėjimo institucijomis kintamuosius **teismai - baznycia** į priklausomų kintamųjų sąrašą („Test Variable List“); iš kairės pusės perkeliame grupuojantį kintamąjį **lytis** („Grouping Variable“); spaudžiame „Define Groups...“ ir įrašome lyties grupių kodus, kokie jie yra duomenų faile: „1“ (t.y., vyrai) ir „0“ (t.y., moterys) → Continue → OK.



Sprendimas su SPSS 15.0



Sintaksės komandos

NPART TESTS

/M-W= teismai partijos armija seimas policija prezid profsaj baznycia

BY lytis(1 0)

/MISSING ANALYSIS.

Gauname tokius rezultatus **rezultatų išvesties lange:**

Ranks

	lytis	N	Mean Rank	Sum of Ranks
Kiek jūs pasitikite teismais?	0 Moteris	574	538,26	308959,50
	1 Vyras	481	515,76	248080,50
	Total	1055		
Kiek jūs pasitikite politinėmis partijomis?	0 Moteris	581	544,35	316266,00
	1 Vyras	485	520,51	252445,00
	Total	1066		
Kiek jūs pasitikite kariuomene?	0 Moteris	578	535,13	309305,50
	1 Vyras	479	521,60	249847,50
	Total	1057		
Kiek jūs pasitikite Seimu?	0 Moteris	592	552,33	326981,00
	1 Vyras	488	526,15	256759,00
	Total	1080		
Kiek jūs pasitikite policija?	0 Moteris	603	559,93	337638,00
	1 Vyras	493	534,52	263518,00
	Total	1096		
Kiek jūs pasitikite Prezidentu?	0 Moteris	593	557,06	330336,00
	1 Vyras	487	520,34	253404,00
	Total	1080		
Kiek jūs pasitikite profsajungomis?	0 Moteris	487	478,82	233186,00
	1 Vyras	401	402,82	161530,00
	Total	888		
Kiek jūs pasitikite Bažnyčia?	0 Moteris	599	556,27	333205,00
	1 Vyras	479	518,53	248376,00
	Total	1078		

Sprendimas su SPSS 15.0

Test Statistics^a

	Kiek jūs pasitikite teismais?	Kiek jūs pasitikite politinėmis partijomis?	Kiek jūs pasitikite kariuomene?	Kiek jūs pasitikite Seimu?	Kiek jūs pasitikite policija?	Kiek jūs pasitikite Prezidentu?	Kiek jūs pasitikite profsąjungomis?	Kiek jūs pasitikite Bažnyčia?
Mann-Whitney U	132159,500	134590,000	134887,500	137443,000	141747,000	134576,000	80929,000	133416,000
Wilcoxon W	248080,500	252445,000	249847,500	256759,000	263518,000	253404,000	161530,000	248376,000
Z	-1,219	-1,290	-,732	-1,400	-1,342	-1,960	-4,466	-2,019
Asymp. Sig. (2-tailed)	,223	,197	,464	,161	,179	,050	,000	,043

a. Grouping Variable: Lytis

Pirmoje lentelėje pateikti atvejų skaičius kiekvienoje grupėje, vidutinis rangas (neparametrinis aritmetinio vidurkio pakaitalas) ir rangų suma.

Antroje lentelėje pateikti reikšmingumai parodo, kad tik pasitikėjimo profsąjungomis ir bažnyčia kintamuosiuose moterų ir vyrų vertinimai skiriasi. Remdamiesi pirmojoje lentelėje pateiktais rangais, galime teigti, kad moterys statistiškai reikšmingai (su 95% statistine garantija) daugiau pasitiki profsąjungomis ir bažnyčia negu vyrai. Pasitikėjimo prezidentu atvejis yra ant statistinio reikšmingumo ribos (t.y., $p=0,05$), tačiau pagal priimtą susitarimą reiktų jį laikyti statistiškai nereikšmingu.

5.3.2. Mano-Vitnio-Vilkoksono rangų sumų kriterijus dviems nepriklausomoms imtims su Stata

Sprendimas su STATA 10.0

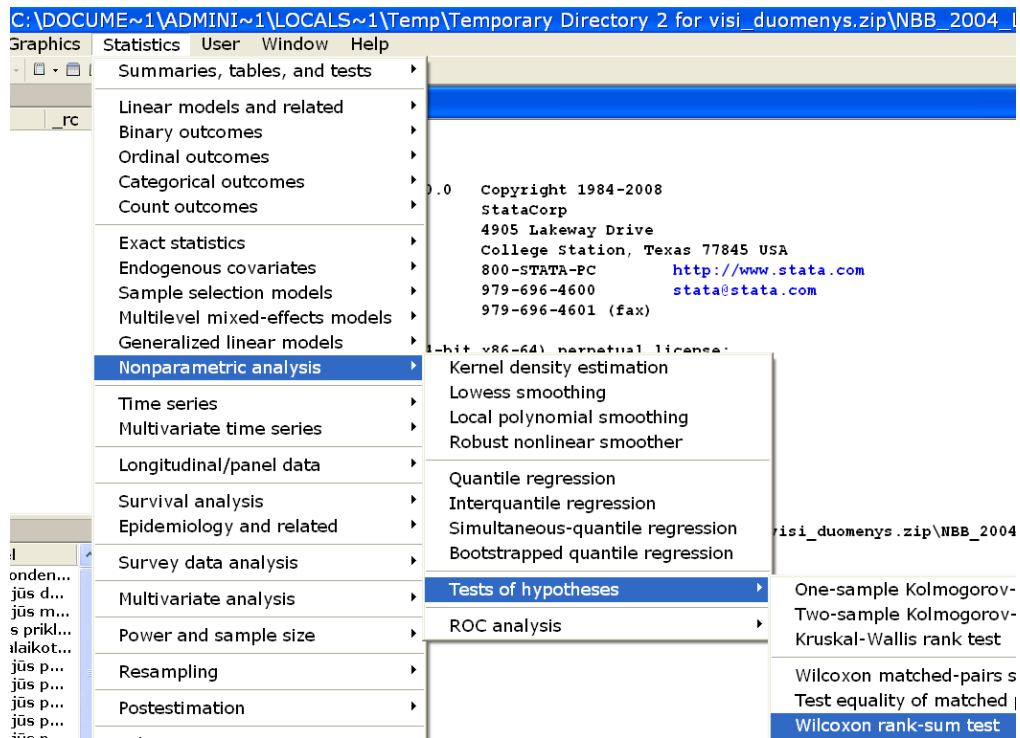
Parengiamieji veiksmai

Nėra

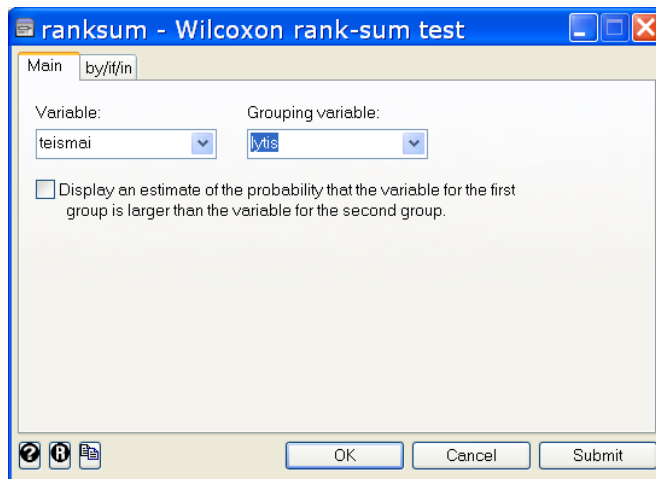
Analizės veiksmai

Meniu punktai: Analyze → Nonparametric analysis → Tests of hypotheses... → Wilcoxon rank-sum test

Sprendimas su STATA 10.0



Parenkame analizės kintamuosius (analizės atliekamos analogiškai, tik įkeliamas atitinkamas kintamasis) ties „Variable“ **teismai** – **baznycia**, o ties „Grouping variable“ parenkame grupuojantį kintamąjį **lytis** ir spaudžiame OK.



Sintaksės komandos

```
ranksum teismai, by(lytis)
ranksum partijos, by(lytis)
ranksum armija, by(lytis)
ranksum seimas, by(lytis)
ranksum policija, by(lytis)
ranksum prezid, by(lytis)
ranksum profsaj, by(lytis)
ranksum baznycia, by(lytis)
```

Gauname tokius rezultatus:

Sprendimas su STATA 10.0

Two-sample wilcoxon rank-sum (Mann-whitney) test

lytis	obs	rank sum	expected
moteris	574	308959.5	303072
vyras	481	248080.5	253968
combined	1055	557040	557040

unadjusted variance 24296272
adjustment for ties -953369.49

adjusted variance 23342903

Ho: teismai(lytis==moteris) = teismai(lytis==vyras)
z = 1.219
Prob > |z| = 0.2230

Rangų skirtumai grupėse nereikšmingi
($p=0,223>0,05$).

Two-sample wilcoxon rank-sum (Mann-whitney) test

lytis	obs	rank sum	expected
moteris	581	316266	309963.5
vyras	485	252445	258747.5
combined	1066	568711	568711

unadjusted variance 25055383
adjustment for ties -1174162.5

adjusted variance 23881220

Ho: partijos(lytis==moteris) = partijos(lytis==vyras)
z = 1.290
Prob > |z| = 0.1972

Rangų skirtumai grupėse nereikšmingi
($p=0,197>0,05$).

Two-sample wilcoxon rank-sum (Mann-whitney) test

lytis	obs	rank sum	expected
moteris	578	309305.5	305762
vyras	479	249847.5	253391
combined	1057	559153	559153

unadjusted variance 24410000
adjustment for ties -972388.01

adjusted variance 23437612

Ho: armija(lytis==moteris) = armija(lytis==vyras)
z = 0.732
Prob > |z| = 0.4642

Rangų skirtumai grupėse nereikšmingi
($p=0,464>0,05$).

Two-sample wilcoxon rank-sum (Mann-whitney) test

lytis	obs	rank sum	expected
moteris	592	326981	319976
vyras	488	256759	263764
combined	1080	583740	583740

unadjusted variance 26024715
adjustment for ties -996807.1

adjusted variance 25027908

Ho: seimas(lytis==moteris) = seimas(lytis==vyras)
z = 1.400
Prob > |z| = 0.1614

Rangų skirtumai grupėse nereikšmingi
($p=0,161>0,05$).

Sprendimas su STATA 10.0

Two-sample wilcoxon rank-sum (Mann-whitney) test

lytis	obs	rank sum	expected
moteris	603	337638	330745.5
vyras	493	263518	270410.5
combined	1096	601156	601156

unadjusted variance **27176255**
adjustment for ties **-815370.24**

adjusted variance **26360885**

Ho: policija(lytis==moteris) = policija(lytis==vyras)
z = **1.342**
Prob > |z| = **0.1795**

Rangų skirtumai grupėse nereikšmingi
(p=0,178>0,05).

Two-sample wilcoxon rank-sum (Mann-whitney) test

lytis	obs	rank sum	expected
moteris	593	330336	320516.5
vyras	487	253404	263223.5
combined	1080	583740	583740

unadjusted variance **26015256**
adjustment for ties **-910296.09**

adjusted variance **25104960**

Ho: prezid(lytis==moteris) = prezid(lytis==vyras)
z = **1.960**
Prob > |z| = **0.0500**

Rangų skirtumai grupėse nereikšmingi (p=0,05).

Two-sample wilcoxon rank-sum (Mann-whitney) test

lytis	obs	rank sum	expected
moteris	487	233186	216471.5
vyras	401	161530	178244.5
combined	888	394716	394716

unadjusted variance **14467512**
adjustment for ties **-460554.22**

adjusted variance **14006958**

Ho: profsaj(lytis==moteris) = profsaj(lytis==vyras)
z = **4.466**
Prob > |z| = **0.0000**

Rangų skirtumai grupėse reikšmingi (p<0,001).

Two-sample wilcoxon rank-sum (Mann-whitney) test

lytis	obs	rank sum	expected
moteris	599	333205	323160.5
vyras	479	248376	258420.5
combined	1078	581581	581581

unadjusted variance **25798980**
adjustment for ties **-1059743.7**

adjusted variance **24739236**

Ho: baznycia(lytis==moteris) = baznycia(lytis==vyras)
z = **2.019**
Prob > |z| = **0.0434**

Rangų skirtumai grupėse reikšmingi
(p=0,043<0,05).

Analizės išvados sutampa su SPSS pavyzdyje gautomis išvadomis: tik pasitikėjimo profsajungomis ir bažnyčia kintamuosiuose moterų ir vyrų vertinimai skiriasi - moterų statistiškai reikšmingai (su 95%

Sprendimas su STATA 10.0

statistine garantija) daugiau pasitiki profsajungomis ir bažnyčia negu vyrai.

5.3.3. Kraskelo-Voliso kriterijus k-nepriklausomoms imtims su SPSS

Kraskelo-Voliso kriterijus k-nepriklausomoms imtims ($k > 2$) yra neparametrinė vienfaktorinės dispersinės analizės alternatyva, taikoma, kai priklausomas kintamasis netenkina normalumo sąlygos.

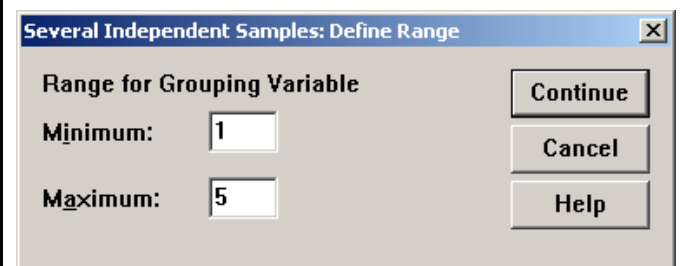
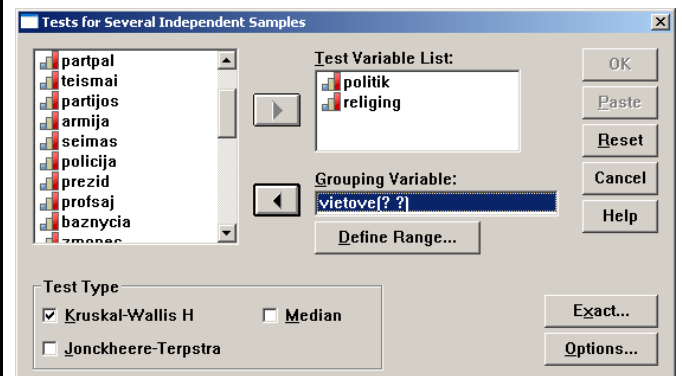
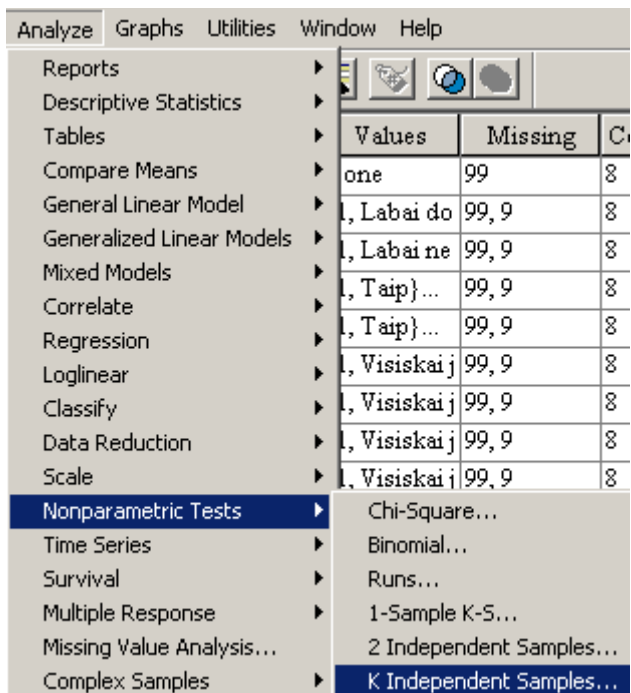
Sprendimas su SPSS 15.0

Parengiamieji veiksmai

Nėra

Analizės veiksmai

Meniu punktai (Kraskelo-Voliso kriterijus k-nepriklausomoms imtims): Analyze → Nonparametric tests → K-Independent Samples... → Pasirenkame priklausomus („Test variable List“) kintamuosius **religing**, **politik**; nurodome grupuojantį („Grouping Variable“) kintamąjį **vietove**, spaudžiame „Define Range“ ir pateikiame grupuojančio kintamojo reikšmių intervalą nuo 1 iki 5 → OK.



Sintaksės komandos

NPARTESTS
/K-W=politik religing BY vietove(1 5)
/MISSING ANALYSIS.

Gauname tokius rezultatus:

Sprendimas su SPSS 15.0

Ranks		
Vietovė	N	Mean Rank
Kiek jūs domitės politika?		
Vilnius	207	571,76
Kaunas, Klaipėda, Šiauliai, Panevėžys	268	502,40
Rajono centras	258	552,05
Kiti miestai	47	512,45
Kaimo vietovė	313	569,85
Total	1093	
Kaip dažnai jūs einate į bažnyčią?		
Vilnius	198	493,95
Kaunas, Klaipėda, Šiauliai, Panevėžys	253	545,74
Rajono centras	252	547,94
Kiti miestai	46	592,90
Kaimo vietovė	310	516,26
Total	1059	

Pirmojoje, rangų lentelėje gauname, kad didžiausias religingumo rangas yra kitų miestų kategorijoje (592,9), kiek mažesnis rajonų centruose, toliau didžiuosiuose miestuose, išskyrus Vilnių, kaimo vietovėse ir, galų gale, mažiausias religingumas stebimas Vilniuje (vidutinis rangas 493,95). Didžiausias domėjimasis politika yra Vilniuje (vidutinis rangas 571,76), toliau seka kaimo vietovės, rajonų centrai, kiti miestai, o mažiausias rangas yra didžiuosiuose, išskyrus Vilnių, miestuose (502,4). Dažniausiai patogiausiai pateikti aritmetinius vidurkius, nes juos galima paprasčiau interpretuoti (pvz., 3,00 vidurkis reikštų, kad vidutiniškai respondentai eina į bažnyčią keletą kartų per metus).

Test Statistics^{a,b}

	Kiek jūs domitės politika?	Kaip dažnai jūs einate į bažnyčią?
Chi-Square	10,712	7,572
df	4	4
Asymp. Sig.	,030	,109

a. Kruskal Wallis Test

b. Grouping Variable: Vietovė

Antroje lentelėje pateikti Kraskelo-Voliso kriterijaus reikšmingumo lygmenys rodo, kad religingumas nesiskiria tarp vietovių grupių ($p=0,109>0,05$), tuo tarpu bent dviejose vietovių grupėse respondentų susidomėjimo politika intensyvumas statistiškai reikšmingai skiriasi ($p=0,030<0,05$). Norėdami iširti, kokiose grupėse šie skirtumai yra reikšmingi, turėtume taikyti neparametrinį Mano-Vitnio-Vilkoksono rangų sumų kriterijų dviems nepriklausomoms imtims kiekvienai grupių porai, kadangi daugkartinių lyginimų dispersinėje analizėje analogo neparametrinėje versijoje SPSS nėra.

5.3.4. Kraskelo-Voliso kriterijus k-nepriklausomoms imtims su Stata

Sprendimas su STATA 10.0

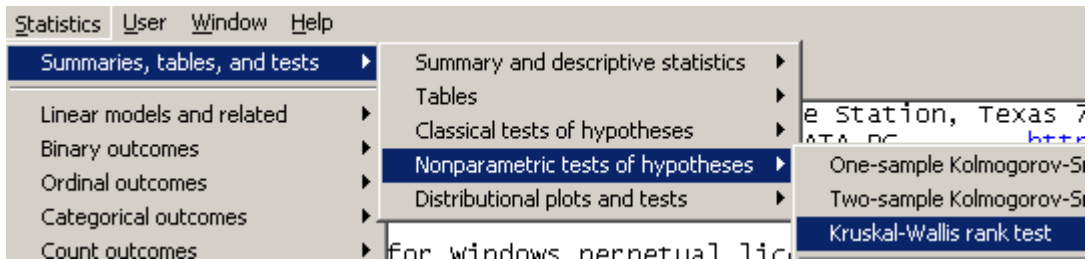
Parengiamieji veiksmai

Nėra

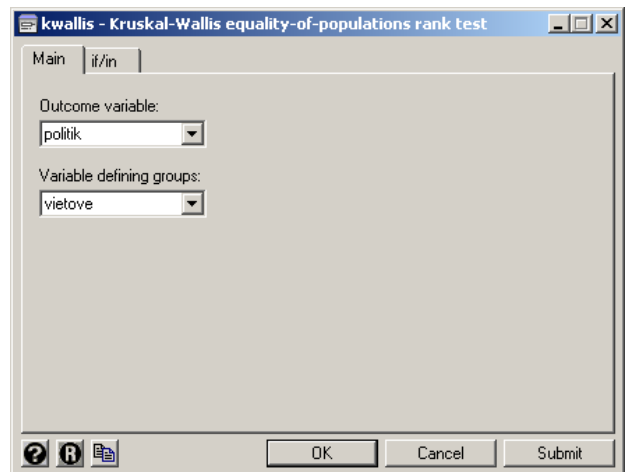
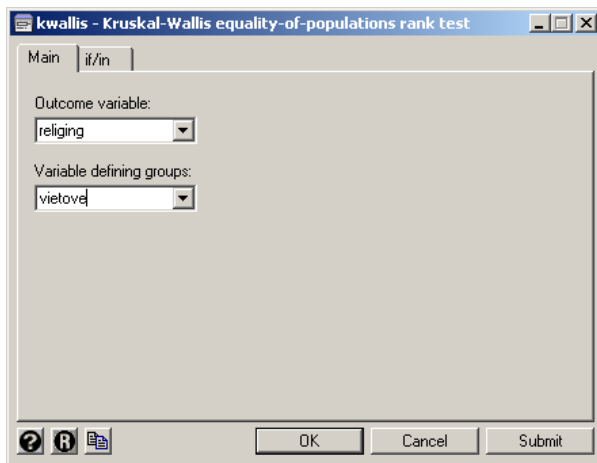
Sprendimas su STATA 10.0

Analizės veiksmai

Meniu punktai (Kraskelo-Voliso rangų sumų kriterijus dviems nepriklausomoms imtims): Statistics → Summaries, tables, & tests → Nonparametric tests of hypothesis → Kraskel-Wallis rank test



Dialogo lange pasirenkame priklausoma kintamąjį (“Outcome variable”) **religing**, nurodome grupuojantį kintamąjį (“Variable defining groups”) **vietove**, spaudžiame „Submit“.



Sintaksės komandos

```
kwallis religing, by(vietove)
```

```
kwallis politik, by(vietove)
```

Gauname tokius rezultatus:

kruskal-wallis equality-of-populations rank test

vietove	obs	Rank Sum
vilnius	198	97802.00
kaunas,	253	138072.00
rajono c	252	138081.50
kiti mie	46	27273.50
kaimo vi	310	160041.00

chi-squared = 6.859 with 4 d.f.
probability = 0.1435

chi-squared with ties = 7.572 with 4 d.f.
probability = 0.1086

kruskal-wallis equality-of-populations rank test

vietove	obs	Rank Sum
vilnius	207	118354.00
kaunas,	268	134642.50
rajono c	258	142428.00
kiti mie	47	24085.00
kaimo vi	313	178361.50

chi-squared = 8.892 with 4 d.f.
probability = 0.0639

chi-squared with ties = 10.712 with 4 d.f.
probability = 0.0300

Lentelėje pateikiami grupių vardai, stebinių skaičius ir rangų suma. Po lentelę pateikiami: chi-kvadrato kriterijaus reikšmė ir jos reikšmingumas; chi-kvadrato, pataisyto atsižvelgiant į surištas poras, kriterijaus reikšmė ir jos reikšmingumas.

Sprendimas su STATA 10.0	
Išvada: nulinė hipotezė apie rangų lygybę grupėse priimama ($p=0,109>0,05$)	Išvada: nulinė hipotezė apie rangų lygybę grupėse atmetama ($p=0,030<0,05$)
Rezultatų interpretacija sutampa su pateiktąja analogiškame SPSS pavyzdyje.	

6. POŽYMIŲ PRIKLAUSOMUMO TYRIMAS. KORELIACINĖ ANALIZĖ

Šiame skyriuje pateikiami pavyzdžiai, kaip su SPSS ir Stata atlikti požymių priklausomumo ir koreliacinę analizę.

6.1. Intervalų skalės koreliacinė analizė

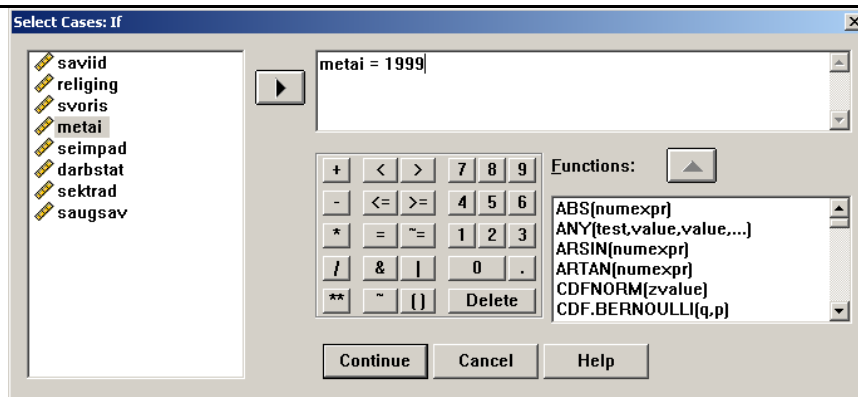
EVT pateikiami išlikimo vs. saviraiškos vertybių ir tradicinės vs. sekuliosios vertybių indeksai (**sektrad** ir **saugsav** kintamieji, atitinkamai). Patyrinėkime, koku būdu šie du kintamieji susiję Lietuvos gyventojų mąstyme 1999 m.: ar kuo didesnė saviraiškos orientacija, tuo didesnė ir sekuliosioji orientacija, ar atvirkščiai, saviraiškos ir sekuliosioji orientacijos yra susijusios priešingai, o gal tarp jų ryšio iš viso nėra? Tokiam tyrimui, kadangi abu kintamieji yra tvarkos skalės su daug reikšmių, naudosime Pisrono ir Spirmeno koreliacijos koeficientus.

6.1.1. Intervalų skalės koreliacijos koeficientai su SPSS

Sprendimas su SPSS 15.0

Parengiamieji veiksmai

Visų pirma, žinodami, kad duomenų faile yra kelių bangų tyrimų duomenys, pasirenkame tik 1999 m. duomenis: Data → Select cases → If condition is satisfied → If...: pasirenkame kintamąjį **metai** ir perkeliame jį iš kairės į dešinę meniu pusę, dešiniajame meniu užrašome „metai = 1999“, spaudžiame „Continue“ ir „OK“.



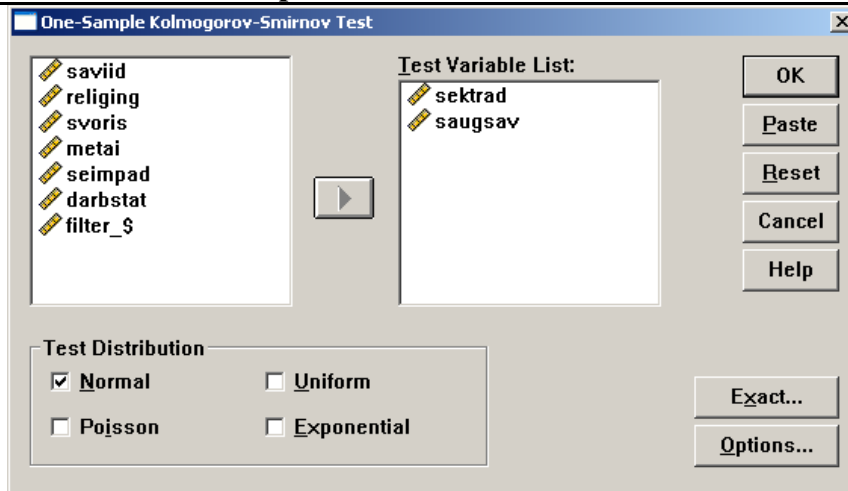
Sintaksės komandos

```
USE ALL.  
COMPUTE filter_$=(metai = 1999).  
VARIABLE LABEL filter_$ 'metai = 1999 (FILTER)'.  
VALUE LABELS filter_$ 0 'Not Selected' 1 'Selected'.  
FORMAT filter_$ (f1.0).  
FILTER BY filter_$.  
EXECUTE .
```

Patyrinę abu kintamuosius per Analyze-Frequencies įsitikintumėte, kad tai yra intervaliniai kintamieji. Kad galėtume skaičiuoti Pisrono koreliacijos koeficientą, reikia dar įsitikinti, kad abu kintamieji turi normalųjį skirstinį.

Meniu punktai: Analyze → Nonparametric Tests → 1 Sample K-S → pasirenkame **survself** ir **tradrat5**, spaudžiame „OK“.

Sprendimas su SPSS 15.0



Sintaksės komandos

NPART TESTS

/K-S(NORMAL)= sekrad saugsav

/MISSING ANALYSIS.

Gauname tokius rezultatus:

One-Sample Kolmogorov-Smirnov Test

		Sekuliarios vs. tradicinės vertybės	Išlikimo vs. saviraiškos vertybės
N		145	145
Normal Parameters ^{a,b}	Mean	,4723241	-,4662841
	Std. Deviation	,91264875	,83604775
Most Extreme Differences	Absolute	,065	,052
	Positive	,065	,052
	Negative	-,045	-,044
Kolmogorov-Smirnov Z		,785	,621
Asymp. Sig. (2-tailed)		,569	,835

a. Test distribution is Normal.

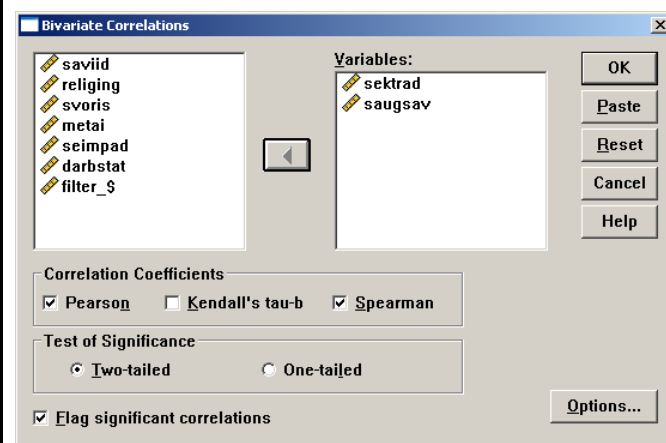
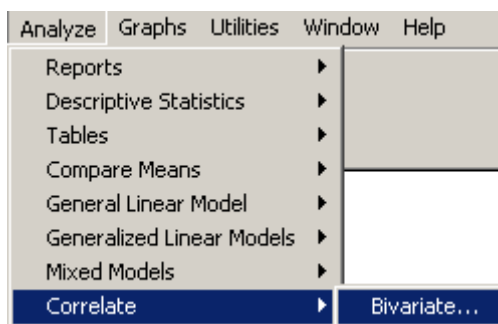
b. Calculated from data.

Asimptotinis reikšmingumo lygmuo rodo, kad rezultatai yra suderinami su nuline (t.y., normalaus skirtingo) hipoteze. Galime skaičiuoti Pirsono koreliacijos koeficientą.

Koreliacinės analizės veiksmas

Meniu punktai: Analyze → Correlate → Bivariate → pasirenkame **survself** ir **tradrat5**, uždedame varneles prie „Pearson“ ir „Spearman“, spaudžiame „OK“.

Sprendimas su SPSS 15.0



Sintaksės komandos

CORRELATIONS

```
/VARIABLES=sekttrad saugsav
/PRINT=TWOTAIL NOSIG
/MISSING=PAIRWISE .
```

NONPAR CORR

```
/VARIABLES=sekttrad saugsav
/PRINT=SPEARMAN TWOTAIL NOSIG
/MISSING=PAIRWISE .
```

Gauname tokius rezultatus:

Correlations

		Sekuliarios vs. tradicinės vertybės	Išlikimo vs. saviraiškos vertybės
Sekuliarios vs. tradicinės vertybės	Pearson Correlation	1	-,354**
	Sig. (2-tailed)		,000
	N	145	145
Išlikimo vs. saviraiškos vertybės	Pearson Correlation	-,354**	1
	Sig. (2-tailed)	,000	
	N	145	145

**. Correlation is significant at the 0.01 level (2-tailed).

Sprendimas su SPSS 15.0

Correlations

			Sekuliarios vs. tradicinės vertybės	Išlikimo vs. saviraiškos vertybės
Spearman's rho	Sekuliarios vs. tradicinės vertybės	Correlation Coefficient	1,000	-,341**
		Sig. (2-tailed)	.	,000
		N	145	145
	Išlikimo vs. saviraiškos vertybės	Correlation Coefficient	-,341**	1,000
		Sig. (2-tailed)	,000	.
		N	145	145

** . Correlation is significant at the 0.01 level (2-tailed).

Matome, kad abu kintamuosius sieja statistikai reikšmingas neigiamas ryšys: Pirsono $r = -0,354$ ($p < 0,001$), Spirmeno $r_o = -0,341$ ($p < 0,001$).

Žemesnės sekuliarių vs. tradicinių vertybių indekso vertės atitinka daugiau tradicinius principus, o aukštesni – sekuliarūs/racionalūs principus, o išlikimo/saviraiškos skalėje didesnės reikšmės žymi didesnė išlikimo vertybių reikšmę respondentui. Remiantis šiomis indeksų savybėmis, gautą statistiškai reikšmingą neigiamą koreliaciją galėtume interpretuoti taip: kuo stipresnės sekuliarios vertybės, tuo stipresnės saviraiškos vertybės *tiriamoje populiacijoje*. Šioje vietoje reiktų atkreipti dėmesį, kad tiriami 145 atvejai iš 1018, visi kiti yra trūkstamos reikšmės, t.y., iškyla abejonių dėl rezultatų statistinio apibendrinimo visai tiriamųjų populiacijai. Vienas iš sprendimo būdų būtų įsitikinti, ar šie 145 atvejai gali būti laikomi atsitiktinai parinktais iš visų apklaustųjų, jeigu taip, tada galime drąsiai sutikti su išvadomis.

Kita problema yra tai, kad šie 145 atsakymai turi būti tam tikru būdu pasverti, suteikiant kiekvienam atvejui svorius, kurie ištaisytų apklausos metu susidariusius respondentų imties iškraipymus:

Data → Weight Cases... → Weight cases by: pasirenkame kintamąjį **svoris**, kuris normuoja imties dydį į 1000 → OK.

Galite įsitikinti, kad dabar kiekvieno atskiro tyrimo imties dydis bus perskaičiuotas į 1000.

Atliekant koreliacijų skaičiavimą su pasvertais duomenimis, gauname, kad Pirsono r sumažėja iki $-0,310$ ($p < 0,001$), o Spirmeno r_o sumažėja iki $-0,271$ ($p < 0,01$).

6.1.2. Intervalų skalės koreliacijos koeficientai su Stata

Sprendimas su STATA 10.0

Parengiamieji veiksmai

Įsitikiname, kad **sektrad** ir **saugsav** kintamųjų skirstiniai yra suderinami su normaliuoju (žr. 3 sk. apie suderinamumo hipotezės tikrinimą Kolmogorovo-Smirnovo kriterijumi).

Sintaksės komandos

```
ksmirnov sektrad = normal(sektrad) if  
metai==1999
```

```
ksmirnov saugsav = normal(saugsav) if  
metai==1999
```

Gauname tokius rezultatus:

Sprendimas su STATA 10.0

One-sample Kolmogorov-Smirnov test against theoretical distribution
normal(sektrad)

Smaller group	D	P-value	Corrected
sektrad:	0.0006	1.000	
Cumulative:	-0.1870	0.000	
Combined K-S:	0.1870	0.000	0.000

Note: ties exist in dataset;
there are 140 unique values out of 145 observations.

One-sample Kolmogorov-Smirnov test against theoretical distribution
normal(saugsav)

Smaller group	D	P-value	Corrected
saugsav:	0.2338	0.000	
Cumulative:	-0.0052	0.992	
Combined K-S:	0.2338	0.000	0.000

Note: ties exist in dataset;
there are 140 unique values out of 145 observations.

Rezultatai rodo, kad kintamųjų skirstiniai nėra normalieji, Priešingai su SPSS gautiems rezultatams.

STATA paketu normalumui tikrinti galima naudoti sktest arba swilk (Šapiro-Vilk kriterijus) komandas. Pakartojus normalumo tikrinimą su komandomis “sktest saugsav if metai==1999” ir “sktest sektrad if metai==1999”, gauname, kad visgi nulinė skirstinių normalumo hipotezė neatmetama ($p > 0,05$):

Skewness/Kurtosis tests for Normality

variable	Pr(Skewness)	Pr(Kurtosis)	adj chi2(2)	joint Prob>chi2
saugsav	0.547	0.936	0.37	0.8297

Skewness/Kurtosis tests for Normality

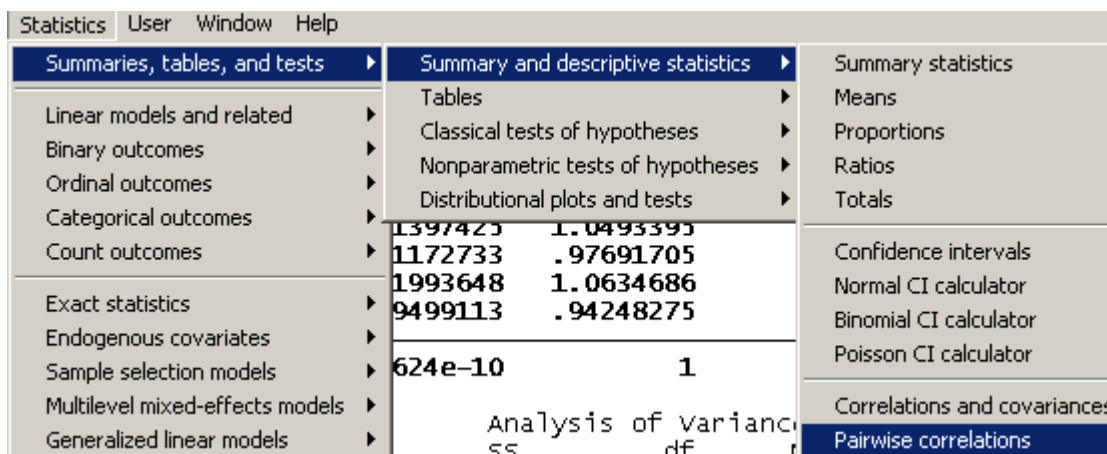
variable	Pr(Skewness)	Pr(Kurtosis)	adj chi2(2)	joint Prob>chi2
sektrad	0.018	0.954	5.48	0.0645

Išvada: **sektrad** ir **saugsav** kintamųjų skirstiniai yra suderinami su normaliuoju skirstiniu, todėl galime pagrįstai skaičiuoti Pirsono tiesinės koreliacijos koeficientą.

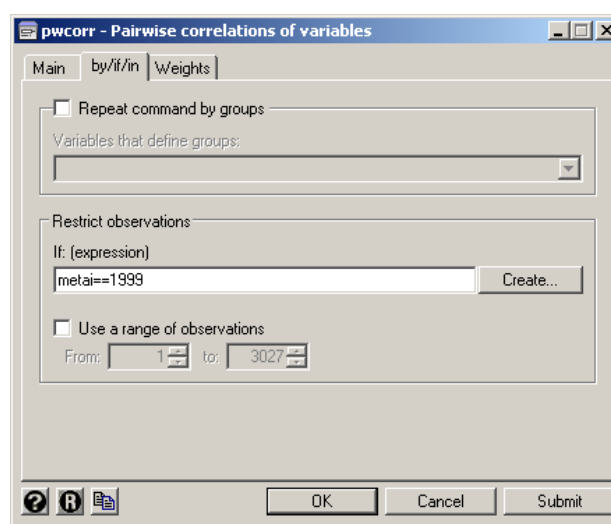
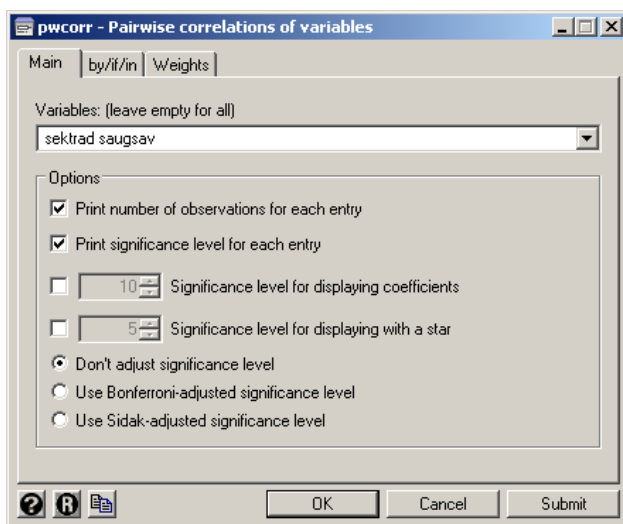
Analizės veiksmiai: Pirsono koreliacijos koeficientas

Meniu punktai: Statistics → Summary, tables, & tests → Summary statistics → Pairwise correlations

Sprendimas su STATA 10.0



Pasirenkame **sektrad** ir **saugsav**, „Main“ lange, „Options“ skiltyje pasirenkame „Print number of observations for each entry“ ir „Print significance level for each entry“. „by/if/in“ lange „Restrict observations“ skiltyje įrašome „metai == 1999“, spaudžiame „Submit“.



Sintaksės komandos

```
pwcorr sektrad saugsav if metai==1999, obs sig
```

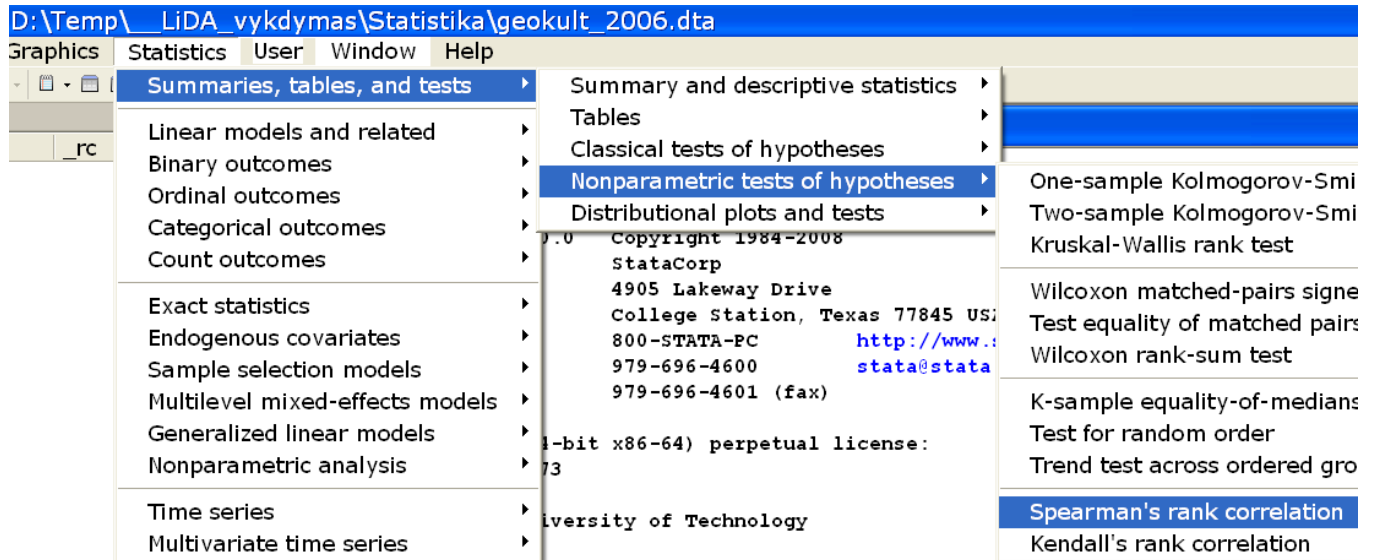
Gauname tokius rezultatus:

	sektrad	saugsav
sektrad	1.0000 145	
saugsav	-0.3541 0.0000 145	1.0000 145

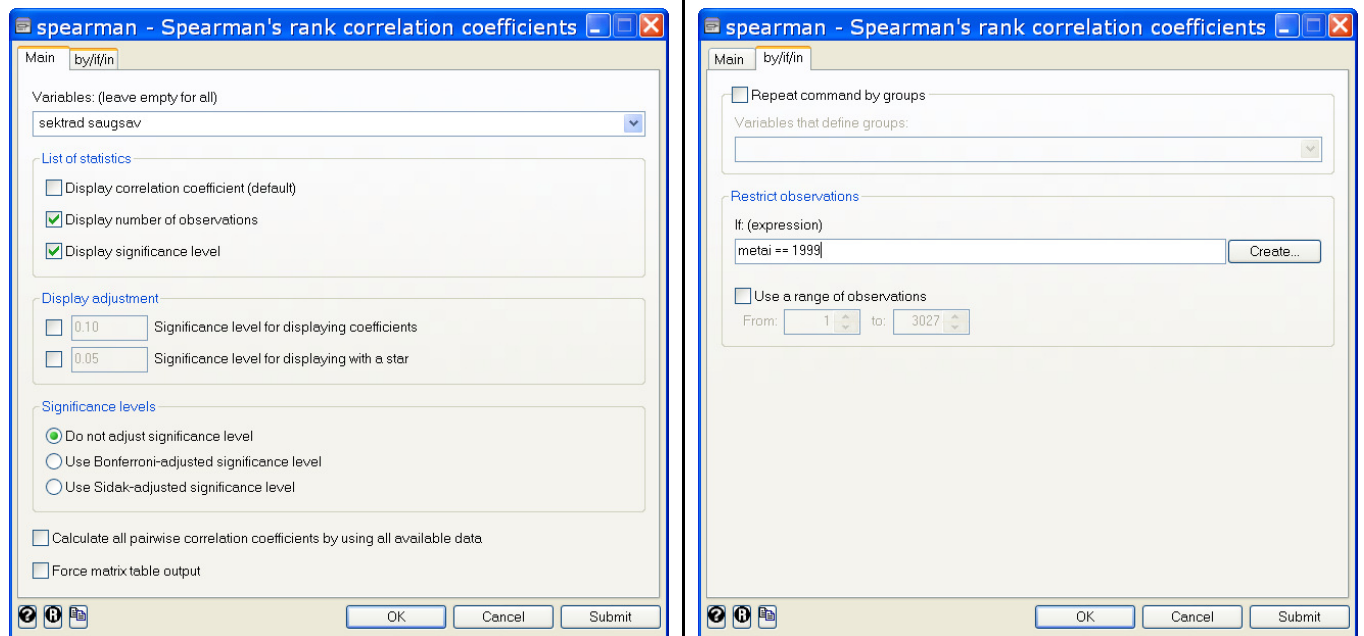
Analizės veiksmi: Spirmeno koreliacijos koeficientas

Sprendimas su STATA 10.0

Meniu punktai: Statistics → Summary, tables, & tests → Nonparametric tests of hypothesis → Spearman's rank correlation



Atsidariusiame lange parenkame kintamuosius **sekt**rad ir **saug**sav ties „Variables...“, o „List of statistics“ skiltyje pasirenkame „Display number of observations“ ir „Display significance level“. Po to „by/if/in“ kortelėje „Restrict observations“ skiltyje ties „If: ...“ įrašome „metai == 1999“ ir spaudžiame „OK“.



Sintaksės komandos

```
spearman sekt'rad saug'sav if metai == 1999, stats(obs p)
```

Gauname tokius rezultatus:

Sprendimas su STATA 10.0

Number of obs =	145
spearman's rho =	-0.3412
Test of Ho: sektrad and saugsav are independent	
Prob > t =	0.0000

Gavome, kad nagrinėjant 1999 m. duomenis (145 stebiniai) Pirsono tiesinės koreliacijos įvertis yra -0,3541 ($p < 0,001$), Spirmeno ryšio koeficientas monotoniškumo prasme yra -0,3412 ($p < 0,001$), t.y., egzistuoja statistiškai reikšmingas ryšys tarp kintamųjų su 95% statistine garantija tiesinio ryšio ir monotoniškumo prasme.

Rezultatų interpretacija sutampa su SPSS pavyzdyje pateiktu apibendrinimu.

6.2. Požymių priklausomumo lentelės. Požymių nepriklausomumo hipotezės tikrinimas (Chi kvadratu kriterijus)

Panagrinėkime, ar priklausymas kokiai nors sporto, meno, bendruomeninei ar labdaringai organizacijai yra susijęs su kurios nors partijos palaikymu (saviidentifikacija su kokia nors partija)? Tirsime šių dvireikšmių kintamųjų iš 2004 m. Naujojo Baltijos barometro tyrimo požymių priklausomumo lentelę:

nvo – Ar jūs priklausote kokiai nors sporto, meno, bendruomeninei ar labdaringai organizacijai?

partpal – Ar jūs palaikote kokią nors partiją?

6.2.1. Požymių priklausomumo 2*2 lentelės su SPSS

Kadangi kintamieji yra dvireikšmiai, nagrinėsime 2*2 požymių priklausomumo lentelę, todėl pasirenkame tik bendrąjį chi-kvadrato kriterijų, kuris tik parodo, ar yra ryšys, ar jo nėra, tačiau neparodo jo stiprumo. Ryšio stiprumui 2*2 lentelėse nustatyti tinka visi intervalų, tvarkos ir vardų skalių koeficientai (šiam pavyzdyje ryšio stiprumo koeficientų nenagrinėjame).

Chi-kvadratas 2*2 požymių priklausomumo lentelėms pateikiamas su Jeito pataisa (Continuity Correction SPSS'e); jeigu tikėtinų reikšmių sąlygos yra neišpildytos (bent viena tikėtina reikšmė yra mažesnė nei 5), reikia remtis Fišerio kriterijumi.

Sprendimas su SPSS 15.0

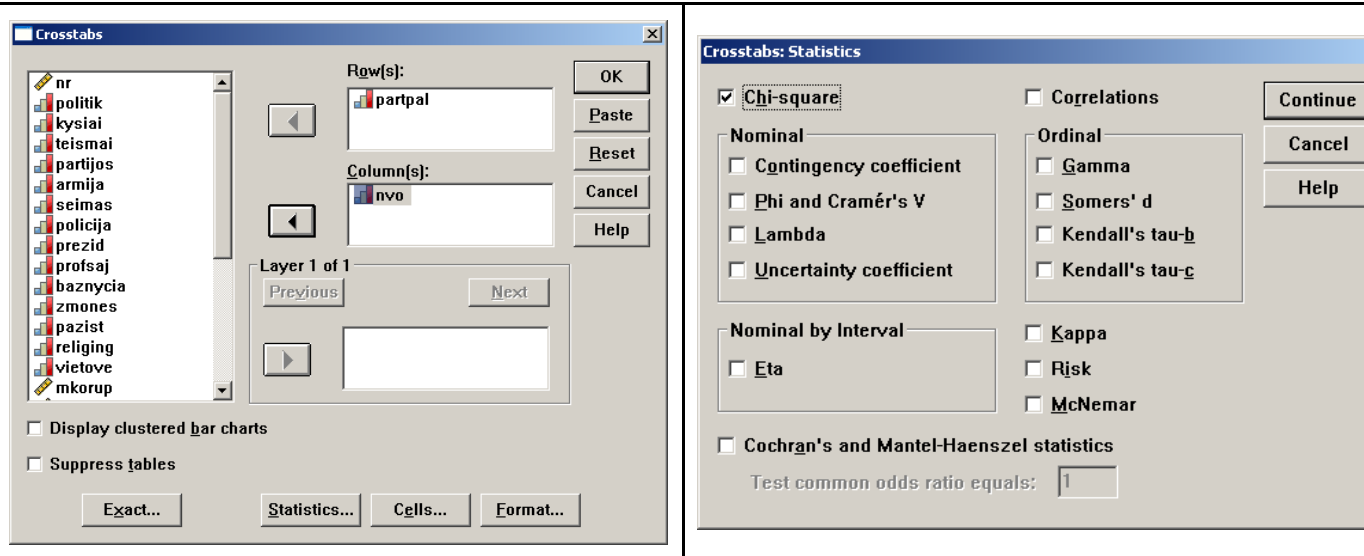
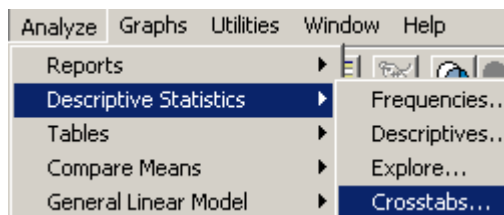
Parengiamieji veiksmai

Nėra

Analizės veiksmai

Meniu punktai: Analize → Descriptive Statistics → Crosstabs...: pasirenkame kintamuosius **nvo** (priklausymas organizacijoms) ir **partpal** (saviidentifikacija su partija) → Statistics...: pažymime „Chi-Square“ → Continue → OK.

Sprendimas su SPSS 15.0



Sintaksės komandos

```
CROSSTABS
/TABLES=partpal BY nvo
/FORMAT=AVALUE TABLES
/STATISTIC=CHISQ
/CELLS= COUNT
/COUNT ROUND CELL .
```

Gauname tokius rezultatus:

Ar palaikote kurią nors partiją? * Ar jūs priklausote kokiai nors sporto, meno, bendruomeninei ar labdarinei organizacijai? Crosstabulation

			Ar jūs priklausote kokiai nors sporto, meno, bendruomeninei ar labdarinei organizacijai?		Total
			Taip	Ne	
Count	Ar palaikote kurią nors partiją?	Taip	38	163	201
		Ne	65	847	912
	Total		103	1010	1113
Expected Count	Ar palaikote kurią nors partiją?	Taip	18,6	182,4	201,0
		Ne	84,4	827,6	912,0
	Total		103,0	1010,0	1113,0
% of Total	Ar palaikote kurią nors partiją?	Taip	3,4%	14,6%	18,1%
		Ne	5,8%	76,1%	81,9%
	Total		9,3%	90,7%	100,0%

Sprendimas su SPSS 15.0

Chi-Square Tests

	Value	df	Asymp. Sig. (2-sided)	Exact Sig. (2-sided)	Exact Sig. (1-sided)
Pearson Chi-Square	27,208 ^b	1	,000		
Continuity Correction ^a	25,823	1	,000		
Likelihood Ratio	22,932	1	,000		
Fisher's Exact Test				,000	,000
Linear-by-Linear Association	27,183	1	,000		
N of Valid Cases	1113				

a. Computed only for a 2x2 table

b. 0 cells (.0%) have expected count less than 5. The minimum expected count is 18,60.

Maži ($<0,05$) chi-kvadrato kriterijaus reikšmingumai antrojoje SPSS rezultatų lentelėje rodo, kad ryšys tarp kintamųjų yra ($p<0,001$), nulinė hipotezė apie požymių nepriklausomumą yra atmesta su 99% statistine garantija. Kadangi daugiausia - 76,1% atvejų (žr. visumos procentus pirmojoje rezultatų lentelėje) - nepalaiko partijos ir nepriklauso partijoms, ryšys yra teigiamas, tuo taip pat galima įsitikinti SPSS meniu užsakius ryšio koeficientus 2*2 lentelėse reikia vadovautis chi-kvadrato su Jeito pataisa („Continuity Correction“) reikšme (25,823) ir jos reikšmingumu ($p<0,001$).

6.2.2. Požymių priklausomumo 2*2 lentelės su Stata

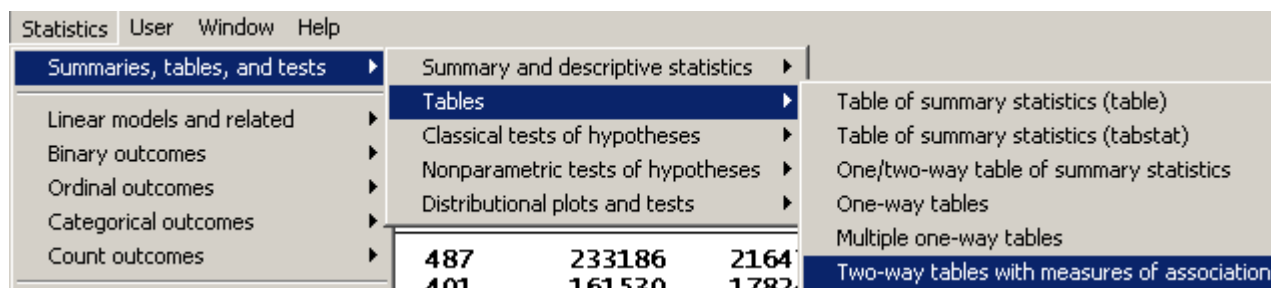
Sprendimas su SPSS 15.0

Parengiamieji veiksmai

Nėra

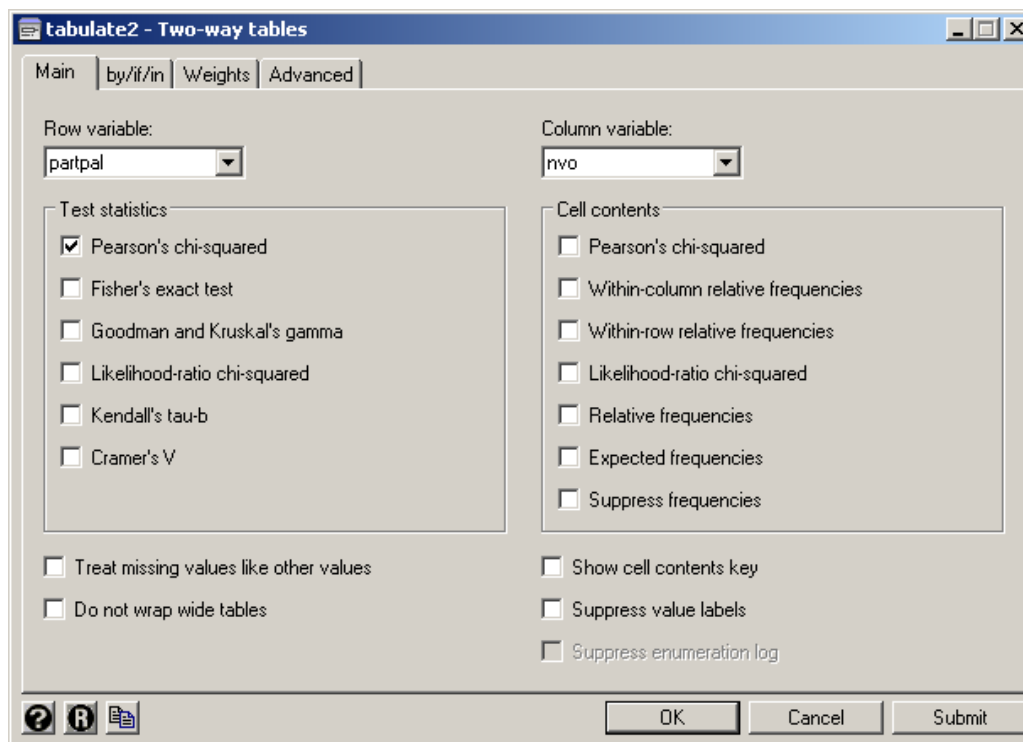
Analizės veiksmai

Meniu punktai: Statistics → Summaries, tables, & tests → Tables → Two-way tables with measure of association



Dialogo lange pasirenkame **partpal** („Row variable“) ir **nvo** („Column variable“), pažymime „Pearson’s chi-squared“, spaudžiame „Submit“

Sprendimas su SPSS 15.0



Sintaksės komandos

```
tabulate partpal nvo, chi2
```

Gauname tokius rezultatus:

ar palaikote <urià nors partiją?	ar jūs priklausote kokiai nors sporto, meno, bendruomeninei ar labdaringai organ		Total
	taip	ne	
taip	38	163	201
ne	65	847	912
Total	103	1,010	1,113

Pearson chi2(1) = 27.2077 Pr = 0.000

Pateikta chi-kvadrato reikšmė sutampa su SPSS rezultatu (žr. aukščiau): nulinė hipotezė apie kintamųjų nepriklausomumą atmetama ($p > 0,001$). STATA, deja, nepateikia chi-kvadrato su Jeito pataisa reikšmės 2*2 lentelėms. Rezultatų paaiškinimą žr. SPSS pavyzdyje.

6.3. Ryšio stiprumo matai tvarkos skalėje

Tradicinės vertybės ir ypač religingumas neretai yra siejami su bendresne pasaulėžiūra, apimančia ir politines orientacijas. Lietuvoje didesnis religingumas turėtų būti susijęs su konservatorių partijomis, kurios save laiko dešiniosiomis. Naudodamiesi EVT duomenų failu, ištirkime, ar saviidentifikavimas politinės kairės-dešinės skalėje (kintamasis saviid) yra susijęs su religingumu, kuris operacionalizuotas (išmatuotas) religing klausimu „Ar dažnai paskutiniu metu einate į bažnyčią ar religinius susirinkimus?“.

6.3.1. Ryšio stiprumo matai tvarkos skalėje su SPSS

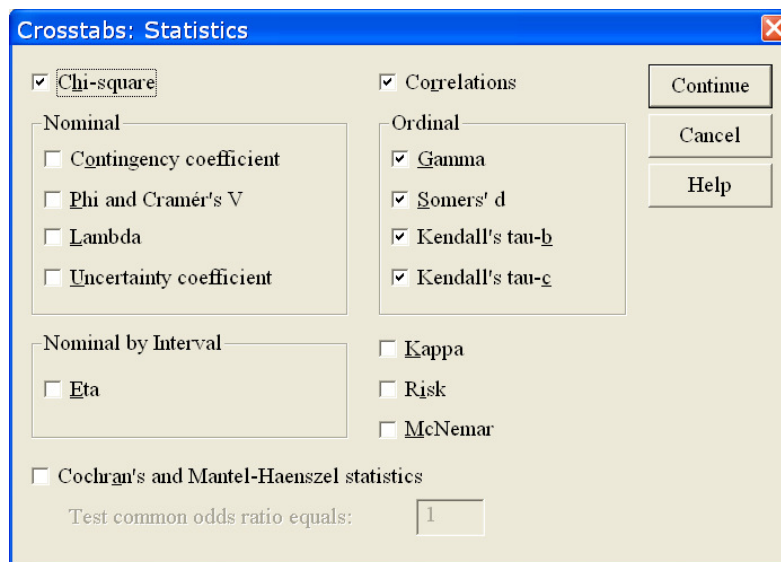
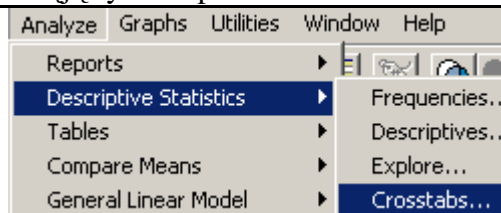
Sprendimas su SPSS 15.0

Parengiamieji veiksmai

Nėra

Analizės veiksmai

Meniu punktai: Analyze → Descriptive Statistics → Crosstabs... Pasirenkame chi-kvadrato kriterijų, koreliacijas ir tvarkos skalės kintamųjų ryšio stiprumo matus.



Sintaksės komandos

```
CROSSTABS
  /TABLES=saviid BY religing
  /FORMAT= AVALUE TABLES
  /STATISTIC=CHISQ CORR GAMMA D BTAU CTAU
  /CELLS= COUNT
  /COUNT ROUND CELL .
```

Sprendimas su SPSS 15.0

Gauname tokius rezultatus:

Saviidentifikavimas politinės kairės-dešinės skalėje * Kiek dažnai paskutiniu metu einate į bažnyčią? Crosstabulation

Count		Kiek dažnai paskutiniu metu einate į bažnyčią?								Total
		Dazniau negu karta per savaitę	Karta per savaitę	Karta per mėnesį	Tik per Kalėdas ir Velykas	Per religines šventes	Karta per metus	Rečiau nei karta per metus	Niekada, beveik niekada	
Saviidentifikavimas	Kaire	2	5	10	15	8	4	6	13	63
politinės	2	1	3	5	11	6	5	2	11	44
kairės-dešinės	3	1	8	18	29	20	7	10	25	118
skalėje	4	0	5	14	22	8	9	12	26	96
	5	4	43	63	157	68	39	46	74	494
	6	6	18	37	53	26	21	20	29	210
	7	2	11	17	27	4	6	8	6	81
	8	6	39	31	37	10	6	3	12	144
	9	6	19	13	19	2	6	2	5	72
	Desine	6	32	19	24	10	5	2	12	110
Total		34	183	227	394	162	108	111	213	1432

Chi-Square Tests

	Value	df	Asymp. Sig. (2-sided)
Pearson Chi-Square	191,615 ^a	63	,000
Likelihood Ratio	186,701	63	,000
Linear-by-Linear Association	85,094	1	,000
N of Valid Cases	1432		

a. 14 cells (17,5%) have expected count less than 5. The minimum expected count is 1,04.

Directional Measures

			Value	Asymp. Std. Error ^a	Approx. T ^b	Approx. Sig.
Ordinal by Ordinal	Somers' d	Symmetric	-,202	,021	-9,794	,000
		Saviidentifikavimas politinės kairės-dešinės skalėje	-,201	,020	-9,794	,000
		Dependent Kiek dažnai paskutiniu metu einate į bažnyčią? Dependent	-,204	,021	-9,794	,000

a. Not assuming the null hypothesis.

b. Using the asymptotic standard error assuming the null hypothesis.

Symmetric Measures

		Value	Asymp. Std. Error ^a	Approx. T ^b	Approx. Sig.
Ordinal by Ordinal	Kendall's tau-b	-,202	,021	-9,794	,000
	Kendall's tau-c	-,192	,020	-9,794	,000
	Gamma	-,243	,025	-9,794	,000
	Spearman Correlation	-,258	,026	-10,099	,000 ^c
Interval by Interval	Pearson's R	-,244	,026	-9,508	,000 ^c
N of Valid Cases		1432			

a. Not assuming the null hypothesis.

b. Using the asymptotic standard error assuming the null hypothesis.

c. Based on normal approximation.

Visų pirma, studijuodami chi-kvadrato kriterijaus lentelę pastebime, kad galime taikyti chi-kvadrato kriterijų (mažiausia tikėtina reikšmė yra $1,04 > 1$ ir tik $17,5\% < 20\%$ langelių turi tikėtinas reikšmes mažesnes negu 5).

Nesimetrinio Somerso-d ryšio koeficiento taškiniai įverčiai rodo, kad ryšys tarpo kintamųjų yra statistiškai reikšmingas ($p < 0,001$), silpnas, neigiamas ir beveik simetriškas (asimetriniai įverčiai yra

Sprendimas su SPSS 15.0

atitinkamai -0,201 ir -0,204).

Visi koeficientai rodo, kad egzistuoja statistiškai reikšmingas, silpnas, neigiamas ryšys tarp saviidentifikavimo politinės kairės-dešinės skalėje ir religingumo, o žvilgsnis į požymių lentelę padeda nustatyti šio ryšio interpretaciją: kuo dešiniau skalėje save identifikuoja Lietuvos gyventojas, tuo jis pasižymi didesniu religingumu (neigiama koreliacija rodo, kad kintamieji yra susiję priešingai, t.y., vienam didėjant, kitas mažėja, tačiau reikia atsižvelgti į kodų reikšmių prasmę: dažniausiai einančių į bažnyčią kategorija žymima skaičiumi „1“, o rečiausiai einančių atsakymai skaičiumi „8“ – aišku, būtų patogiau, jeigu didesnis lankymo dažnis būtų koduotas didesnėmis reikšmėmis, tada gautume teigiamą koreliaciją; interpretuojant ryšį tarp kintamųjų būtina atkreipti dėmesį į kintamojo kodavimą).

Jeigu koeficientų dydžiai labai skirtingi tarpusavyje, jais remiantis būtų galima patikslinti bendrąją išvadą, kokia prasme ryšys yra stipriausias, pvz., monotoniškumo, surištų porų ir pan. Tačiau verta atkreipti dėmesį, a) kad koeficientai gali turėti skirtingą matą (skiriasi reikšmių intervalai, pvz., [0;1], [-1;1]), todėl gali būti nekorektiška juos tiesiogiai lyginti tarpusavyje ir b) kad nors taškiniai koeficientų įverčiai skiriasi, tačiau jų pasikliautiniai intervalai gali persidengti ir statistiškai reikšmingo skirtumo tarp jų gali nebūti. SPSS pateikia asimptotinę standartinę paklaidą (Asymp. Std. Error), pagal kurią galime suskaičiuoti 95% pasikliautinąjį intervalą: [koeficiento taškinis įvertis - 2* asimptotinė standartinė paklaida; koeficiento taškinis įvertis + 2* asimptotinė standartinė paklaida]. Nagrinėjamu atveju gauname, kad nei vieno pasikliautinąjo intervalo minimali vertė neviršija nė vieno kito pasikliautinąjo intervalo maksimalios reikšmės, t.y., visi pasikliautiniai intervalai persidengia ir todėl koeficientus galime laikyti tarpusavyje nesiskiriančiais.

6.3.2. Ryšio stiprumo matai tvarkos skalėje su Stata

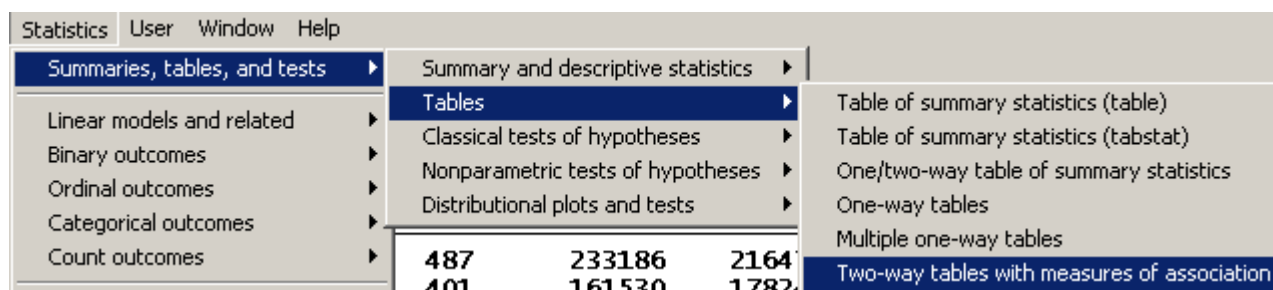
Sprendimas su STATA 10.0

Parengiamieji veiksmai

Nėra

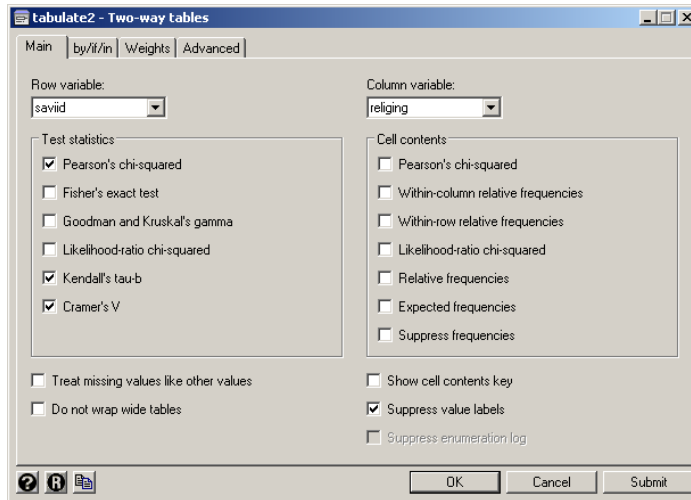
Analizės veiksmai

Meniu punktai: Statistics → Summaries, tables, & tests → Tables → Two-way tables with measure of association



Dialogo lange pasirenkame **saviid** („Row variable“) ir **relig** („Column variable“), pažymime „Pearson’s chi-squared“, „Kendall’s tau-b“, „Cramer’s V“, bei „Suppress value labels“, spaudžiame „Submit“

Sprendimas su STATA 10.0



Sintaksės komandos

tabulate saviid religing, chi2 nolabel taub V

Gauname tokius rezultatus:

Saviidentifikavimas politinės kairės-dešinės skalėje	kiek dažnai paskutiniu metu einate į bažnyčią?						Total
	1	2	3	4	5	6	
1	2	5	10	15	8	4	63
2	1	3	5	11	6	5	44
3	1	8	18	29	20	7	118
4	0	5	14	22	8	9	96
5	4	43	63	157	68	39	494
6	6	18	37	53	26	21	210
7	2	11	17	27	4	6	81
8	6	39	31	37	10	6	144
9	6	19	13	19	2	6	72
10	6	32	19	24	10	5	110
Total	34	183	227	394	162	108	1,432

Saviidentifikavimas politinės kairės-dešinės skalėje	kiek dažnai paskutiniu metu einate į bažnyčią?		Total
	7	8	
1	6	13	63
2	2	11	44
3	10	25	118
4	12	26	96
5	46	74	494
6	20	29	210
7	8	6	81
8	3	12	144
9	2	5	72
10	2	12	110
Total	111	213	1,432

Pearson chi2(63) = **191.6155** Pr = **0.000**
 Cramér's V = **0.1383**
 Kendall's tau-b = **-0.2023** ASE = **0.021**

Išvada: nulinė kintamųjų požymių nepriklausomumo hipotezė atmesta ($p < 0,001$). Surištų porų prasme, egzistuoja neigiamas ryšys religingumo ir politinės saviidentifikacijos (Kendalo tau-b = -0,202). Kitus ryšio koeficientus reikia užsakinėti iš kitų STATA meniu.

6.4. Ryšio stiprumo matai vardų skalėje

Pateikdami ryšio stiprumo matų vardų skalėje analizės pavyzdžius panagrinėsime, ar šeimyninis statusas (**seimpad**) ir darbinis statusas (**darbstat**) yra susiję (naudojame EVT duomenų failą)?

6.4.1. Ryšio stiprumo matai vardų skalėje su SPSS

Sprendimas su SPSS 15.0

Parengiamieji veiksmai

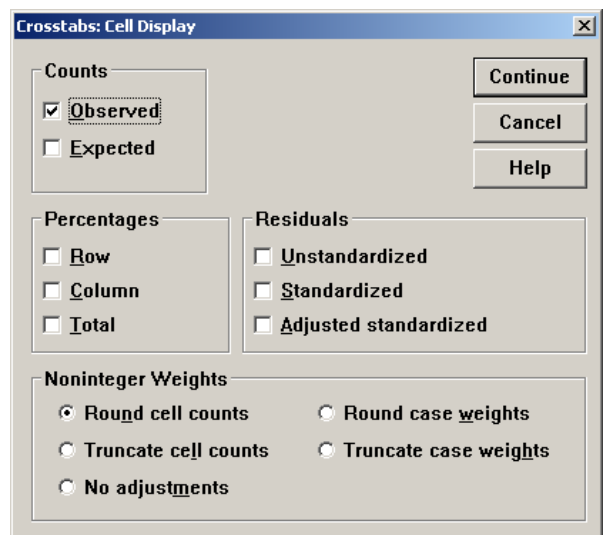
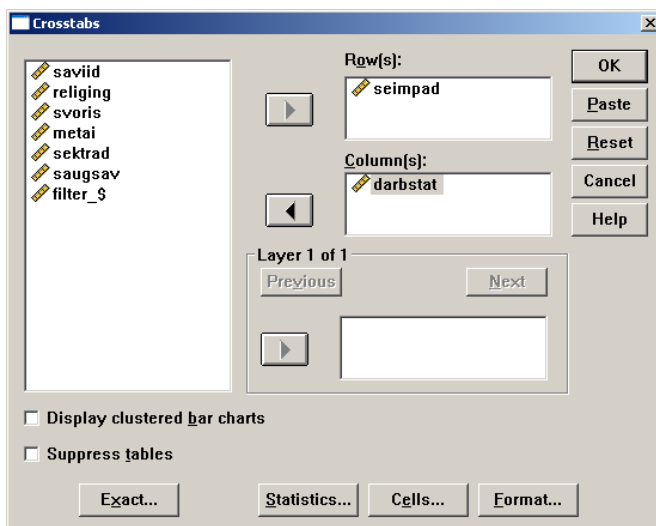
Nėra

Analizės veiksmai

Meniu punktai: Analize → Descriptive Statistics → Crosstabs

Požymių priklausomumo meniu pasirinkę **seimpad** ir **darbstat** kintamuosius kituose meniu pradinės analizės etape užsakome tik chi-kvadrato statistiką (nes kol kas nežinome, ar mūsų duomenys tinkami chi-kvadrato ir juo paremtų ryšio matų skaičiavimui). Žemiau pateikti pagrindiniai pasirinkimų langai ir rezultatų lentelės.

Neretai naudojamos šios taisyklės tinkamo chi-kvadrato skaičiavimui: nė vienas langelis negali turėti tikėtinų reikšmės mažesnės negu 1 ir nedaugiau kaip 20% langelių gali turėti tikėtiną reikšmę mažesnę nei 5, o 2*2 lentelėse – nė vienas iš langelių neturi turėti tikėtinų reikšmių, mažesnių už 5.



Sprendimas su SPSS 15.0

Crosstabs: Statistics

☒ Chi-square

☐ Contingency coefficient

☐ Phi and Cramér's V

☐ Lambda

☐ Uncertainty coefficient

☐ Correlations

☐ Gamma

☐ Somers' d

☐ Kendall's tau-b

☐ Kendall's tau-c

☐ Nominal by Interval

☐ Eta

☐ Kappa

☐ Risk

☐ McNemar

☐ Cochran's and Mantel-Haenszel statistics

Test common odds ratio equals: 1

Continue

Cancel

Help

Sintaksės komandos

SPLIT FILE

OFF.

CROSSTABS

/TABLES=seimpad BY darbstat

/FORMAT=AVALUE TABLES

/STATISTIC=CHISQ

/CELLS= COUNT

/COUNT ROUND CELL .

Gauname tokius rezultatus:

Šeimyninė padėtis * Darbinis statusas Crosstabulation

Count		Darbinis statusas								Total
		Pilna darbo diena	Ne pilna darbo diena	Savarankiskai dirbantis	Pensijoje	Namu seimininke	Studentas/-e	Bedarbis/-e	Kita	
Šeimyninė padėtis	Vedes/-usi	815	311	76	355	85	12	163	48	1865
	Gyvena kartu, nevede	12	2	5	4	1	2	7	3	36
	Issiskyres/-usi	86	31	10	17	3	0	31	2	180
	Gyvena atskirai	28	4	1	10	1	0	6	0	50
	Naslys/-e	64	23	0	171	4	2	8	0	272
	Nevedes/-usi	183	77	12	26	2	206	76	11	593
Total		1188	448	104	583	96	222	291	64	2996

Chi-Square Tests

	Value	df	Asymp. Sig. (2-sided)
Pearson Chi-Square	1260,679 ^a	35	,000
Likelihood Ratio	1043,437	35	,000
Linear-by-Linear Association	121,712	1	,000
N of Valid Cases	2996		

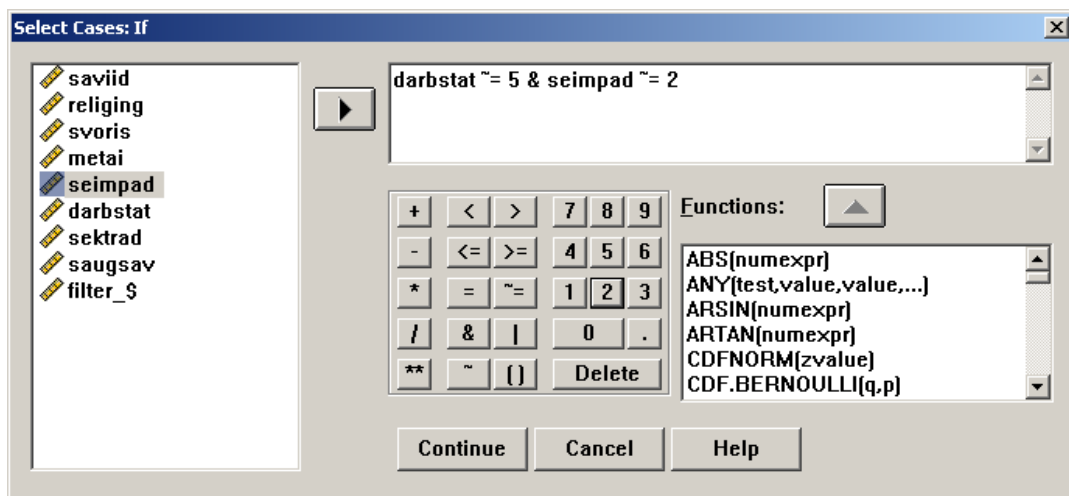
a. 11 cells (22,9%) have expected count less than 5. The minimum expected count is ,77.

Sprendimas su SPSS 15.0

Chi-kvadrato kriterijaus lentelės apačioje esantis užrašas mus perspėja, kad, deja, šios sąlygos nėra išpildytos: 22,9% langelių turi tikėtinas reikšmes mažesnes negu 5, o minimali tikėtina reikšmė yra 0,77. Esant tokiai situacijai, siūlomos tokios išeitys: ištrinti arba išmesti iš analizės (panaudojant Select Cases galimybę iš Data meniu SPSS) kategorijas, kuriose yra nedaug atvejų arba sujungti panašias kategorijas į vieną naują, kurioje būtų daugiau atvejų.

Į analizę neįtraukiame 5-os kategorijos („namų šeimininkės“) iš **seimpad** ir 2-os kategorijos („nevedę, gyvenantys kartu“) iš **darbstat**. Alternatyviai, buvo galima prijungti šias kategorijas prie kitų, artimiausių pagal prasmę kategorijų. Dėl to rezultatai bus apibendrinami tik populiacijai be šių kategorijų.

Meniu punktai: Data → Select Cases: If



Sintaksės komandos

```
USE ALL.  
COMPUTE filter_$=(darbstat ~= 5 & seimpad ~= 2).  
VARIABLE LABEL filter_$ 'darbstat ~= 5 & seimpad ~= 2 (FILTER)'.  
VALUE LABELS filter_$ 0 'Not Selected' 1 'Selected'.  
FORMAT filter_$ (f1.0).  
FILTER BY filter_$.  
EXECUTE .
```

Pakartojame analizę:

Sprendimas su SPSS 15.0

Crosstabs: Cell Display X

Counts

☒ Observed

☒ Expected

Percentages

☒ Row

☒ Column

☒ Total

Residuals

☐ Unstandardized

☐ Standardized

☒ Adjusted standardized

Noninteger Weights

☒ Round cell counts ☐ Round case weights

☐ Truncate cell counts ☐ Truncate case weights

☐ No adjustments

Continue

Cancel

Help

Gauname tokius rezultatus:

Šeimyninė padėtis * Darbinis statusas Crosstabulation

			Darbinis statusas							Total
			Pilna darbo diena	Ne pilna darbo diena	Savarankiskai dirbantis	Pensijoje	Studentas/-e	Bedarbis/-e	Kita	
Count	Šeimyninė padėtis	Vedes/-usi	815	311	76	355	12	163	48	1780
		Issiskyres/-usi	86	31	10	17	0	31	2	177
		Gyvena atskirai	28	4	1	10	0	6	0	49
		Naslys/-e	64	23	0	171	2	8	0	268
		Nevedes/-usi	183	77	12	26	206	76	11	591
	Total		1176	446	99	579	220	284	61	2865
Expected Count	Šeimyninė padėtis	Vedes/-usi	730,6	277,1	61,5	359,7	136,7	176,4	37,9	1780,0
		Issiskyres/-usi	72,7	27,6	6,1	35,8	13,6	17,5	3,8	177,0
		Gyvena atskirai	20,1	7,6	1,7	9,9	3,8	4,9	1,0	49,0
		Naslys/-e	110,0	41,7	9,3	54,2	20,6	26,6	5,7	268,0
		Nevedes/-usi	242,6	92,0	20,4	119,4	45,4	58,6	12,6	591,0
	Total		1176,0	446,0	99,0	579,0	220,0	284,0	61,0	2865,0
% within Šeimyninė padėtis	Šeimyninė padėtis	Vedes/-usi	45,8%	17,5%	4,3%	19,9%	,7%	9,2%	2,7%	100,0%
		Issiskyres/-usi	48,6%	17,5%	5,6%	9,6%	,0%	17,5%	1,1%	100,0%
		Gyvena atskirai	57,1%	8,2%	2,0%	20,4%	,0%	12,2%	,0%	100,0%
		Naslys/-e	23,9%	8,6%	,0%	63,8%	,7%	3,0%	,0%	100,0%
		Nevedes/-usi	31,0%	13,0%	2,0%	4,4%	34,9%	12,9%	1,9%	100,0%
	Total		41,0%	15,6%	3,5%	20,2%	7,7%	9,9%	2,1%	100,0%
% within Darbinis statusas	Šeimyninė padėtis	Vedes/-usi	69,3%	69,7%	76,8%	61,3%	5,5%	57,4%	78,7%	62,1%
		Issiskyres/-usi	7,3%	7,0%	10,1%	2,9%	,0%	10,9%	3,3%	6,2%
		Gyvena atskirai	2,4%	,9%	1,0%	1,7%	,0%	2,1%	,0%	1,7%
		Naslys/-e	5,4%	5,2%	,0%	29,5%	,9%	2,8%	,0%	9,4%
		Nevedes/-usi	15,6%	17,3%	12,1%	4,5%	93,6%	26,8%	18,0%	20,6%
	Total		100,0%	100,0%	100,0%	100,0%	100,0%	100,0%	100,0%	100,0%
% of Total	Šeimyninė padėtis	Vedes/-usi	28,4%	10,9%	2,7%	12,4%	,4%	5,7%	1,7%	62,1%
		Issiskyres/-usi	3,0%	1,1%	,3%	,6%	,0%	1,1%	,1%	6,2%
		Gyvena atskirai	1,0%	,1%	,0%	,3%	,0%	,2%	,0%	1,7%
		Naslys/-e	2,2%	,8%	,0%	6,0%	,1%	,3%	,0%	9,4%
		Nevedes/-usi	6,4%	2,7%	,4%	,9%	7,2%	2,7%	,4%	20,6%
	Total		41,0%	15,6%	3,5%	20,2%	7,7%	9,9%	2,1%	100,0%
Adjusted Residual	Šeimyninė padėtis	Vedes/-usi	6,6	3,6	3,1	-,5	-18,0	-1,7	2,7	
		Issiskyres/-usi	2,1	,7	1,7	-3,6	-4,0	3,5	-1,0	
		Gyvena atskirai	2,3	-1,4	-,5	,0	-2,0	,6	-1,0	
		Naslys/-e	-6,0	-3,3	-3,3	18,7	-4,5	-4,0	-2,5	
		Nevedes/-usi	-5,6	-1,9	-2,1	-10,7	27,9	2,7	-,5	

Chi-Square Tests

	Value	df	Asymp. Sig. (2-sided)
Pearson Chi-Square	1176,834 ^a	24	,000
Likelihood Ratio	984,951	24	,000
Linear-by-Linear Association	146,854	1	,000
N of Valid Cases	2865		

a. 5 cells (14,3%) have expected count less than 5. The minimum expected count is 1,04.

Directional Measures

			Value	Asymp. Std. Error ^a	Approx. T ^b	Approx. Sig.
Nominal by Nominal	Lambda	Symmetric	,117	,012	9,286	,000
		Šeimyninė padėtis	,179	,012	13,554	,000
		Dependent				
	Goodman and Kruskal tau	Darbinis statusas	,077	,014	5,229	,000
		Dependent				
		Šeimyninė padėtis	,159	,007		,000 ^c
	Uncertainty Coefficient	Darbinis statusas	,070	,006		,000 ^c
		Dependent				
		Šeimyninė padėtis	,128	,007	16,685	,000 ^d
		Dependent	,158	,009	16,685	,000 ^d
		Darbinis statusas	,107	,006	16,685	,000 ^d

a. Not assuming the null hypothesis.

b. Using the asymptotic standard error assuming the null hypothesis.

c. Based on chi-square approximation

d. Likelihood ratio chi-square probability.

Symmetric Measures

		Value	Approx. Sig.
Nominal by Nominal	Phi	,641	,000
	Cramer's V	,320	,000
	Contingency Coefficient	,540	,000
N of Valid Cases		2865	

a. Not assuming the null hypothesis.

b. Using the asymptotic standard error assuming the null hypothesis.

Šį kartą SPSS meniu užsakome platesnį duomenų analizės spektrą: tikėtinus dažnius, procentines išraiškas stulpeliuose, eilutėse ir bendras kiekvieno langelio procentines išraiškas, taip pat galime užsakyti pataisytas standartizuotas liekanas, kurias galima interpretuoti kaip Z skaičių pavidalu pateiktus teigiamus arba neigiamus nukrypimus, lyginant stebimas su tikėtinomis reikšmėmis (kuo didesnis nuokrypis, tuo daugiau nukrypstama nuo nulinės kintamųjų nepriklausomumo hipotezės; taip pat, dideli nuokrypiai padeda identifikuoti, kurioje vietoje yra didžiausi neatitikimai).

Šį kartą gavome, kad 14,3% langelių turi tikėtinas reikšmes mažesnes negu 5, o minimali tikėtina reikšmė yra 1,04, t.y., reikalavimai chi-kvadrato kriterijaus skaičiavimui yra išpildyti. Chi-kvadrato kriterijaus lentelė taip pat parodo, kad nulinė hipotezė apie darbinio ir šeimyninio statuso nepriklausomumą yra atmetama ($p < 0,001$).

Kryptiniai ryšio matai, skirti vardų skalės kintamiesiems, parodo, kokią dalimi sumažėja klaida nuspėjant vieno kintamojo kategoriją žinant kito kintamojo kategoriją, todėl pateikiamos trys galimybės: a) simetrinis (vidutinis) ryšio stiprumas, b) ryšio stiprumas, kai pirmasis kintamasis yra priklausomas (jo reikšmės nuspėjame remdamiesi antro kintamojo reikšmėmis), c) ryšio stiprumas, kai antrasis kintamasis yra priklausomas (jo reikšmės nuspėjame remdamiesi pirmojo kintamojo reikšmėmis).

Lambda, Gudmano ir Kraskelo tau, Nuspėjamumo koeficiento („Uncertainty coefficient“) taškiniai įverčiai yra statistiškai reikšmingi ($p < 0,001$), darbinį statusą kur kas geriau galima nuspėti, žinant respondento šeimyninę padėtį (egzistuoja asimetriškas ryšys tarp kintamųjų): klaida sumažėja 17,9% - 15,8% (rezultatai skiriasi, priklausomai nuo analizei naudojamo koeficiento) šeimyninės padėties kategorijas nuspėjant darbinio statuso kategorijomis, o nuspėjant darbinio statuso kategorijas šeimyninės padėties kategorijomis klaida sumažėja 7% - 10,7% (priklausomai nuo analizei naudojamo koeficiento). Simetriniai matų, paremtų chi-kvadratu - Kramerio V, Kontingencijos koeficiento C (Pirsono kontingencijos koef.) taškiniai įverčiai yra statistiškai reikšmingi ($p < 0,001$), rodo vidutinio stiprumo ryšį tarp šeimyninio ir darbinio statuso (0,320 ir 0,540, atitinkamai). Koeficientas Fi yra skirtas 2*2 požymių lentelėms, todėl rekomenduojama naudotis Kramerio V, kuris yra Fi koeficiento apibendrinimas $n \times m$ požymių lentelėms.

Verta atkreipti dėmesį, kad didžiausi nuokrypiai (remiantis pataisytų standartizuotų liekanų reikšmėmis) yra nevedusių studentų (27,9, t.y., yra žymiai daugiau, nei tikėtina), našlių pensijoje (18,7), vedusių studentų (-18,0, t.y., yra žymiai mažiau negu tikėtina), nevedusių pensijoje (-10,7) langeliuose. Šie rezultatai, panašu, kad neprieštarauja mums pažįstamai realybei, bei nurodo, kurios kategorijos daugiausia prisideda prie nuspėjamumo klaidos sumažinimo.

6.4.2. Ryšio stiprumo matai vardų skalėje su Stata

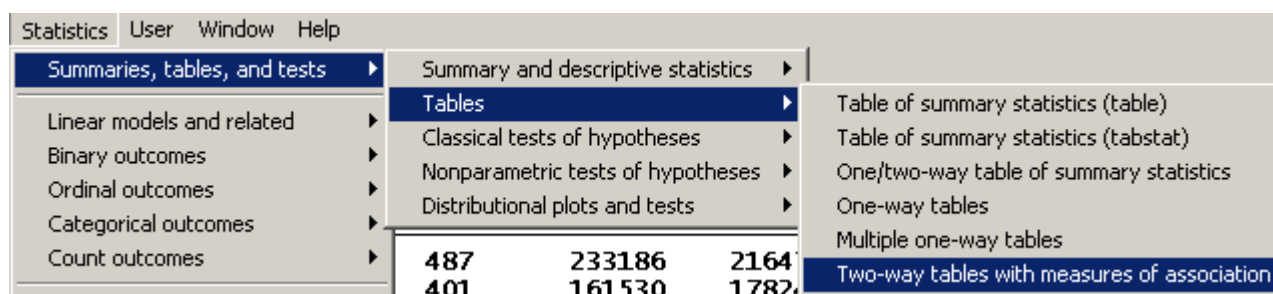
Sprendimas su SPSS 15.0

Parengiamieji veiksmai

Nėra

Analizės veiksmai

Meniu punktai: Statistics → Summaries, tables, & tests → Tables → Two-way tables with measure of association



Dialogo lange pasirenkame **seimpad** („Row variable“) ir **darbstat** („Column variable“), pažymime „Pearson’s chi-squared“, „Cramer’s V“, bei „Supress value labels“. Tai pat, iš analizės išmetame 5-ą **seimpad** atsakymų kategoriją („namų šeiminkės“) ir 2-ą **darbstat** kategoriją („nevedę, gyvenantys kartu“) dėl priešasčių, aptartų analogiškame SPSS pavyzdyje. Spaudžiame „Submit“.

Sprendimas su SPSS 15.0

Sintaksės komandos

```
tabulate seimpad darbstat if seimpad!= 2 & darbstat != 5, chi2 nolabel v
```

Gauname tokius rezultatus:

Šeimyninė padėtis	Darbinis statusas						Total
	1	2	3	4	6	7	
1	815	311	76	355	12	163	1,780
3	86	31	10	17	0	31	177
4	28	4	1	10	0	6	49
5	64	23	0	171	2	8	268
6	183	77	12	26	206	76	591
Total	1,176	446	99	579	220	284	2,865

Šeimyninė padėtis	Darbinis statusas	Total
	8	
1	48	1,780
3	2	177
4	0	49
5	0	268
6	11	591
Total	61	2,865

Pearson chi2(24) = 1.2e+03 Pr = 0.000
Cramér's V = 0.3205

Chi-kvadrato kriterijaus reikšmė ir jos statistinis reikšmingumas parodo, kad nulinė hipotezė apie darbinio ir šeimyninio statuso nepriklausomumą yra atmetama ($p < 0,001$). Egzistuoja vidutinio stiprumo ryšys tarp kintamųjų (Kramerio $V = 0,321$).

Pastebime, kad STATA pakete mažiau galimybių užsakyti ryšio koeficientus naudojantis vartotojo meniu. Taip pat, reikia pačiam vartotojui pasitikrinti chi-kvadrato skaičiavimo prielaidas (žr. SPSS pavyzdį).

7. REGRESINĖ ANALIZĖ

Šiame skyriuje pateikiami pavyzdžiai, kaip su SPSS ir Stata atlikti regresinę analizę. Bus naginėjamas sudėtingesnis – daugialypės regresijos – atvejis.

7.1. Daugialypė tiesinė regresinė analizė

Panagrinėkime, kaip pasitikėjimas Lietuvos institucijomis priklauso nuo kitų nuostatų ir demografinių kintamųjų.

Nagrinėsime, kaip pasitikėjimas Lietuvos institucijomis gali būti aiškinamas demografiniais atributais (lytimi ir amžiumi), šeimos pajamų lygiu ir kitomis subjektyviomis orientacijomis (esamos ekonominės sistemos vertinimas ir korupcijos lygio vertinimas).

Pasitikėjimo institucijomis lygį veikia esamo politinio ir ekonominio režimo veiksmingumas ir įvairios kultūrinės įtakos. Politinio ir ekonominio režimo veiksmingumo įtaką, be kitų kintamųjų, galime apytiksliai įvertinti objektyviais matais (pvz., šeimos pajamų lygis) ir sistemos lygio gyventojų subjektyviais vertinimais (pvz., esamos ekonominės sistemos, korupcijos lygio vertinimais ir kt.). Kultūrinės įtakos pasireiškia per skirtumus, susijusius su skirtingomis kartomis, kurios patyrė nevienodą socializacinę įtaką, priešinga lytimi, kadangi ši skirtis parodo istoriškai ir kultūriškai įsišaknijusius vyro ir moters elgesio ir suvokimo stereotipus bei kitus socio-ekonominis ir demografinius veiksnus.

Remdamiesi Naujojo Baltijos Barometro 2004 m. tyrimo duomenų pagrindu sudarytu pavyzdžių failu „NBB Lietuva 2004.sav“ („NBB Lietuva 2004.dta“) įvertinkime pasitikėjimo institucijomis priklausomybę nuo Lietuvos dabartinės ekonominės sistemos vertinimo (ekonom), amžiaus (amzius), šeimos pajamų (pajamos), lyties (lytis), suvokiamo korupcijos lygio (korupcija).

Bendrasis regresinės analizės algoritmas:

parenkame teorinį modelį (kokie veiksniai ir kaip turėtų paaiškinti pasitikėjimą institucijomis), empirinį modelį (kokie indikatoriai operacionalizuoja teoriškai identifikuotus veiksnus), statistinį modelį (daugialypė tiesinė regresija)

atliekame regresinės analizės procedūrą ir tikriname statistinio modelio sąlygas:

tiesinio ryšio stiprumą tarp priklausomo ir nepriklausomų kintamųjų (duomenų sklaidos diagrama, Pirsono koreliacijos koef.)

liekanų skirstinių normalumo arba priklausomo ir nepriklausomųjų intervalinių kintamųjų skirstinių normalumo sąlygą

liekanų sąlyginės dispersijos pastovumą

interpretuojame statistinės analizės metu gautus įverčius

7.1.1. Daugialypė tiesinė regresinė analizė su SPSS

7.1.1.1 Kintamųjų paruošimas regresinei analizei

Prieš atliekant regresinę analizę aprašomosios statistikos pagalba (SPSS'e per **Analyze → Descriptive statistics → Frequencies...** ir **Descriptives...** meniu, STATA – per **Statistics → Summaries, tables, and tests → Summary and descriptive statistics → Summary statistics**) verta įsitikinti, kokios yra kintamųjų skaitinės charakteristikos, ypač atkreipiant dėmesį į reikšmių intervalą, kintamojo kodavimo kryptį (ar didelės kodų reikšmės reiškia aukštą požymio lygį ar

atvirkščiai, žemą požymio lygį), o taip pat patenkančių į analizę atvejų skaičių kiekvieno kintamojo atveju (įsitikinant, kad kiekviename kintamajame yra sąlyginai nedaug trūkstamų reikšmių). Siūlome šią dalį atlikti savarankiškai, remiantis pirmos dalies teorija ir pavyzdžiais.

Priklausomi kintamieji regresinei analizei buvo sukurti faktorinės analizės būdu (žr. skyrelį apie faktorinę analizę): **pasidem** (pasitikėjimas demokratinės valstybės institucijomis), **pasitrad** (pasitikėjimas tradicinio autoriteto institucijomis).

Sprendimas su SPSS 15.0

Parengiamieji veiksmai

Tiesinė regresinė analizė, paremta mažiausių kvadratų metodu, turi griežtus reikalavimus į analizę įtraukiamiems kintamiesiems: paklaidos turi turėti normalųjį skirstinį arba priklausomo intervalinio ir nepriklausų intervalinių kintamųjų skirstinys turi būti daugiamatis normalusis. Atliekant regresinę analizę, reikia patikrinti, ar į analizę įtraukti intervaliniai kintamieji tenkina normalumo sąlygą.

Patikriname, ar priklausomų ir nepriklausomų intervalinių kintamųjų skirtiniai yra suderinami su normaliuoju skirstiniu: **Analyze → Nonparametric tests → 1 Sample K-S...**, pasirenkame intervalinius **pasidem**, **pasitrad** priklausomus kintamuosius ir nepriklausomus kintamuosius **ekonom**, **amzius**, **pajamos**, **lytis** (detaliau apie suderinamumo hipotezės tikrinimą žr. vienfaktorinės dispersinės analizės pavyzdį) → „OK“.

One-Sample Kolmogorov-Smirnov Test

		Pasitikėjimas demokratinės valstybės institucijomis	Pasitikėjimas tradicinio autoriteto institucijomis	Neigiamas ekonominės sistemos vertinimas	Respondento amžius	Šeimos pajamos
N		1113	1113	1086	1113	1053
Normal Parameter ^{a,b}	Mean	,0000000	,0000000	8,85	48,62	4,87
	Std. Deviation	1,00000000	1,00000000	4,471	17,452	1,835
Most Extreme Differences	Absolute	,027	,039	,180	,058	,140
	Positive	,027	,024	,180	,053	,140
	Negative	-,013	-,039	-,080	-,058	-,099
Kolmogorov-Smirnov Z		,917	1,313	5,935	1,928	4,538
Asymp. Sig. (2-tailed)		,369	,064	,000	,001	,000

a. Test distribution is Normal.

b. Calculated from data.

Sintaksės komandos

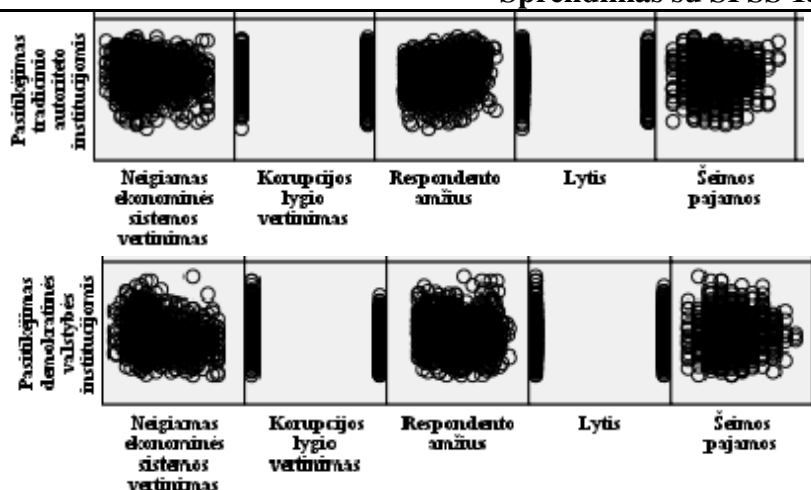
NPARTESTS

/K-S(NORMAL)= pasidem pasitrad ekonom amzius pajamos

/MISSING ANALYSIS.

Nubraižome kiekvieno priklausomo ir nepriklausomo kintamojo taškinės diagramas, kad įsitikintumėme, kad tiesinis ryšys egzistuoja ir kad nėra akivaizdžių išskirčių. Kai nagrinėjame daug kintamųjų, patogu braižyti matricų formos diagramas: **Graphs → Legacy dialogs → Scatter/Dot → Matrix Scatter → Define** - pasirenkame nagrinėjamus kintamuosius **pasidem**, **pasitrad**, **ekonom**, **korupcij**, **amzius**, **lytis**, **pajamos** → **OK**. Nagrinėjamu atveju, šios diagramos akivaizdžiai nepademonstruoja tiesinio ryšio (pateikta tik dalis matricų formos diagramos), tačiau didelių išskirčių duomenyse nepastebime.

Sprendimas su SPSS 15.0



GRAPH

/SCATTERPLOT(MATRIX)=ekonom korupcij amzius lytis pajamos pasidem pasitrad

/MISSING=VARIABLEWISE .

Patikriname priklausomo (-ų) ir nepriklausomų kintamųjų tiesinio ryšio egzistavimo sąlygą kitu, ne grafiniu būdu: **Analyze → Correlate → Bivariate...**, pasirenkame **pasidem, pasitrad, ekonom, korupcij, amzius, lytis, pajamos** kintamuosius ir Pirsono koreliacijos koeficientą (pažymėtas pagal nutylėjimą) → OK.

Correlations

		Pasitikėjimas demokratinės valstybės institucijomis	Pasitikėjimas tradicinio autoriteto institucijomis
Pearson Correlation	Neigiamas ekonominės sistemos vertinimas	-,241**	-,119**
	Korupcijos lygio vertinimas	-,249**	-,077*
	Respondento amžius	-,112**	,222**
	Lytis	-,043	-,058
	Šeimos pajamos	,051	-,042

**. Correlation is significant at the 0.01 level (2-tailed).

*. Correlation is significant at the 0.05 level (2-tailed).

Sintaksės komandos

CORRELATIONS

/VARIABLES=pasidem pasitrad ekonom korupcij amzius lytis pajamos

/PRINT=TWOTAIL NOSIG

/MISSING=PAIRWISE .

Analizės veiksmai: atliekame regresinę analizę

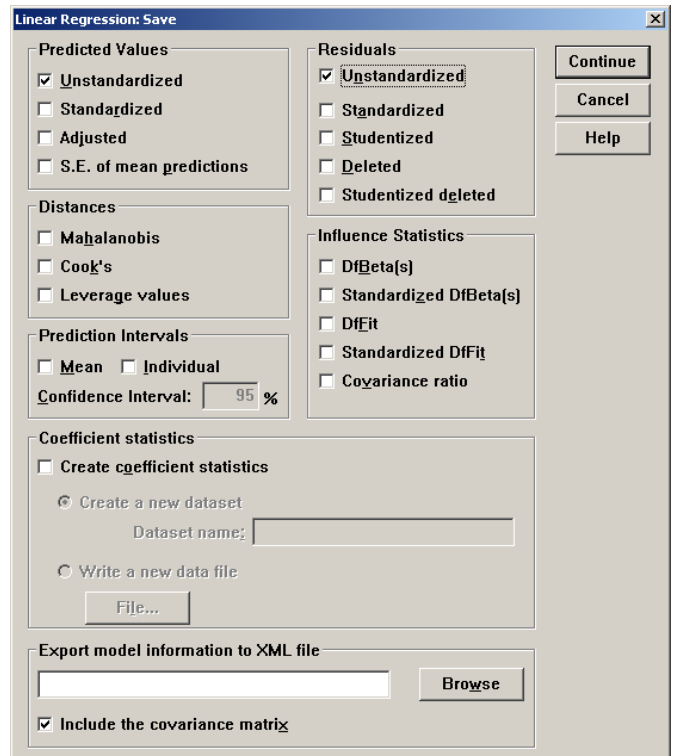
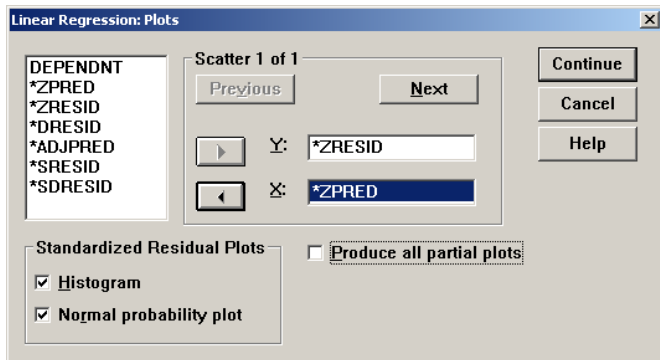
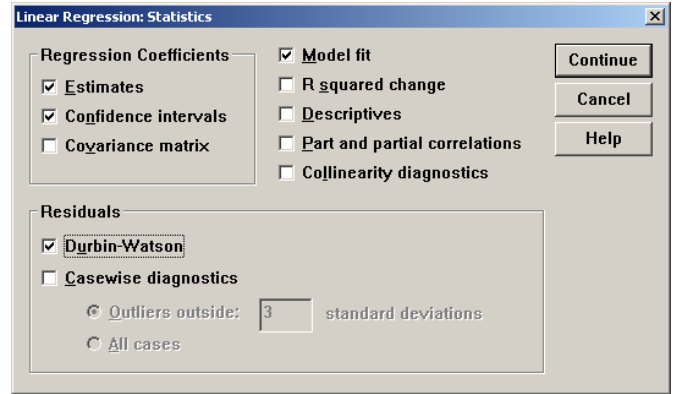
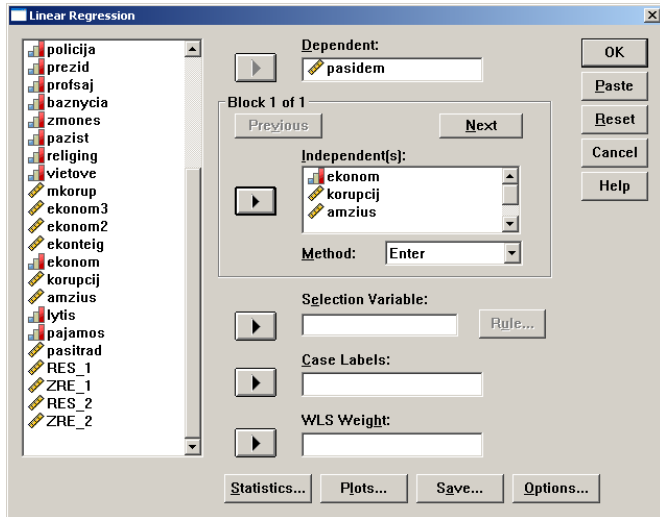
Meniu punktai: Analyze→Regression→Linear...

Į „Dependant“ sąrašą perkeliame **pasidem** (kartojant analizę pakeičiame į **pasitrad**) kintamąjį, į „Independent(s)“ sąrašą perkeliame kintamuosius **ekonom, korupcija** ir **amzius**; „Statistics“ pažymime „Estimates“, „Confidence intervals“, „Model fit“, „Durbin-Watson“ – spaudžiame „Continue“; „Plots“ lange pasirenkame *ZPRED ir perkeliame į X ašį, o *ZRESID – į Y ašį, „Standardized Residual Plots“ skyrelyje pažymime „Histogram“ ir „Normal probability plot“, spaudžiame „Continue“; „Save“ lange pažymime „Predicted values: Unstandardized“ ir „Residuals Unstandardized“, spaudžiame „Continue“.

Sprendimas su SPSS 15.0

→ spaudžiame „OK“.

Meniu punktai:



Sintaksės komandos

```
REGRESSION
/DESCRIPTIVES MEAN STDDEV CORR SIG N
/MISSING LISTWISE
/STATISTICS COEFF OUTS CI R ANOVA COLLIN TOL
/CRITERIA=PIN(.05) POUT(.10)
/NOORIGIN
/DEPENDENT pasidem
/METHOD=ENTER ekonom korupcij amzius
/PARTIALPLOT ALL
```

```
REGRESSION
/DESCRIPTIVES MEAN STDDEV CORR SIG N
/MISSING LISTWISE
/STATISTICS COEFF OUTS CI R ANOVA COLLIN TOL
/CRITERIA=PIN(.05) POUT(.10)
/NOORIGIN
/DEPENDENT pasitrad
/METHOD=ENTER ekonom korupcij amzius
/PARTIALPLOT ALL
```

Sprendimas su SPSS 15.0

/SCATTERPLOT=(*ZRESID , *ZPRED) (pasidem , *ZPRED) (*ZRESID , pasidem)
/RESIDUALS DURBIN HIST(ZRESID) NORM(ZRESID)
/SAVE RESID ZRESID .

/SCATTERPLOT=(*ZRESID , *ZPRED) (pasitrad , *ZPRED) (*ZRESID , pasitrad)
/RESIDUALS DURBIN HIST(ZRESID) NORM(ZRESID)
/SAVE RESID ZRESID .

Gauname tokius rezultatus :

Model Summary^b

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	Durbin-Watson
1	,340 ^a	,116	,113	,92215035	1,731

a. Predictors: (Constant), Respondento amžius, Ekonominės sistemos vertinimas, Korupcijos lygio vertinimas

b. Dependent Variable: Pasitikėjimas demokratinės valstybės institucijomis

Model Summary^b

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	Durbin-Watson
1	,264 ^a	,070	,067	,97183214	1,833

a. Predictors: (Constant), Respondento amžius, Ekonominės sistemos vertinimas, Korupcijos lygio vertinimas

b. Dependent Variable: Pasitikėjimas tradicinio autoriteto institucijomis

ANOVA^b

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	116,424	3	38,808	45,637	,000 ^a
	Residual	891,179	1048	,850		
	Total	1007,602	1051			

a. Predictors: (Constant), Respondento amžius, Ekonominės sistemos vertinimas, Korupcijos lygio vertinimas

b. Dependent Variable: Pasitikėjimas demokratinės valstybės institucijomis

ANOVA^b

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	74,266	3	24,755	26,211	,000 ^a
	Residual	989,792	1048	,944		
	Total	1064,057	1051			

a. Predictors: (Constant), Respondento amžius, Ekonominės sistemos vertinimas, Korupcijos lygio vertinimas

b. Dependent Variable: Pasitikėjimas tradicinio autoriteto institucijomis

Coefficients^a

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	95% Confidence Interval for B		Collinearity Statistics	
		B	Std. Error	Beta			Lower Bound	Upper Bound	Tolerance	VIF
1	(Constant)	,797	,100		7,999	,000	,602	,993		
	Neigiamas ekonominės sistemos vertinimas	-,045	,006	-,204	-6,962	,000	-,057	-,032	,980	1,021
	Korupcijos lygio vertinimas	-,445	,061	-,214	-7,259	,000	-,565	-,325	,974	1,027
	Respondento amžius	-,006	,002	-,103	-3,527	,000	-,009	-,003	,992	1,008

a. Dependent Variable: Pasitikėjimas demokratinės valstybės institucijomis

Coefficients^a

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	95% Confidence Interval for B		Collinearity Statistics	
		B	Std. Error	Beta			Lower Bound	Upper Bound	Tolerance	VIF
1	(Constant)	-,369	,105		-3,515	,000	-,575	-,163		
	Neigiamas ekonominės sistemos vertinimas	-,026	,007	-,117	-3,894	,000	-,040	-,013	,980	1,021
	Korupcijos lygio vertinimas	-,174	,065	-,082	-2,700	,007	-,301	-,048	,974	1,027
	Respondento amžius	,013	,002	,228	7,619	,000	,010	,017	,992	1,008

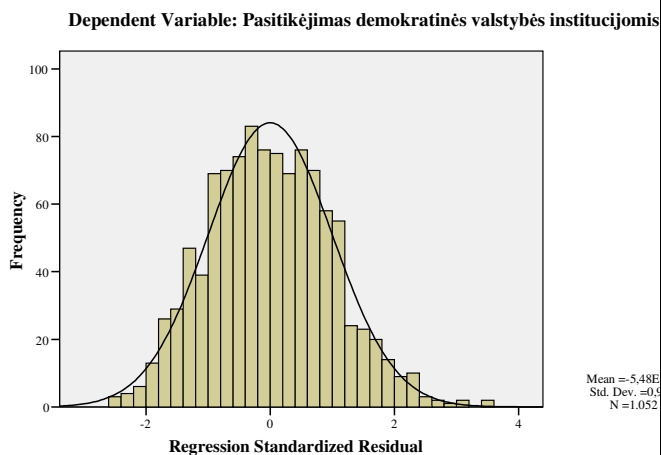
a. Dependent Variable: Pasitikėjimas tradicinio autoriteto institucijomis

Regresijos prielaidų tikrinimas

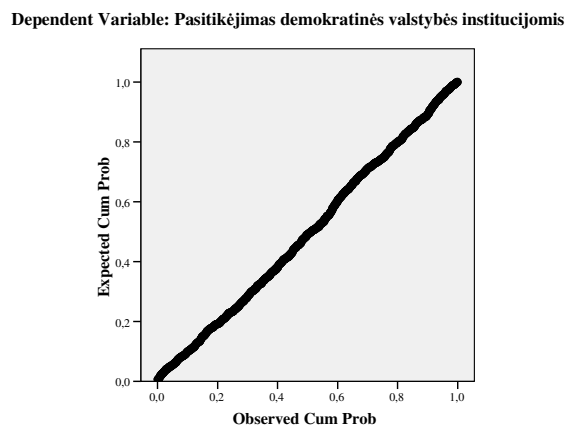
Įvertinime liekanų skirstinio normalumą grafiškai ir remdamiesi Kolmogorovo-Smirnovo suderinamumo kriterijumi

Sprendimas su SPSS 15.0

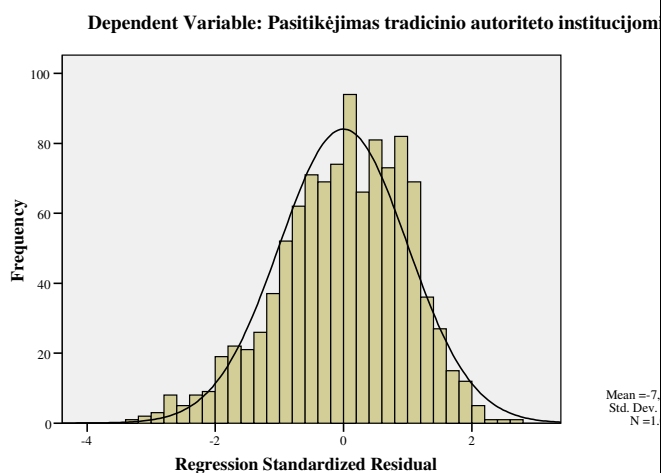
Histogram



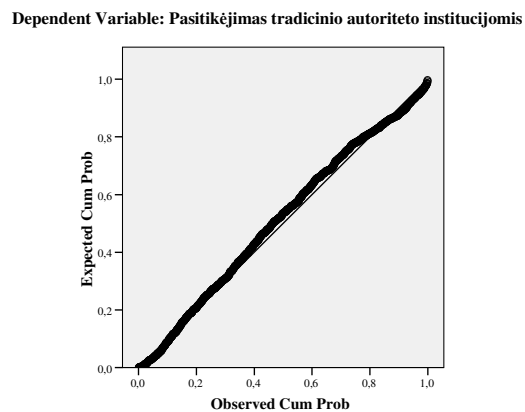
Normal P-P Plot of Regression Standardized Residual



Histogram



Normal P-P Plot of Regression Standardized Residual



Regresinės analizės procedūros metu išsaugotų liekanų normalumą galime ištirti atlikdami neparametrinį Kolmogorovo-Smirnovo kriterijaus tikrinimą RES_1 ir RES_2 kintamiesiems (detaliau apie suderinamumo hipotezės tikrinimą žr. vienfaktorinės dispersinės analizės pavyzdį).

Sprendimas su SPSS 15.0

One-Sample Kolmogorov-Smirnov Test

		Unstandardized Residual	Unstandardized Residual
N		1052	1052
Normal Parameters ^{a,b}	Mean	,0000000	,0000000
	Std. Deviation	,92083330	,97044414
Most Extreme Differences	Absolute	,022	,041
	Positive	,022	,026
	Negative	-,013	-,041
Kolmogorov-Smirnov Z		,716	1,319
Asymp. Sig. (2-tailed)		,684	,062

a. Test distribution is Normal.

b. Calculated from data.

Sintaksės komandos

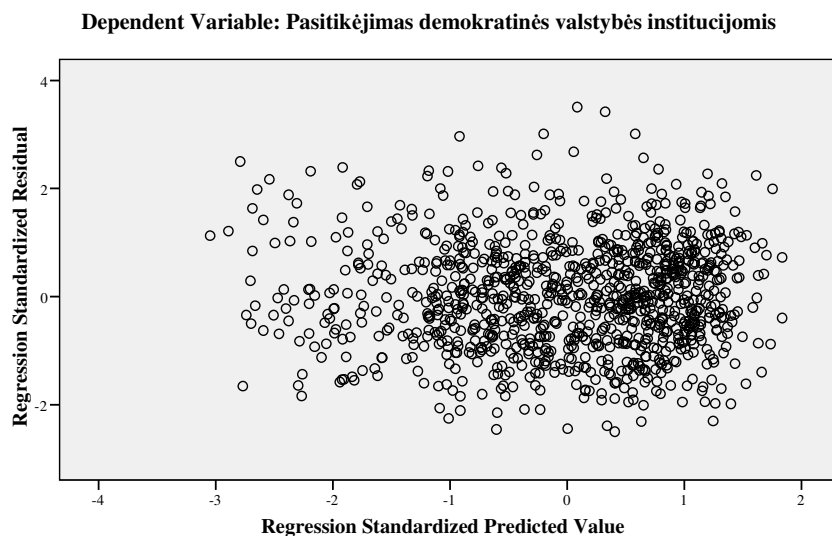
NPART TESTS

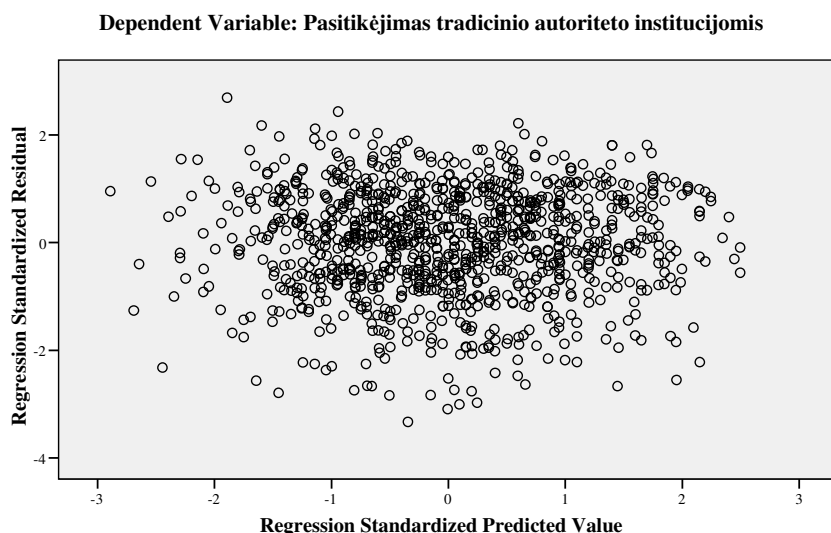
/K-S(NORMAL)= RES_1 RES_2

/MISSING ANALYSIS.

Vizualiai įvertinime liekanų dispersijos pastovumą (homoskedastiškumą):

Scatterplot



Scatterplot

7.1.1.2 Pradinių regresijos prielaidų tikrinimas

Asimptotinis reikšmingumo lygmuo tik priklausomiems intervaliniams kintamiesiems yra daugiau negu 0,05, t.y., tik priklausomų kintamųjų skirstinys yra suderinamas su normaliuoju. Griežtesnioji sąlyga, kad *visi* intervaliniai kintamieji turi būti normalaus skirstinio, yra netenkinama. Vadinasi, atlikus regresinę analizę reikės tikrinti, ar tenkinama silpnesnioji - standartizuotų liekanų skirstinio normalumo - sąlyga.

Alternatyviai, galima bandyti nenormalaus skirstinio intervalinius kintamuosius transformuoti į tinkamą tiesinei regresinei analizei funkcinę formą, pvz., ištraukti kvadratinę šaknį, kelti įvairiais laipsniais, perkoduoti į dvireikšmius kintamuosius ir pan.

Pastebime, kad pasitikėjimo tiek demokratinėmis, tiek tradicinio autoriteto institucijomis lygis nėra susijęs su gyventojų šeimos pajamomis ir lytimi, todėl šių kintamųjų nėra prasmės įtraukti į regresinę analizę. Taip pat, reikia atkreipti dėmesį, kad **ekonom** kintamajame užkoduotas neigiamas esamos ekonominės sistemos vertinimas, todėl koreliacija su pasitikėjimu yra neigiama. Atitinkamai, reikia nepamiršti išvadose nurodyti, šie kintamieji nenuspėja arba nepaaiškina pasitikėjimo institucijomis lygio.

Tad į tolesnę tiesinę regresinę analizę įtraukiame respondento amžių ir kyšių ėmimo paplitimo (nedidelis ar didelis) vertinimą.

7.1.1.3 Regresinės analizės atlikimas

Nagrinėjame pasitikėjimo demokratinėmis (antru atveju – tradicinio autoriteto) institucijomis priklausomybę nuo neigiamo ekonomikos sistemos vertinimo, respondento amžiaus ir korupcijos lygio suvokimo.

Lentelėse „Modelio santrauka“ („Model Summary“) išvesti bendrieji regresijos modelio parametrai. Mažas koreguotas R^2 reiškia, kad modelis nėra labai geras, nes paaiškina tik apie 11% priklausomo kintamojo sklaidos apie jo vidurkį pirmu atveju ir apie 7% antru atveju.

Lentelėje „ANOVA“ aukštas reikšmingumo lygmuo („Sig.“) parodo, kad modelis tinkamas tiesinei regresinei analizei.

Lentelėje „Regresijos koeficientai“ („Coefficients“) pateikti regresinių lygčių nestandartizuoti („B“) ir standartizuoti („beta“) koeficientai yra statistškai reikšmingi, t.y., turi nepriklausomos įtakos priklausomojo kintamojo lygiui. Tolerancijos (angl. „tolerance“) įverčiai parodo, kad nepriklausomi kintamieji yra silpnai tarpusavyje susiję (nėra multikolinearumo): tolerancijos kriterijus yra toli nuo 0 (0 reikštų, kad kintamasis yra tiesinė kitų nepriklausomų kintamųjų kombinacija ir nė kiek neprideda prie priklausomojo kintamojo nuspėjimo), visi kiti į analizę įtraukti kintamieji (pirmo pavyzdžio atveju – korupcijos lygio vertinimo ir ekonomikos vertinimo kintamieji) paaiškina 0,8% $[(1-0,992)*100\%]$ respondento amžiaus dispersijos; korupcijos lygio ir ekonomikos vertinimo kintamųjų atveju kiek didesnė – 2 – 2,6 % šių kintamųjų dispersijos dalis – paaiškina kiti kintamieji. VIF (angl. „variance inflator factor“ – dispersijos mažėjimo daugiklis) tiesiogiai išplaukia iš tolerancijos įverčio, kuo jis didesnis, tuo yra daugiau tarpusavyje susiję nepriklausomi kintamieji.

Užrašome regresijos lygtis abiemis pasitikėjimo institucijomis atvejams:

- Pasitikėjimas demokratinėmis institucijomis = $-0,045 * \text{neigiamas ekonominės sistemos vertinimas} - 0,445 * \text{korupcijos lygio vertinimas} - 0,006 * \text{respondento amžius}$
- Pasitikėjimas tradicinio autoriteto institucijomis = $-0,026 * \text{neigiamas ekonominės sistemos vertinimas} - 0,174 * \text{korupcijos lygio vertinimas} + 0,013 * \text{respondento amžius}$

Galime daryti išvadą, kad pasitikėjimo abiemis institucijų grupėms lygį mažina neigiamas ekonominės sistemos vertinimas (pasitikėjimas sumažėja 0,045 balo demokratinėmis institucijomis, 0,026 balo tradicinio autoriteto institucijomis, neigiamam ekonominės sistemos vertimui padidėjus 1 vertinimo balu), pasitikėjimo lygis demokratinėmis institucijomis stipriai priklauso nuo suvokiamo korupcijos lygio (pasitikėjimas demokratinėmis institucijomis sumažėja 0,445 balo, jei gyventojas mano, kad korupcija yra paplitusi), pasitikėjimas tradicinio autoriteto institucijomis nuo suvokiamo korupcijos lygio priklauso silpniau, bet taip pat neigiamai; vyresni gyventojai linkę mažiau pasitikėti demokratinėmis institucijomis (dešimt metų didesnio amžiaus gyventojų pasitikėjimas demokratinėmis institucijomis sumažėja 0,06 balo), didėjant gyventojų amžiui, jie daugiau pasitiki tradicinio autoriteto institucijomis (dešimt metų padidėjus respondentų amžiui, jų pasitikėjimas tradicinio autoriteto institucijomis padidėja 0,15 balo).

7.1.1.4 Regresijos prielaidų tikrinimas

Liekanų skirstinio normalumą įvertinime grafiškai ir patikrindami suderinamumo hipotezę. Nagrinėjamų pavyzdžių atveju randame, kad liekanos pasiskirstę pagal normalųjį skirstinį ($p > 0,05$) su 95% statistine garantija.

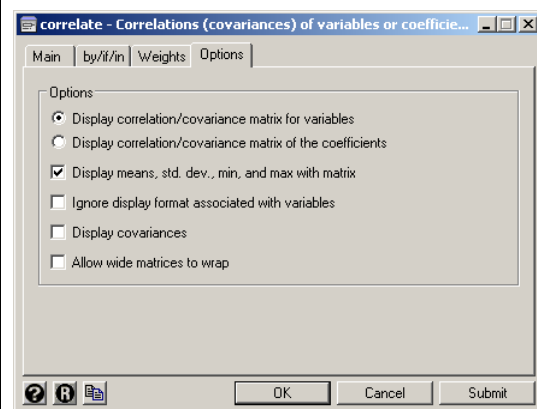
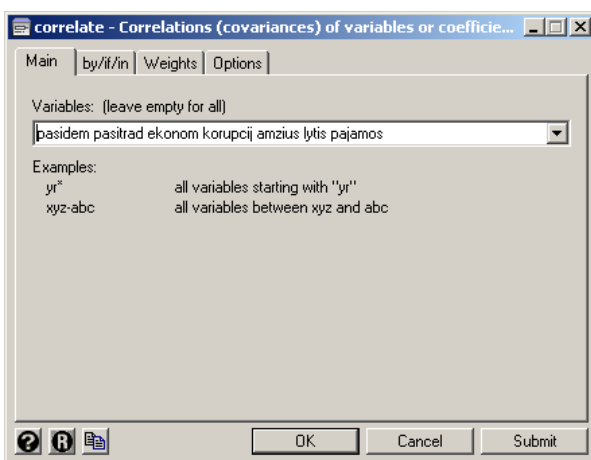
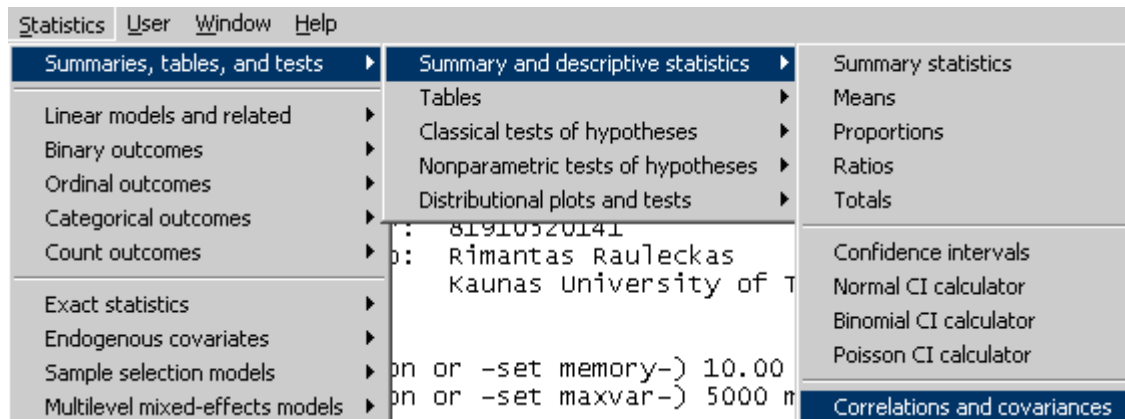
Liekanų dispersijos pastovumo (homoskedastiškumo) įvertinimą atliekame grafiškai (galima naudoti ir formalius metodus). Vizualiai vertinant abiejų pavyzdžių liekanų dispersijų pastovumą, galima teigti, kad nėra liekanų debesies formos asimetrijos ar kitos aiškios priklausomybės nuspėtų priklausomo kintamojo reikšmių atžvilgiu.

7.1.2. Daugialypė tiesinė regresinė analizė su Stata

Sprendimas su Stata SE 10.0

Parengiamieji veiksmai: tiesinio ryšio tarp kintamųjų stiprumas

Sprendimas su Stata SE 10.0



Sintaksės komandos

correlate pasidem pasitrad ekonom korupcij amzius lytis pajamos, means

Gauname tokius rezultatus :

variable	Mean	Std. Dev.	Min	Max
pasidem	-.0180815	.9797884	-2.49005	3.236419
pasitrad	-.007805	1.006565	-3.120634	2.291396
ekonom	8.88833	4.509429	1	21
korupcij	.3239437	.4682144	0	1
amzius	48.55332	17.22365	18	90
lytis	.4466801	.4973991	0	1
pajamos	4.910463	1.839762	1	11

	pasidem	pasitrad	ekonom	korupcij	amzius	lytis	pajamos
pasidem	1.0000						
pasitrad	-0.0137	1.0000					
ekonom	-0.2426	-0.1275	1.0000				
korupcij	-0.2515	-0.0803	0.1493	1.0000			
amzius	-0.1357	0.2100	0.0369	0.0959	1.0000		
lytis	-0.0328	-0.0588	-0.0644	0.0440	-0.0184	1.0000	
pajamos	0.0634	-0.0337	-0.1813	-0.1791	-0.2902	0.0691	1.0000

Sprendimas su Stata SE 10.0

[vertiname intervalinių kintamųjų normalumą

Žr. Kolmogorovo-Smirnovo suderinamumo kriterijaus taikymo pavyzdį 3 sk. Skaitytojas gali patikrinti, kad tik **pasidek** ir **pasitrad** kintamųjų atveju nulinė hipotezė apie skirtingo normalumą yra neatmetama.

Sintaksės komandos

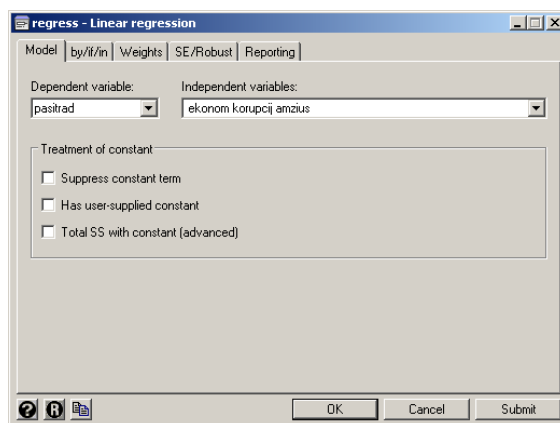
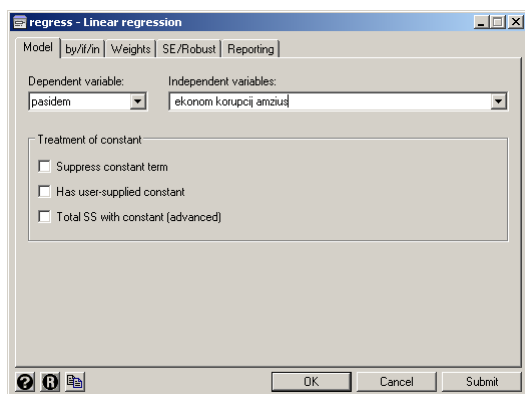
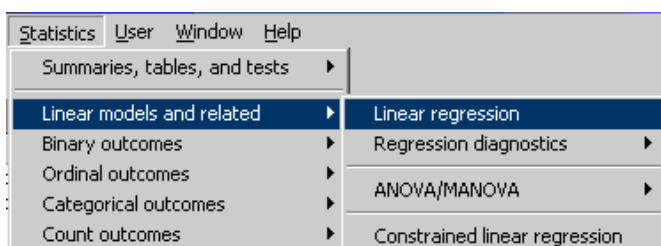
ksmirnov ekonom = normal(ekonom)

ksmirnov amzius = normal(amzius)

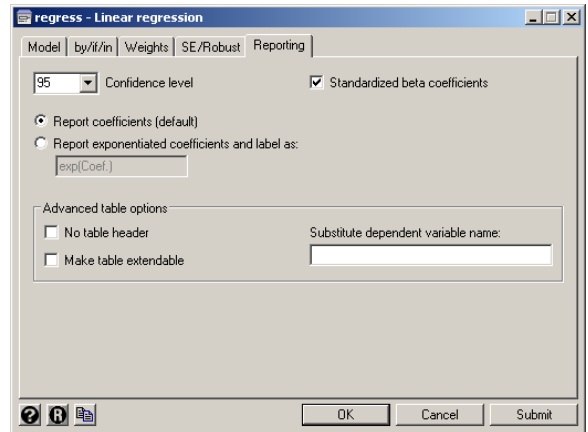
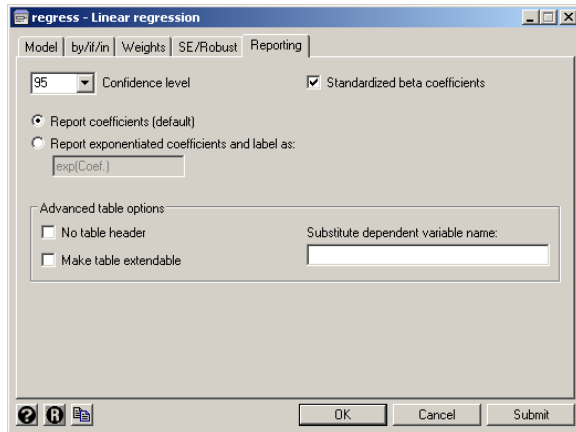
....

Analizės veiksmai: atliekame regresinę analizę

Atliekame regresinės analizės procedūrą su **pasidem** kintamuoju. Antrą kartą darydami analizę pakeičiame priklausomąjį kintamąjį į **pasitrad**. Statistics → Linear models and related → Linear regression. Pasirenkame priklausomą **pasidem** (antroje analizėje – **pasitrad**) kintamuosius ir nepriklausomus **ekonom**, **korupcij**, **amzius** kintamuosius. Užsakome standartizuotus (beta) koeficientus „Reporting“ skiltyje, spaudžiame „Submit“:



Sprendimas su Stata SE 10.0



Sintaksės komandos

regress pasidem ekonom korupcij amzius, beta

regress pasitrad ekonom korupcij amzius, beta

Gauname tokius rezultatus lange:

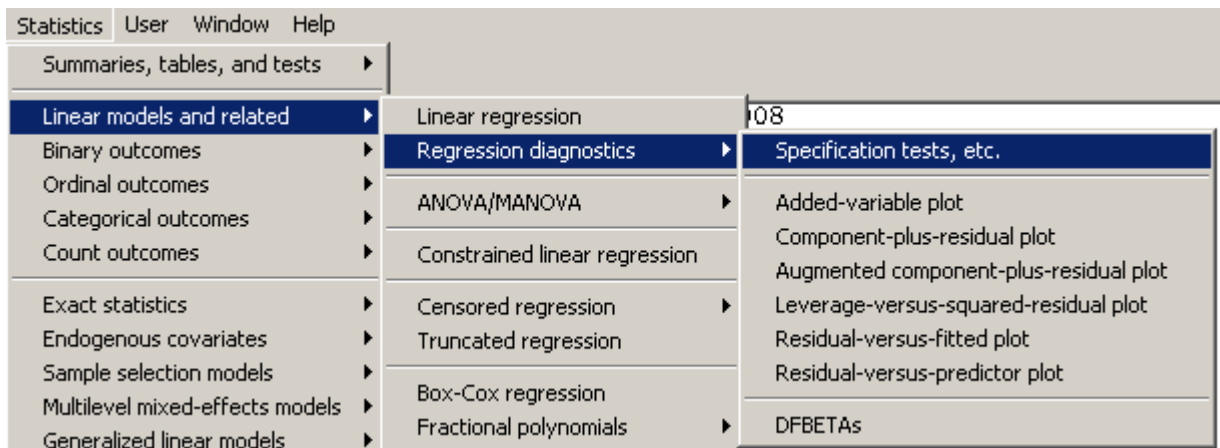
Source	SS	df	MS		
Model	116.423844	3	38.8079481	Number of obs =	1052
Residual	891.178605	1048	.850361264	F(3, 1048) =	45.64
Total	1007.60245	1051	.958708325	Prob > F =	0.0000
				R-squared =	0.1155
				Adj R-squared =	0.1130
				Root MSE =	.92215

pasidem	Coef.	Std. Err.	t	P> t	Beta
ekonom	-.0446826	.0064177	-6.96	0.000	-.2043495
korupcij	-.4447678	.0612684	-7.26	0.000	-.2136683
amzius	-.0058278	.0016525	-3.53	0.000	-.102858
_cons	.7971932	.0996636	8.00	0.000	.

Source	SS	df	MS		
Model	74.2657473	3	24.7552491	Number of obs =	1052
Residual	989.791689	1048	.944457719	F(3, 1048) =	26.21
Total	1064.05744	1051	1.01242382	Prob > F =	0.0000
				R-squared =	0.0698
				Adj R-squared =	0.0671
				Root MSE =	.97183

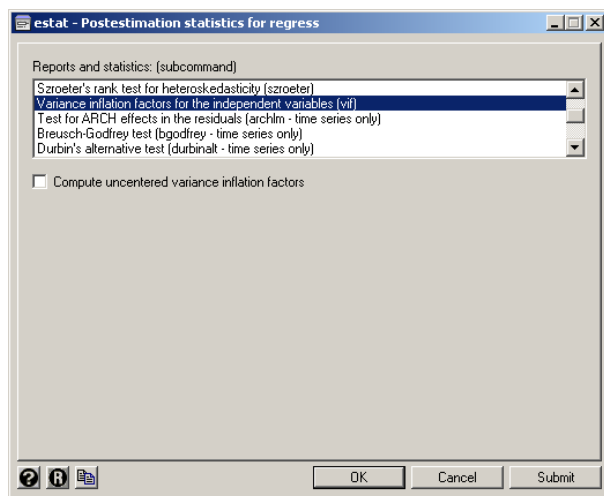
pasitrad	Coef.	Std. Err.	t	P> t	Beta
ekonom	-.0263349	.0067634	-3.89	0.000	-.1172004
korupcij	-.1743474	.0645693	-2.70	0.007	-.081505
amzius	.013269	.0017415	7.62	0.000	.227894
_cons	-.3691544	.1050331	-3.51	0.000	.

Atliekame regresinės analizės prielaidų tikrinimą:



[vertiname multikolinearumą (VIF kriterijus)]

Sprendimas su Stata SE 10.0



Variable	VIF	1/VIF
korupcij	1.03	0.974152
ekonom	1.02	0.979692
amzius	1.01	0.992144
Mean VIF	1.02	

Sintaksės komandos

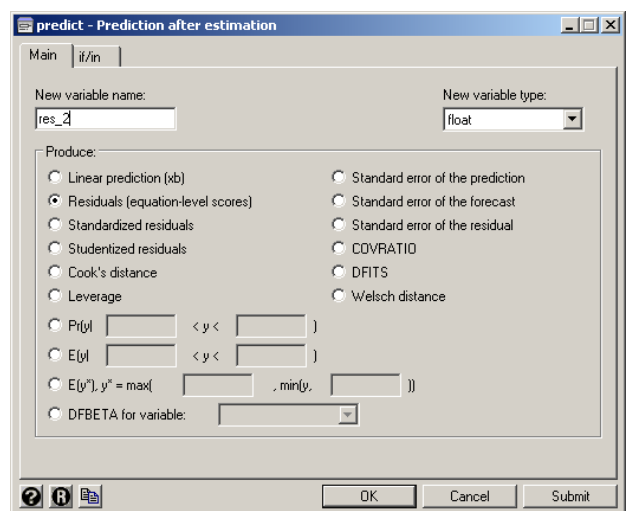
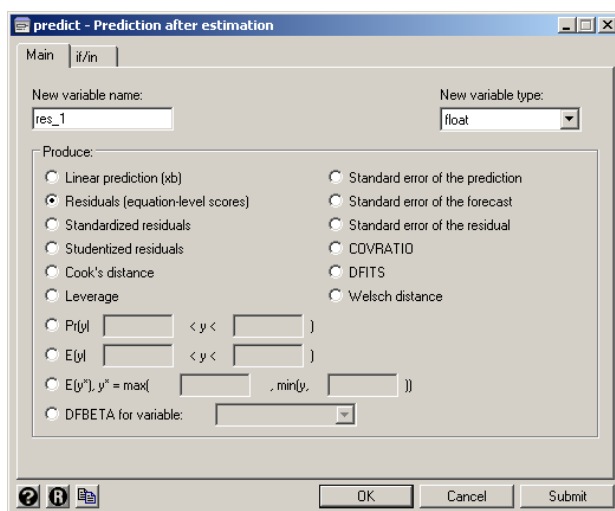
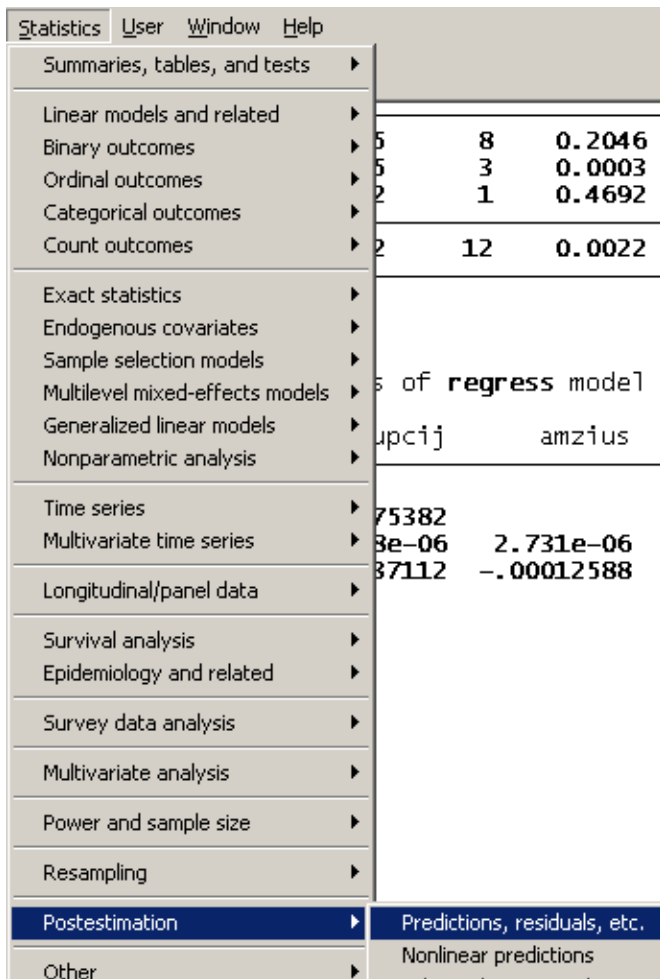
estat vif

Išvada: liekanų autokoreliacija yra labai silpna

Patikriname liekanų skirstinio normalumo sąlygą:

Išsaugojame regresijos liekanas duomenų faile:

Sprendimas su Stata SE 10.0



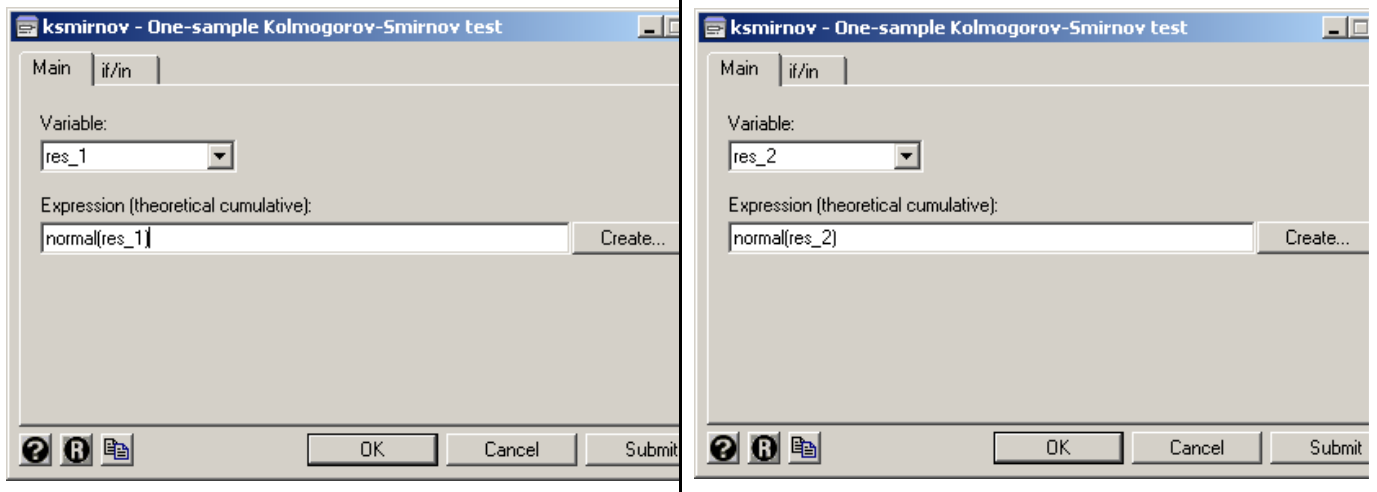
Sintaksės komandos

predict res_1, residuals

predict res_2, residuals

Taikome Kolmogorovo-Smirnovo neparametrinį kriterijų:

Sprendimas su Stata SE 10.0



Sintaksės komandos

ksmirnov res_1 = normal(res_1)

ksmirnov res_2 = normal(res_2)

One-sample kolmogorov-smirnov test against theoretical distribution
normal(res_1)

Smaller group	D	P-value	Corrected
res_1:	0.0305	0.141	
Cumulative:	-0.0261	0.239	
Combined K-S:	0.0305	0.281	0.267

Note: ties exist in dataset;
there are 1051 unique values out of 1052 observations.

One-sample kolmogorov-smirnov test against theoretical distribution
normal(res_2)

Smaller group	D	P-value	Corrected
res_2:	0.0329	0.103	
Cumulative:	-0.0385	0.044	
Combined K-S:	0.0385	0.088	0.082

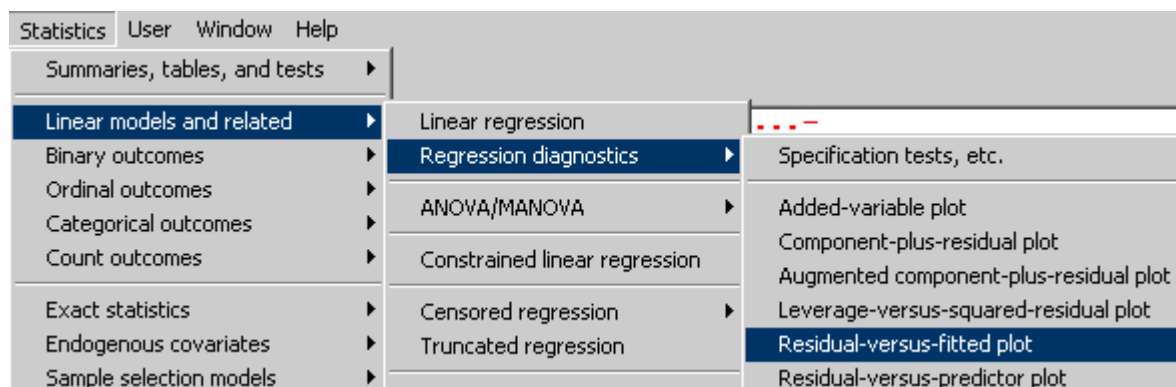
Note: ties exist in dataset;
there are 1051 unique values out of 1052 observations.

Patikriname liekanų skirstinio suderinamumo hipotezę remdamiesi Kolmogorovo-Smirnovo neparametriniu kriterijumi (žr. daugiau vienfaktorinės dispersinės analizės pavyzdį):

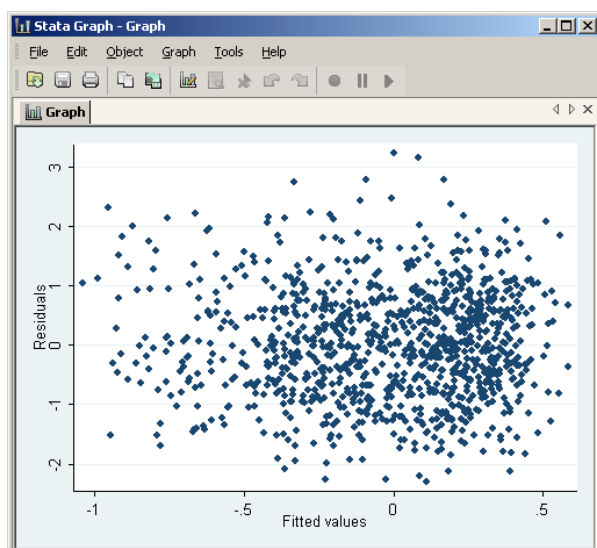
Išvada: abejais atvejais liekanų skirstinys yra normalusis ($p > 0,05$).

Liekanų priklausomybės nuo nuspėto priklausomo kintamojo reikšmių grafikas sąlyginės liekanų dispersijos pastovumui įvertinti:

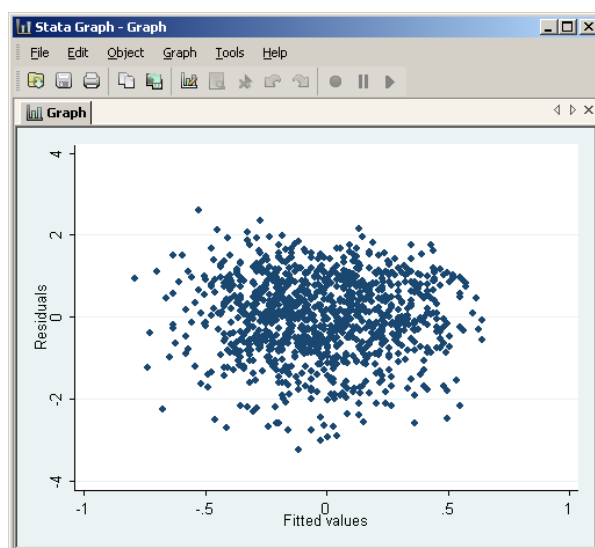
Sprendimas su Stata SE 10.0



pasidem



pasitrad



Sintaksės komandos

rvfplot

rvfplot

8. KLAŠTERINĖ ANALIZĖ

Šiame skyriuje pateikiami pavyzdžiai, kaip su SPSS ir Stata atlikti klasterinę analizę. Bus nagrinėjamas tik hierarchinio klasterizavimo metodas.

8.1. Hierarchinė klasterinė analizė

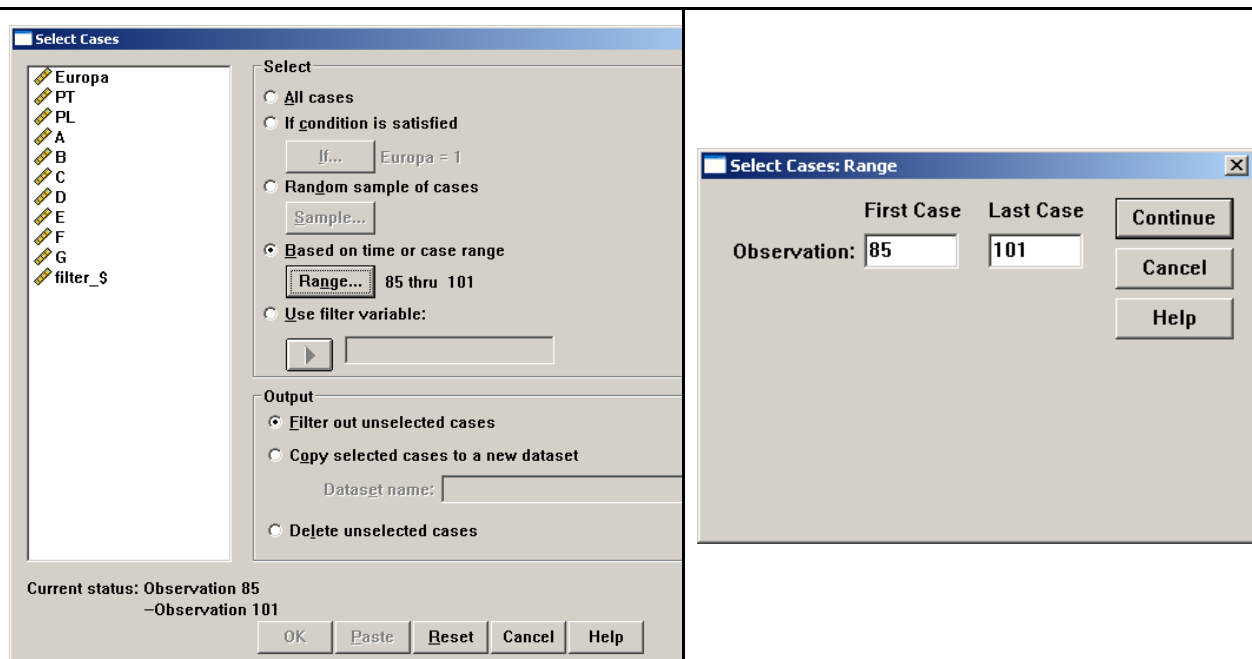
Ištirsime, kokias grupes sudaro šalys pagal 2007 m. Freedom House pateiktus pasaulio šalių laisvės tyrimo duomenis, naudojant hierarchinį klasterizavimo metodą. Tikimės, kad remiantis šešiomis įvairias laisves ir teises įvertinančiais, galime sugrupuoti šalis pagal jose esamą liberalios demokratijos lygį į mažai demokratiškas ir daugiau demokratiškas.

8.1.1. Hierarchinė klasterinė analizė su SPSS

Sprendimas su SPSS 15.0

Parengiamieji veiksmai

Visų pirma, pasirenkame dalį šalių, kad būtų galima vaizdžiau pateikti analizės rezultatus: Data→Select Cases →Based on time or case range: pasirenkame atvejus nuo 85 iki 101 (t.y., šalys, pagal abėcėlę prasidedančios J-L pirmomis raidėmis) →Continue →OK.

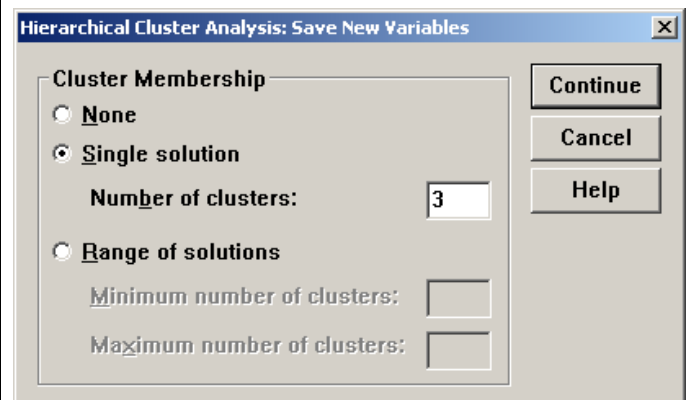
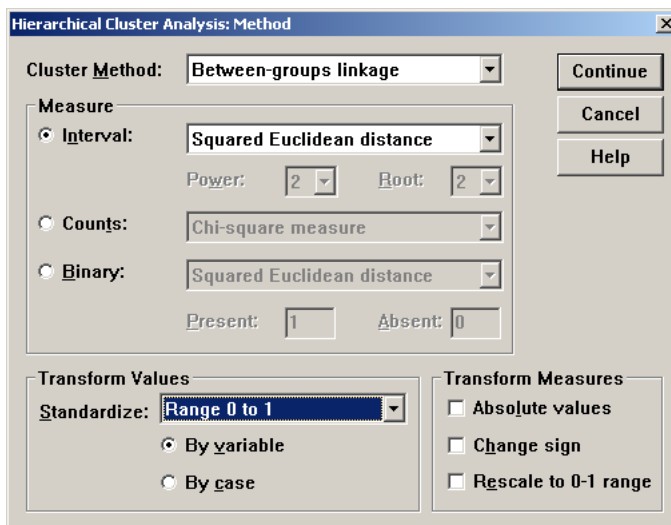
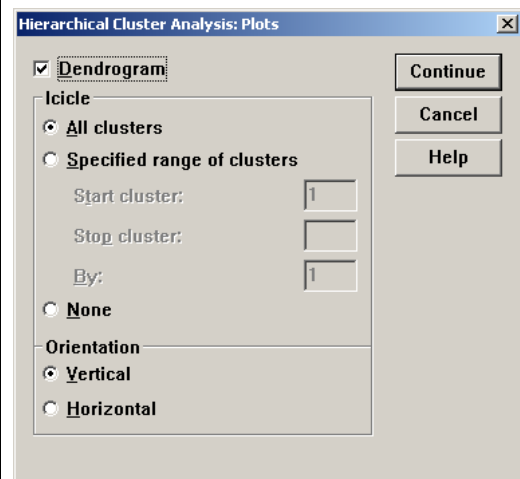
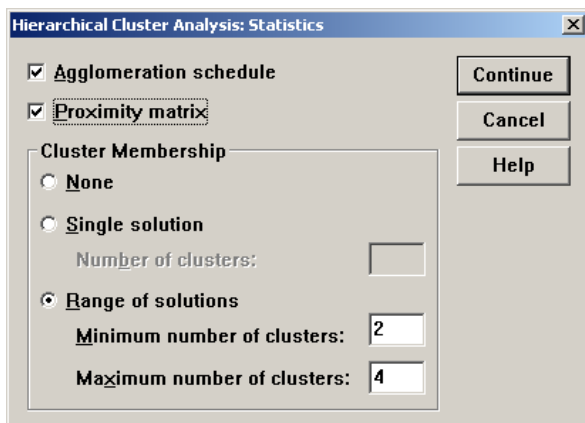
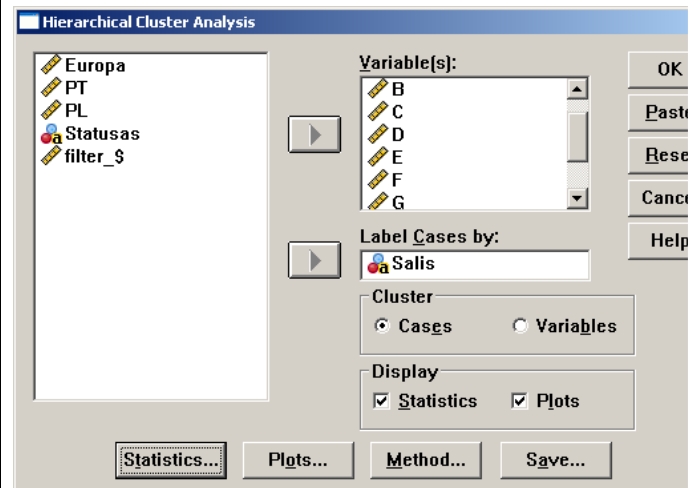
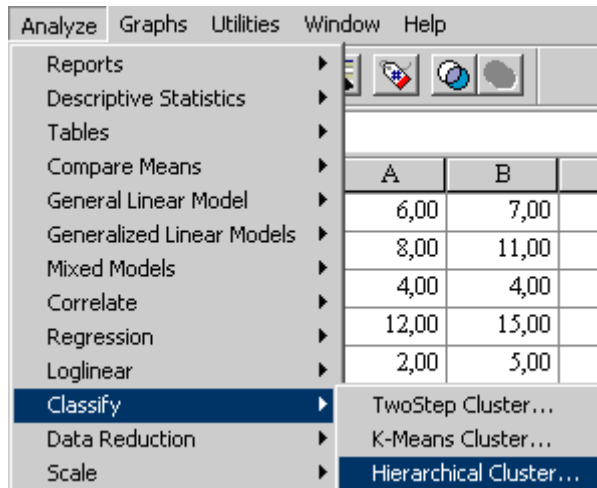


Klasterinės analizės veiksmai

Meniu punktai: Analyze→Classify→Hierarchical cluster... → „Variable(s)“: perkeliame kintamuosius **a, b, c, d, e, f, g** iš kairės pusės sąrašo į dešinį sąrašą; „Label Cases by“: pasirenkame kintamąjį **salis**; „Plots“: pasirenkame „Dendrogram“, „Icicle: All clusters“ → Continue → „Statistics“: nužymime

Sprendimas su SPSS 15.0

„Agglomeration Schedule“, pažymime „Proximity matrix“ ir „Cluster membership: Range of solutions“: mažiausiai 2 sprendiniai, daugiausiai 4 sprendiniai → Continue → „Method“: „Cluster method: Between groups linkage“, „Transform measures: rescale to 0-1 scale“ → Continue → „Save“: pasirenkame „Single solution: Number of clusters“: 3 → Continue – OK.



Sprendimas su SPSS 15.0

Sintaksės komandos

```

PROXIMITIES a b c d e f g
/MATRIX OUT ('C:\spssclus.tmp')
/VIEW= CASE
/MEASURE= SEUCLID
/PRINT NONE
/ID= salis
/STANDARDIZE= VARIABLE RESCALE .
CLUSTER
/MATRIX IN ('C:\spssclus.tmp')
/METHOD BAVERAGE
/ID=salis
/PRINT CLUSTER(2,4)
/PRINT DISTANCE
/PLOT DENDROGRAM VICICLE
/SAVE CLUSTER(3) .
ERASE FILE= 'C:\spssclus.tmp'.
    
```

Gauname tokius rezultatus :

Proximity Matrix																	
	Squared Euclidean Distance																
Case	1:Jamaica	2:Japan	3:Jordan	4:Kazakhstan	5:Kenya	6:Kiribati	7:Kuwait	8:Kyrgyzstan	9:Laos	10:Latvia	11:Lebanon	12:Lesotho	13:Liberia	14:Libya	15:Liechtenstein	16:Lithuania	17:Luxembourg
1:Jamaica	,000	,323	,982	1,527	,166	,355	,873	,730	3,199	,215	,539	,093	,287	3,669	,702	,258	,702
2:Japan	,323	,000	1,864	2,621	,717	,078	1,677	1,631	4,677	,110	1,391	,350	,884	5,184	,178	,058	,178
3:Jordan	,982	1,864	,000	,094	,599	2,093	,139	,118	,706	1,939	,139	,850	,371	,979	2,825	1,958	2,825
4:Kazakhstan	1,527	2,621	,094	,000	,958	2,875	,294	,230	,316	2,702	,365	1,404	,615	,507	3,801	2,739	3,801
5:Kenya	,166	,717	,599	,958	,000	,701	,450	,311	2,284	,594	,282	,170	,076	2,777	1,373	,638	1,373
6:Kiribati	,355	,078	2,093	2,875	,701	,000	1,879	1,722	5,012	,051	1,465	,446	,986	5,599	,194	,027	,194
7:Kuwait	,873	1,677	,139	,294	,450	1,879	,000	,133	1,028	1,761	,227	,659	,284	1,380	2,696	1,779	2,696
8:Kyrgyzstan	,730	1,631	,118	,230	,311	1,722	,133	,000	1,013	1,558	,069	,681	,193	1,355	2,558	1,635	2,558
9:Laos	3,199	4,677	,706	,316	2,284	5,012	1,028	1,013	,000	4,833	1,295	2,989	1,736	,071	6,246	4,859	6,246
10:Latvia	,215	,110	1,939	2,702	,594	,051	1,761	1,558	4,833	,000	1,279	,354	,874	5,401	,204	,023	,204
11:Lebanon	,539	1,391	,139	,365	,282	1,465	,227	,069	1,295	1,279	,000	,536	,227	1,657	2,129	1,356	2,129
12:Lesotho	,093	,350	,850	1,404	,170	,446	,659	,681	2,989	,354	,536	,000	,278	3,525	,814	,337	,814
13:Liberia	,287	,884	,371	,615	,076	,986	,284	,193	1,736	,874	,227	,278	,000	2,105	1,693	,905	1,693
14:Libya	3,669	5,184	,979	,507	2,777	5,599	1,380	1,355	,071	5,401	1,657	3,525	2,105	,000	6,826	5,444	6,826
15:Liechtenstein	,702	,178	2,825	3,801	1,373	,194	2,696	2,558	6,246	,204	2,129	,814	1,693	6,826	,000	,164	,000
16:Lithuania	,258	,058	1,958	2,739	,638	,027	1,779	1,635	4,859	,023	1,356	,337	,905	5,444	,164	,000	,164
17:Luxembourg	,702	,178	2,825	3,801	1,373	,194	2,696	2,558	6,246	,204	2,129	,814	1,693	6,826	,000	,164	,000

This is a dissimilarity matrix

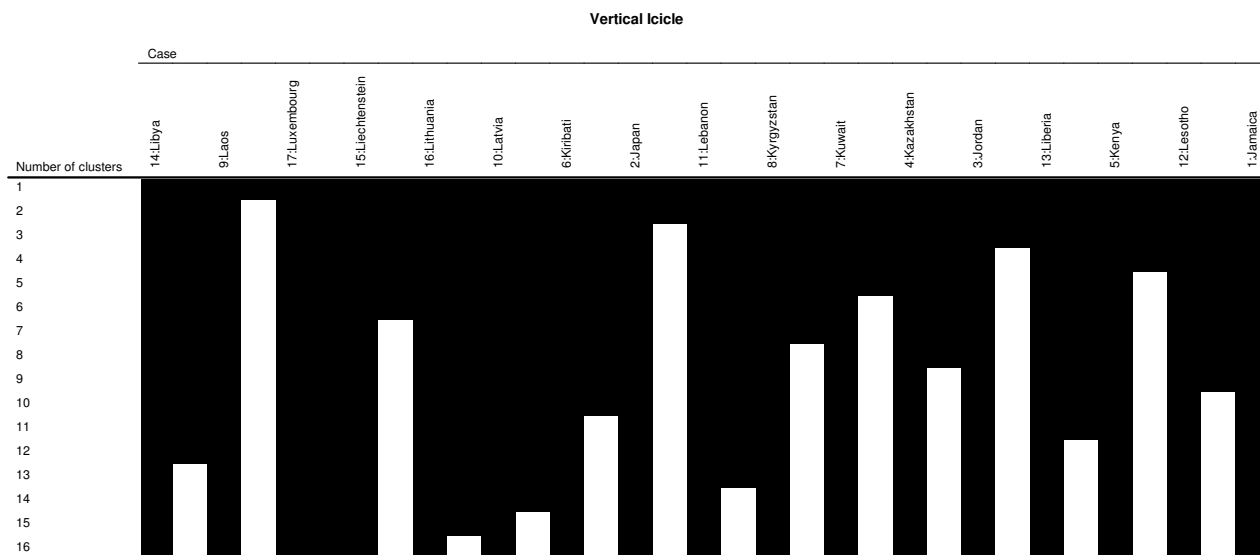
Agglomeration Schedule

Stage	Cluster Combined		Coefficients	Stage Cluster First Appears		Next Stage
	Cluster 1	Cluster 2		Cluster 1	Cluster 2	
1	15	17	,000	0	0	11
2	10	16	,023	0	0	3
3	6	10	,039	0	2	7
4	8	11	,069	0	0	10
5	9	14	,071	0	0	16
6	5	13	,076	0	0	13
7	2	6	,082	0	3	11
8	1	12	,093	0	0	13
9	3	4	,094	0	0	12
10	7	8	,180	0	4	12
11	2	15	,185	7	1	15
12	3	7	,214	9	10	14
13	1	5	,225	8	6	14
14	1	3	,653	13	12	15
15	1	2	1,556	14	11	16
16	1	9	3,323	15	5	0

Cluster Membership

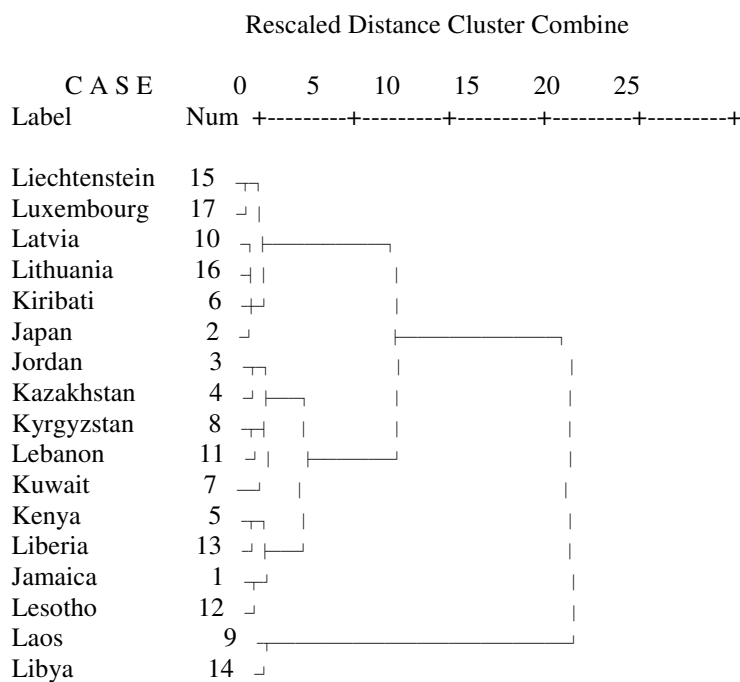
Case	4 Clusters	3 Clusters	2 Clusters
1:Jamaica	1	1	1
2:Japan	2	2	1
3:Jordan	3	1	1
4:Kazakhstan	3	1	1
5:Kenya	1	1	1
6:Kiribati	2	2	1
7:Kuwait	3	1	1
8:Kyrgyzstan	3	1	1
9:Laos	4	3	2
10:Latvia	2	2	1
11:Lebanon	3	1	1
12:Lesotho	1	1	1
13:Liberia	1	1	1
14:Libya	4	3	2
15:Liechtenstein	2	2	1
16:Lithuania	2	2	1
17:Luxembourg	2	2	1

Sprendimas su SPSS 15.0



***** HIERARCHICAL CLUSTER ANALYSIS *****

Dendrogram using Average Linkage (Between Groups)



Pradiniai klasterinės analizės pasirinkimai

Kintamųjų standartizavimas reikalingas tam, kad kintamųjų matavimo skalė būtų suvienodinta, nes skaičiuojant atstumus tarp atveju didesnės skalės kintamieji turės didesnę įtaką negu mažesnės skalės, pvz., rinkimų procesas vertinamas skalėje nuo 0 iki 12, o įstatymų valdžia – nuo 0 iki 16. Galima įvairiai standartizuoti kintamuosius, pvz., paverčiant nuo Z skaičiais su vidurkiu 0 ir

dispersija 1 ar transformuojant kiekvieną kintamąjį taip, kad jo reikšmės būtų tarp 0 ir 1. Galima eksperimentuoti su keliais transformacijos variantais.

Pasirenkame dažniausiai naudojamą klasterizuojant atvejus ir intuityviai priimtina atstumo matą - Euklidinio atstumo kvadratą.

Jungimo metodas apibrėžia, kaip bus sujungti klasteriai ir atvejai. Patariama bandyti kelis klasterio sudarymo būdus, o jeigu rezultatai labai skiriasi, duomenyse greičiausiai nėra atskirų klasterių. Dažniausiai patariama rinktis vidutinės jungties (SPSS-„between group linkage“, STATA – „average linkage“) jungimo metodą.

Rezultatai

Tais atvejais, kada yra nedaug atvejų, analizuojant klasterius gali praversti artumo matricos (Proximity Matrix), jungimo lentelė (Agglomeration Schedule), klasterių narystės (Cluster Membership) lentelės, dendrograma, klasterių grafikas (Vertical Icicle).

Jungimo lentelė parodo, kokie klasteriai ir kelintame žingsnyje yra sujungiami, atstumas tarp jų, sekančio žingsnio, kurio metu naujasis klasteris jungiamas su kitu artimiausiu klasteriu, numeris. Ši lentelė retai naudojama analizuojant duomenis.

Klasterių narystės lentelė dažniausiai yra pagrindinis klasterinės analizės rezultatas ir parodo pasirinkto klasterių skaičiaus sudėtį (iš SPSS meniu galima neužsakyti nė vieno, užsakyti visus, užsakyti keletą sprendinių). Nagrinėjamu atveju pasirinkti 2,3 ir 4 klasterių sprendiniai. Iš klasterių narystės pastebime (taip pat, žr. dendrograma), kad gana aiškiai atsiskiria grupės šalių, kurios laikomos demokratiškomis, mažiau demokratiškomis ir nedemokratiškomis.

Dendrograma ir klasterių grafikas vaizdžiai parodo, kokia seka jungiami klasteriai, klasterių narystė, o taip pat padeda nuspręsti, koks klasterių skaičius yra priimtinausias. Grafinis pateikimas padeda aiškiau pavaizduoti duomenyse esančius klasterius negu klasterių narystės lentelė.

Atlikus reikiamus pasirinkimus SPSS meniu, gauname lentelę, kurioje pateikta šalių priklausomybė grupėms (klasteriams), kai klasterių skaičius yra 2, 3 ir 4. Įdomiausias yra 3 klasterių sprendinys, nes jis atskiria laisvų, pusiau laisvų ir nelaisvų šalių grupes. Įdomu palyginti šį 3 klasterių sprendinį su Freedom House pateiktu laisvos, iš dalies laisvos ir nelaisvos šalies statusu. Krostabuliacijos rezultatai (žr. pirmąją porinę lentelę) parodo, kad geriausiai rezultatai sutampa su Freedom House pateiktu statusu pusiau laisvų šalių atveju; laisvų šalių atveju du atvejus (Jamaiką ir Lesoto) klasterizavimo algoritmas priskyrė pusiau laisvoms šalims, o vieną nelaisvą šalį (Kazakstaną) priskyrė pusiau laisvoms šalims.

Į analizę įtraukus visas 193 šalis, krostabuliacijos rezultatai (žr. antrąją porinę lentelę) parodo, kad geriausiai rezultatai sutampa su Freedom House pateiktu statusu nelaisvų šalių atveju; laisvų šalių atveju vieną atvejį klasterizavimo algoritmas priskyrė pusiau laisvoms šalims, o prasčiausiai sekėsi atvejus priskirti pusiau laisvoms šalims – 8 atvejai priskirti nelaisvoms, o 3 – laisvoms šalims. Toks pat rezultatas gaunamas, pakartojus klasterizavimą su nestandartizuotais kintamaisiais.

<i>Porinė lentelė (17-a šalių):</i>	<i>Porinė lentelė (193 šalys):</i>
-------------------------------------	------------------------------------

Average Linkage (Between Groups)

* Statusas: F-laisva (Free); PF-iš dalies
laisva (Partly Free); NF-nelaisva (Not Free) Crosstabulation

Count

		Statusas: F-laisva (Free); PF-iš dalies laisva (Partly Free); NF-nelaisva (Not Free)			Total
		F	NF	PF	
		Average Linkage (Between Groups)	1	2	
	2	6	0	0	6
	3	0	2	0	2
Total		8	3	6	17

Average Linkage (Between Groups)

* Statusas: F-laisva (Free); PF-iš dalies
laisva (Partly Free); NF-nelaisva (Not Free) Crosstabulation

Count

		Statusas: F-laisva (Free); PF-iš dalies laisva (Partly Free); NF-nelaisva (Not Free)			Total
		F	NF	PF	
		Average Linkage (Between Groups)	1	0	
	2	1	0	49	50
	3	88	0	3	91
Total		89	44	60	193

Nr.	Šalis	Statusas	CLU3_1
85	Jamaica	F	1
86	Japan	F	2
87	Jordan	PF	1
88	Kazakhstan	NF	1
89	Kenya	PF	1
90	Kiribati	F	2
91	Kuwait	PF	1
92	Kyrgyzstan	PF	1
93	Laos	NF	3
94	Latvia	F	2
95	Lebanon	PF	1
96	Lesotho	F	1
97	Liberia	PF	1
98	Libya	NF	3
99	Liechtenstein	F	2
100	Lithuania	F	2
101	Luxembourg	F	2

8.1.2. Hierarchinė klasterinė analizė su Stata

STATA leidžia transformuoti kintamuosius per Data>Create New Variable meniu. Kartodami klasterinės analizės SPSS pavyzdį, paliksime kintamuosius nestandartizuotus, nes matėme, kad šių duomenų atveju standartizavimas į [0;1] intervalą klasterizavimo rezultatui įtakos nedarė, be to, standartizuoti esamus kintamuosius į [0;1] intervalą su STATA reikia daryti rankiniu būdu.

Sprendimas su Stata SE 10.0

Parengiamieji veiksmai

Nėra

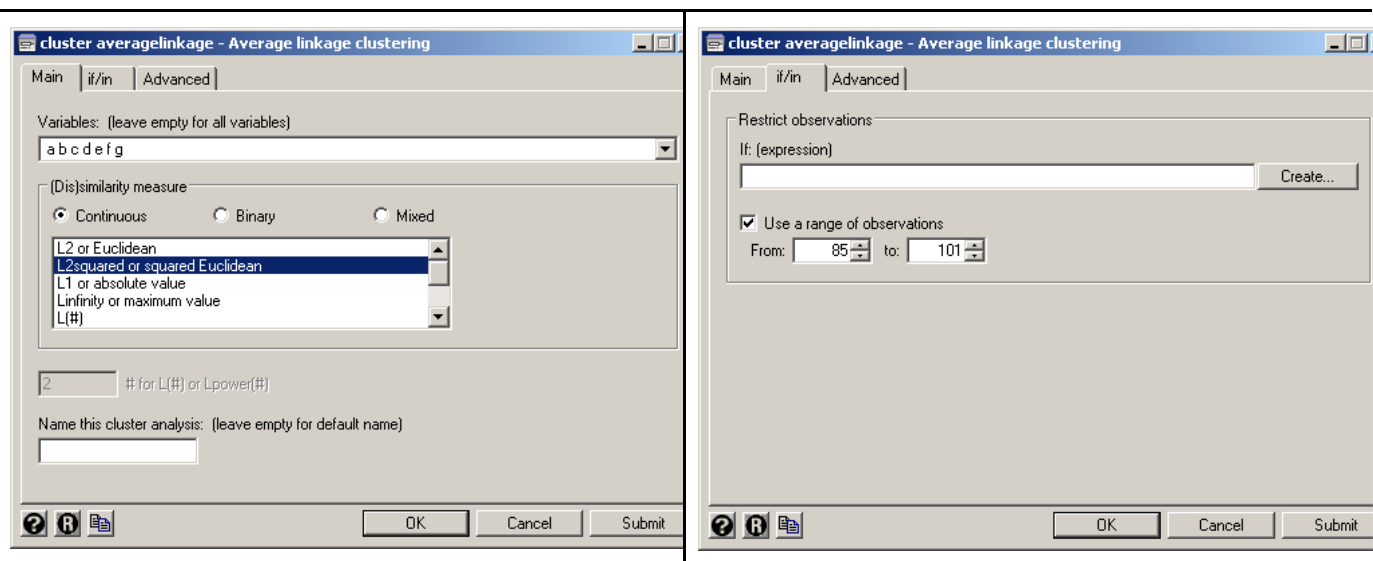
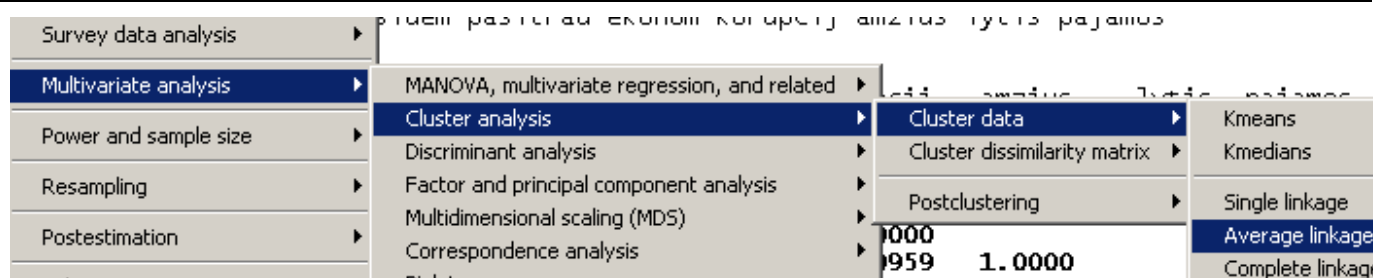
Sintaksės komandos

Klasterinės analizės veiksmai

Meniu punktai: Statistics → Multivariate analysis → Cluster analysis → Cluster data: pasirenkamas vidutinis klasterių jungimo metodas („Average linkage“); pasirenkami kintamieji, pagal kuriuos bus klasterizuoti atvejai (kintamieji **a - g**); nurodomas artumo matas („L2squared or squared Euclidean“ - Euklidinio atstumo kvadratas) → „in/if“ meniu nurodome, kad klasterizavimui pasirenkame ne visus duomenis, o tik atvejus nuo 85 iki 101 → OK → Submit (vietoje OK, kad nereiktų iš naujo kartoti visų

Sprendimas su Stata SE 10.0

pasirinkimų, jei norėsite pakartoti analizę pakeitę kurį nors nustatymą).

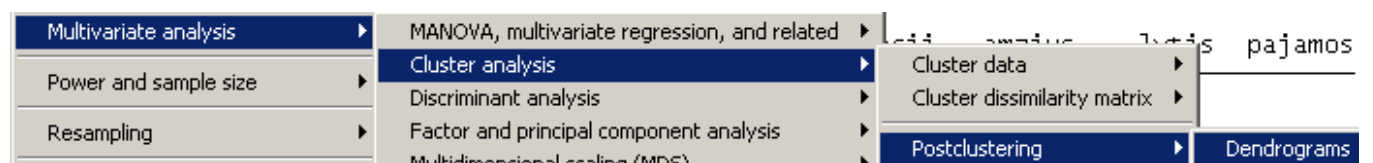


STATA sukuria tris naujus kintamuosius `_clus_1_id`, `_clus_1_ord`, `_clus_1_hgt`, kuriais remiantis bus daroma visa likusi klasterinė analizė. Šie kintamieji tėra tarpinis rezultatas, duomenų interpretavimui praktiškai nenaudojamas. Remiantis šiais kintamaisiais, vėliau bus braižoma dendrograma ir sukurtas klasterių narystės kintamasis.

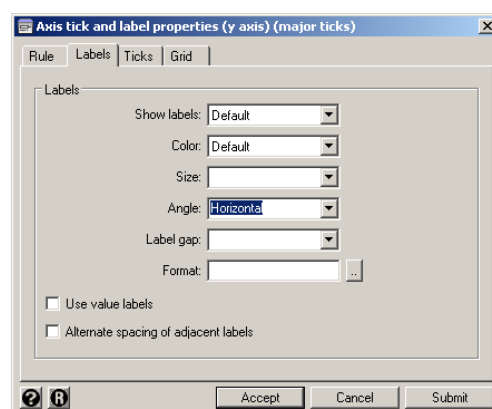
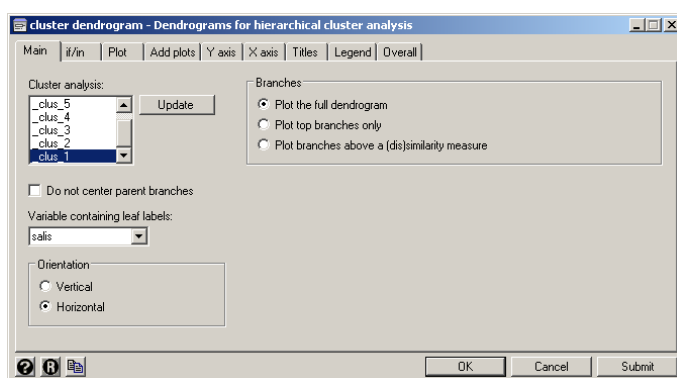
Sintaksės komandos

`cluster averagelinkage a b c d e f g in 85/101, measure(L2squared)`

Toliau, braižome dendrogramą: Statistics → Multivariate analysis → Cluster analysis → Postclustering → Dendrograms → Dendrogramos lange pasirenkame kintamąjį **clus_1** (arba reikiamą klasterinės analizės sprendinį, jeigu kelis kartus kartojo klasterizavimo procedūrą); pasirenkame, kad dendrogramoje atvejų pavadinimai būtų paimti iš kintamojo **salis** (pasirenkame šį kintamąjį iš „**variable containing leaf labels**“ meniu), taip pat, patogumo dėlei, pasirenkame horizontalią dendrogramos orientaciją ir horizontalų atvejų pavadinimų Y ašyje pateikimą → Submit.



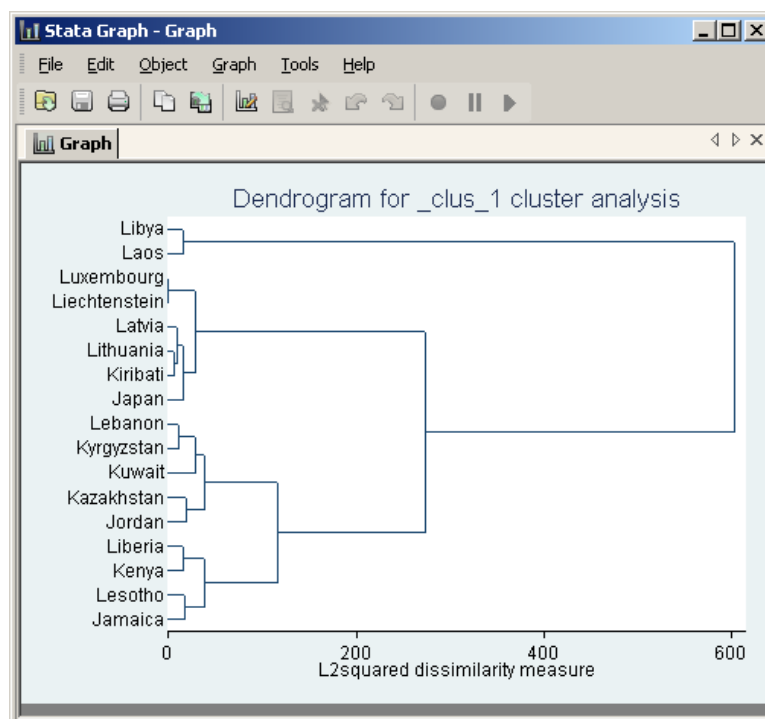
Sprendimas su Stata SE 10.0



Sintaksės komandos

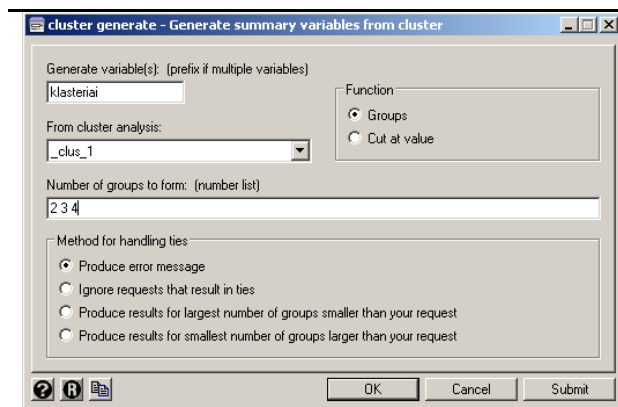
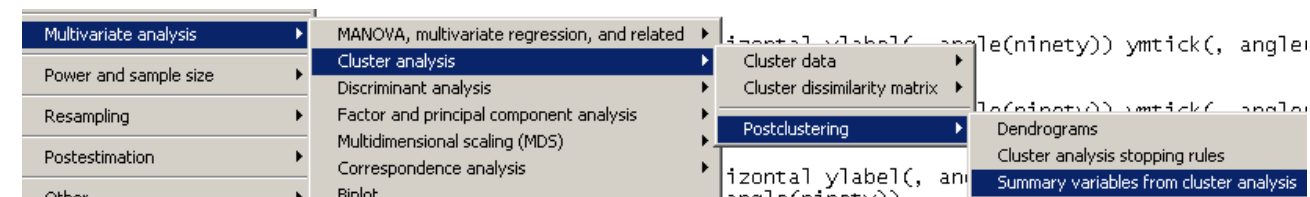
```
cluster dendrogram _clus_1, labels(salis) horizontal ylabel(, angle(horizontal))
```

Gauname tokius rezultatus:



Sukuriame vieną arba kelis klasterių narystės kintamuosius: Statistics → Multivariate analysis → Cluster analysis → Postclustering → Summary variables from cluster analysis: įrašome naujo kintamojo pavadinimą **klasteriai**; nurodome, kelių klasterių ir kiek sprendinių norėtume gauti: „Number of groups to form: (number list)“ įrašome „2 3 4“; iš „From cluster analysis“ meniu pasirenkame, kurios klasterinės analizės (jeigu jų buvo atliekama daugiau) rezultatais naudotis, kuriant kintamuosius: clus_1 → Submit. Po šios operacijos, STATA duomenų faile atsiranda trys nauji kintamieji „klasteriai2“, „klasteriai3“, „klasteriai4“.

Sprendimas su Stata SE 10.0



Sintaksės komandos

cluster generate klasteriai = groups(2 3 4), name(_clus_1) ties(error)

Naujai sukurta 3 klasterių kintamąjį **klasteriai3** galime palyginti su kintamuoju **statusas**, nubraižydami požymių priklausomumo (krostabuliacijų) lentelę : Statistics → Summaries, tables, and tests → Tables → Two-way tables with measures of association.

Sintaksės komandos

tabulate klasteriai3 statusas

Rezultatų langas				
klasteriai3	statusas			Total
	F	NF	PF	
1	2	1	6	9
2	6	0	0	6
3	0	2	0	2
Total	8	3	6	17

Visą analizę pakartojus su visais 193 atvejais, gauname:

visi	statusas			Total
	F	NF	PF	
1	0	44	8	52
2	1	0	48	49
3	88	0	4	92
Total	89	44	60	193

Atlikus lentelėje pateiktas klasterinės analizės procedūras ir nubraižę dendrogramą, pastebime, kad šalys sudaro demokratiškų ir nedemokratiškų šalių grupes – trijų klasterių lygyje Libija ir Laosas sudaro daugiausiai nutolusią nuo kitų klasterių grupę, o tai pat aiški demokratiškų šalių grupė (Liuksemburgas - Japonija) ir tarpinė, Libano-Jamaikos, grupė.

Požymių priklausomumo (krostabuliacijų) lentelė parodo, kad 17 šalių atveju, rezultatai yra identiški SPSS rezultatams, o klasterizuojant 193 atvejus, pusiau laisvos šalys suklasifikuojamos vienu atveju blogiau (48 atvejai STATA vietoje 49 atvejų SPSS'e iš 60 pusiau laisvų šalių).

9. FAKTORINĖ ANALIZĖ

Šiame skyriuje pateikiami pavyzdžiai, kaip su SPSS ir Stata atlikti faktirinę analizę. Bus nagrinėjamas tik tiriamosios faktorinės analizės atvejis

9.1. Tiriamoji faktorinė analizė

Faktorinė analizė leidžia gauti vieną ar kelis apibendrintus kintamuosius iš kelių ar keliolikos panašių savo prasme kintamųjų. Faktorinės analizės būdu galima sumažinti nagrinėjamų kintamųjų skaičių m iki n , taip kad $n < m$. Faktorinės analizės atlikimui reikia kintamųjų interkoreliacijų matricos. Iš to išplaukia, kad atliekant šią analizę, kintamieji turi būti normalieji, nes faktorinei analizei naudojama Pirsono koreliacijos koeficientų matrica. Jeigu kintamieji nėra normalaus skitinio, reikia sugeneruoti tokių sąlygų išpildymo nereikalaujančių koreliacijos koeficientų (pvz. Spirmeno) matricą ir ją naudoti faktorinei analizei.

Atliksime nenormalaus skitinio kintamųjų tiriamąją (angl. „explorative“) pasitikėjimo institucijomis faktorinę analizę naudodamiesi standartiniu faktorinės analizės meniu, trūkstamus duomenis (neatsakytus respondentų klausimų kodus) pakeisdami to kintamojo vidurkiais. Pasitikėjimas įvairiomis institucijomis Naujojo Baltijos Barometro (apklausa atlikta rinkos tyrimų bendrovės „Vilmorus“) tyrime įvertinamas klausiant respondentų, kiek pasitikima įvairiomis institucijomis: teismais, politinėmis partijomis, armija, Seimu, policija, Prezidentu, profsąjungomis, Bažnyčia.

9.1.1. Tiriamoji faktorinė analizė su SPSS

Sprendimas su SPSS 15.0	
Parengiamieji veiksmai	
Nėra	
Sintaksės komandos	
Analizės veiksmai	
Meniu punktai: Analyze → Data Reduction → Factor...	
Pagrindiniame lange pasirenkame kintamuosius teismai, partijos, armija, seimas, policija, prezid, profsaj, baznycia .	
„Descriptives“ lange pažymime „Initial solution“ ; „Coefficients“ , „Significance levels“ , „KMO and Bartlett’s test of sphericity“ , „Anti-Image“ .	
„Extraction“ lange pažymime „Scree plot“ ir „Loading plot(s)“ .	
„Rotation“ lange pasirenkame Varimax sūkį.	
„Options“ lange pažymime „Sorted by size“ .	

Sprendimas su SPSS 15.0

The image displays six SPSS 15.0 dialog boxes for Factor Analysis:

- Analyze > Data Reduction > Factor...** (Menu path)
- Factor Analysis** (Main dialog box):
 - Variables: kysiai, nvo, partpal, zmones, pazist, religing, vietove, mkorup, ekonom3, ekonom2, ekonteig, ekonom, korupcij, amzius
 - Variables: teismai, partijos, armija, seimas, policija, prezid, profsaj, baznycia
- Factor Analysis: Descriptives** (Sub-dialog box):
 - Statistics: ☐ Univariate descriptives, ☒ Initial solution
 - Correlation Matrix: ☒ Coefficients, ☐ Inverse, ☒ Significance levels, ☐ Reproduced, ☐ Determinant, ☒ Anti-image, ☒ KMO and Bartlett's test of sphericity
- Factor Analysis: Extraction** (Sub-dialog box):
 - Method: Principal components
 - Analyze: ☒ Correlation matrix, ☐ Covariance matrix
 - Display: ☒ Unrotated factor solution, ☒ Scree plot
 - Extract: ☒ Eigenvalues over: 1, ☐ Number of factors:
 - Maximum Iterations for Convergence: 25
- Factor Analysis: Rotation** (Sub-dialog box):
 - Method: ☐ None, ☐ Quartimax, ☒ Varimax, ☐ Equamax, ☐ Direct Oblimin, ☐ Promax
 - Delta: 0, Kappa: 4
 - Display: ☒ Rotated solution, ☒ Loading plot(s)
 - Maximum Iterations for Convergence: 25
- Factor Analysis: Options** (Sub-dialog box):
 - Missing Values: ☒ Exclude cases listwise, ☐ Exclude cases pairwise, ☐ Replace with mean
 - Coefficient Display Format: ☒ Sorted by size, ☐ Suppress absolute values less than: .10

Sintaksės komandos

FACTOR

```

/VARIABLES teismai partijos armija seimas policija prezid profsaj baznycia
/MISSING LISTWISE /ANALYSIS teismai partijos armija seimas policija prezid
profsaj baznycia
/PRINT INITIAL CORRELATION SIG KMO AIC EXTRACTION ROTATION
/FORMAT SORT
/PLOT EIGEN ROTATION
/CRITERIA MINEIGEN(1) ITERATE(25)
/EXTRACTION PC
/CRITERIA ITERATE(25)
/ROTATION VARIMAX
  
```

Sprendimas su SPSS 15.0

/METHOD=CORRELATION .

Gauname tokius rezultatus e:

Correlation Matrix

		Kiek jūs pasitikite teismais?	Kiek jūs pasitikite politinėmis partijomis?	Kiek jūs pasitikite kariuomene?	Kiek jūs pasitikite Seimu?	Kiek jūs pasitikite policija?	Kiek jūs pasitikite Prezidentu?	Kiek jūs pasitikite profsajungomis?	Kiek jūs pasitikite Bažnyčia?
Correlation	Kiek jūs pasitikite teismais?	1,000	,419	,219	,397	,377	,220	,273	,128
	Kiek jūs pasitikite politinėmis partijomis?	,419	1,000	,235	,583	,352	,306	,255	,185
	Kiek jūs pasitikite kariuomene?	,219	,235	1,000	,317	,295	,383	,203	,324
	Kiek jūs pasitikite Seimu?	,397	,583	,317	1,000	,443	,390	,259	,200
	Kiek jūs pasitikite policija?	,377	,352	,295	,443	1,000	,289	,240	,222
	Kiek jūs pasitikite Prezidentu?	,220	,306	,383	,390	,289	1,000	,225	,279
	Kiek jūs pasitikite profsajungomis?	,273	,255	,203	,259	,240	,225	1,000	,166
	Kiek jūs pasitikite Bažnyčia?	,128	,185	,324	,200	,222	,279	,166	1,000

Lentelėje „Koreliacijų matrica“ („Correlation Matrix“) peržiūrime, ar nėra neigiamų koreliacijų, kas rodytų, kad kai kuriuos kintamuosius reiktų perkoduoti atvirkščiai, o svarbiausia, peržiūrime, ar koreliacijos statistiškai reikšmingai skiriasi nuo nulio. Nagrinėjamu atveju, visos koreliacijos yra teigiamos ir statistiškai reikšmingai skiriasi nuo nulio.

KMO and Bartlett's Test

Kaiser-Meyer-Olkin Measure of Sampling Adequacy.		,827
Bartlett's Test of Sphericity	Approx. Chi-Square	1353,130
	df	28
	Sig.	,000

Lentelėje „KMO and Bartlett's Test“ nurodyto „Kaiser-Meyer-Olkin Measure of Sampling Adequacy“ kriterijaus vertė turi būti $> 0,5$ (pavyzdyje gavome 0,827). Tad gavome, kad 82,7% dispersijos kintamuosiuose gali paaiškinti du išskirti faktoriai.

„Bartlett's Test of Sphericity“ reikšmingumo lygmuo („Sig.“) turi būti $< 0,05$ (pavyzdyje gavome $< 0,0001$). Tikrinama nulinė hipotezė, kad kintamųjų koreliacijų matrica yra vienetinė, t.y., kad kintamieji yra nesusiję vienas su kitu. Nulinė hipotezė atmetama. Šis rezultatas visada turėtų sutapti su koreliacijų matricos (žr. aukščiau) analizės rezultatais, kurie akivaizdžiau parodo, ar kintamieji yra tarpusavyje reikšmingai susiję.

Sprendimas su SPSS 15.0

Anti-image Matrices

	Kiek jūs pasitikite teismais?	Kiek jūs pasitikite politinėmis partijomis?	Kiek jūs pasitikite kariuomene?	Kiek jūs pasitikite Seimu?	Kiek jūs pasitikite policija?	Kiek jūs pasitikite Prezidentu?	Kiek jūs pasitikite profsąjungomis?	Kiek jūs pasitikite Bažnyčia?
Anti-image Correlation								
Kiek jūs pasitikite teismais?	,849 ^a	-,213	-,046	-,112	-,196	,000	-,138	,023
Kiek jūs pasitikite politinėmis partijomis?	-,213	,789 ^a	,008	-,422	-,050	-,062	-,070	-,045
Kiek jūs pasitikite kariuomene?	-,046	,008	,826 ^a	-,095	-,103	-,230	-,056	-,211
Kiek jūs pasitikite Seimu?	-,112	-,422	-,095	,790 ^a	-,217	-,178	-,043	,004
Kiek jūs pasitikite policija?	-,196	-,050	-,103	-,217	,866 ^a	-,060	-,068	-,085
Kiek jūs pasitikite Prezidentu?	,000	-,062	-,230	-,178	-,060	,849 ^a	-,078	-,135
Kiek jūs pasitikite profsąjungomis?	-,138	-,070	-,056	-,043	-,068	-,078	,899 ^a	-,061
Kiek jūs pasitikite Bažnyčia?	,023	-,045	-,211	,004	-,085	-,135	-,061	,821 ^a

^a. Measures of Sampling Adequacy(MSA)

Lentelėje „Anti-image Matrices“ „Anti-image Correlations“ dalyje ištirštinėje esančios koreliacijos yra kintamojo stebėjimų tinkamumo matai (MSA- Measure of Sampling Adequacy) ir turi būti ne mažesnės negu 0,05. Mūsų atveju, visų kintamųjų stebėjimai yra tinkami faktorinei analizei.

Communalities

	Initial	Extraction
Kiek jūs pasitikite teismais?	1,000	,555
Kiek jūs pasitikite politinėmis partijomis?	1,000	,612
Kiek jūs pasitikite kariuomene?	1,000	,578
Kiek jūs pasitikite Seimu?	1,000	,634
Kiek jūs pasitikite policija?	1,000	,457
Kiek jūs pasitikite Prezidentu?	1,000	,501
Kiek jūs pasitikite profsąjungomis?	1,000	,252
Kiek jūs pasitikite Bažnyčia?	1,000	,596

Extraction Method: Principal Component Analysis.

Lentelėje „Kintamųjų bendrumai“ („Communalities“) pateikti bendrumai parodo, kuri kintamojo dispersijos dalis yra paaiškinama faktoriais (tais, kurių tikrinės reikšmės yra didesnės negu 1, jeigu analizės metu nenurodyta kitaip). Kintamieji, kurių paaiškinta faktoriais sklaidos dalis yra nedidelė (patartina ieškoti mažiausių reikšmių, kurios aiškiai išsiskiria iš kitų bendrumų), yra potencialūs kandidatai neįtraukti į faktorinę analizę, kadangi jie yra pernelyg „toli“ nuo bet kurio kito kintamojo ir todėl juos empiriškai mažiau pagrįsta grupuoti su kažkuria panašių kintamųjų grupe. Ši rezultatą taip pat galima patikrinti klasterinės analizės būdu.

Pavyzdyje gavome, kad išsiskiria paaiškinta pasitikėjimo profesinėmis sąjungomis sklaidos dalis (0,252 lyginant su kitu mažiausiu bendrumu 0,457 bei aukščiausiu bendrumu 0,634).

Sprendimas su SPSS 15.0

Total Variance Explained

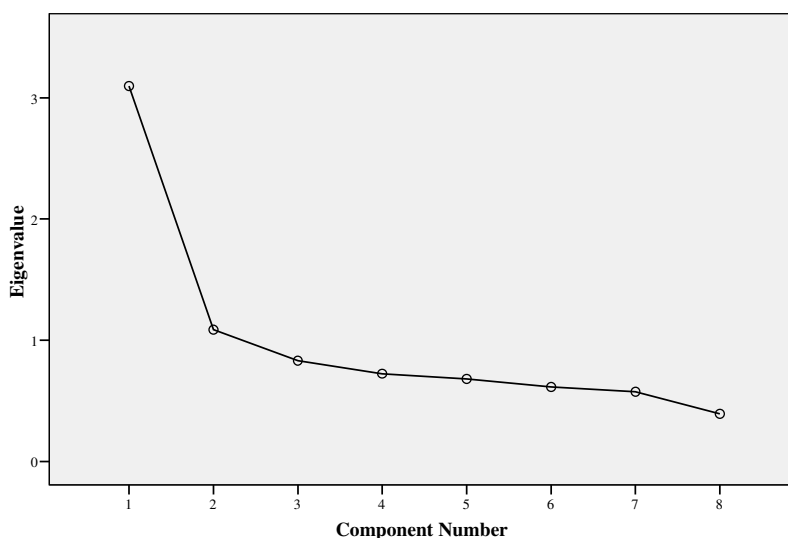
Component	Initial Eigenvalues			Extraction Sums of Squared Loadings			Rotation Sums of Squared Loadings		
	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %
1	3,097	38,718	38,718	3,097	38,718	38,718	2,461	30,765	30,765
2	1,086	13,578	52,297	1,086	13,578	52,297	1,723	21,532	52,297
3	,831	10,387	62,684						
4	,723	9,041	71,725						
5	,681	8,513	80,239						
6	,614	7,672	87,911						
7	,575	7,184	95,094						
8	,392	4,906	100,000						

Extraction Method: Principal Component Analysis.

Lentelėje „Total Variance Explained“ atkreipiame dėmesį į faktorių paaiškinamą visų kintamųjų sklaidos dalį. Kuo didesnė kumuliatyvi sklaidos dalis, kurią kintamuosiuose paaiškina išskirti faktoriai, tuo faktorinė analizė yra sėkmingesnė. Gavome, kad du faktoriai, kurių tikrinės reikšmės yra 2,981 (šis faktorius paaiškina 37,259 % visų kintamųjų dispersijos) ir 1,075 (šis faktorius paaiškina 13,433 % visų kintamųjų dispersijos), kartu šie du faktoriai paaiškina 50,692 % visų kintamųjų sklaidos. Atlikus pradinių faktorių pasukimą (jų struktūros optimizavimą, maksimizuojant faktorių svorių dispersiją), pirmo faktoriaus paaiškinta dispersijos dalis sumažėjo iki 29,184 %, o antrojo faktoriaus paaiškinama dalis padidėjo iki 21,508 %, tačiau bendra faktorių paaiškinama dispersijos dalis išliko ta pati (žr. lentelės dalį „Rotation Sums of Squared Loadings“).

Neretai didesnę dalį sklaidos galima gauti išmetus iš analizės mažai su faktoriais susijusius kintamuosius, žr. aukščiau apie lentelę „Kintamųjų bendrumai“.

Scree Plot



Faktorių tikrinių reikšmių grafike („Scree Plot“) pavaizduotos lygčių sistemos tikrinės reikšmės, leidžiančios nuspręsti, koks faktorių skaičius geriausiai aprašo duomenis. Vienas iš faktorių skaičiaus pasirinkimo tolesnei analizei variantų – toliau tirti visus faktorius, kurių tikrinės vertės yra >1 (šis variantas SPSS pasirenkamas pagal nutylėjimą), o kitas būdas – grafiškai nustatyti, kurioje vietoje tikrinės reikšmės daro paskutinį didelį šuolį žemyn ir toliau pradeda mažėti labai nedaug – tas lūžio taškas ir nurodo galimą faktorių skaičių (t.y., jeigu aiškiai atsiskiria didesnių ir mažesnių tikrinių reikšmių faktorių grupės, pasirenkama tik pirmoji faktorių grupė). Šiuo atveju toliau reiktų tirti arba du faktorius (jų tikrinės reikšmės >1) arba, remiantis grafiniu faktorių skaičiaus nustatymo būdu, užtektų nagrinėti tik pirmąjį faktorių. Kadangi mums įdomu išsiaiškinti, ar galime pasitikėjimą įvairiomis

Sprendimas su SPSS 15.0

institucijomis supaprastintai įvertinti vienu apibendrintu pasitikėjimo institucijomis indeksu ar skale, bet kuris faktorius, kurio tikrinė vertė didesnė negu 1 yra įdomus, nes parodo, kad egzistuoja viena ar daugiau nestebima latentinė dimensija, kuria galima aprašyti sąlyginai didelę visų kintamųjų klaidos dalį, t.y., aprašyti tiriamus duomenis. Kartais, atsižvelgus į faktorių sudėtį ir interpretaciją, gali reikėti užsakyti daugiau arba mažiau faktorių, negu remiantis tikrinės reikšmės kriterijumi.

Component Matrix^a

	Component	
	1	2
Kiek jūs pasitikite Seimu?	,764	-,225
Kiek jūs pasitikite politinėmis partijomis?	,706	-,337
Kiek jūs pasitikite policija?	,665	-,120
Kiek jūs pasitikite teismais?	,624	-,407
Kiek jūs pasitikite Prezidentu?	,623	,337
Kiek jūs pasitikite kariuomene?	,585	,485
Kiek jūs pasitikite profsąjungomis?	,497	-,071
Kiek jūs pasitikite Bažnyčia?	,456	,623

Extraction Method: Principal Component Analysis.

a. 2 components extracted.

Lentelė „Faktorių svorių matrica“ („Component Matrix“) parodo koreliacijas tarp kintamųjų ir faktorių. Ši lentelė dažniausiai traktuojama kaip tarpinis rezultatas, nes neretai tas pats kintamasis labai gerai susijęs (turi didelį svorį) keliuose ar net visuose faktoriuose, todėl juos sunkiau interpretuoti. Tam tikru būdu transformavus pradinis faktorius, galima gauti geriau interpretuojamus rezultatus (žr. toliau „Pasuktų faktorių svorių matrica“). Tačiau verta pastebėti, kad pirmajame nepasuktame faktoriuje didžiausius svorius turi pasitikėjimo Seimu, politinėmis partijomis, policija kintamieji, o antrajame didžiausi teigiami yra pasitikėjimo armija, prezidentu ir bažnyčia kintamųjų svoriai – panašūs rezultatai, tik daug aiškesne forma bus gauti pasuktus faktorius faktorių erdvėje.

Sprendimas su SPSS 15.0

Rotated Component Matrix^a

	Component	
	1	2
Kiek jūs pasitikite politinėmis partijomis?	,773	,118
Kiek jūs pasitikite Seimu?	,758	,244
Kiek jūs pasitikite teismais?	,745	,015
Kiek jūs pasitikite policija?	,617	,275
Kiek jūs pasitikite profsąjungomis?	,450	,221
Kiek jūs pasitikite Bažnyčia?	,027	,772
Kiek jūs pasitikite kariuomene?	,211	,730
Kiek jūs pasitikite Prezidentu?	,325	,629

Extraction Method: Principal Component Analysis.

Rotation Method: Varimax with Kaiser Normalization.

a. Rotation converged in 3 iterations.

Lentelė „Pasuktų faktorių svorių matrica“ („Rotated Component Matrix“) yra bene pagrindinis analizės rezultatas, jeigu visos kitos sąlygos (nėra nesusijusių kintamųjų ir kt., žr. aukščiau) yra išpildytos. Faktoriai buvo pasukti tokiu būdu, kad mažiausios kintamųjų koreliacijos su nepasuktais faktoriais sumažėjo, o didžiausios išaugo. Gauname, kad pasitikėjimas politinėmis partijomis, Seimu, teismais, policija ir profsąjungomis sudaro pirmą faktorių, o pasitikėjimas bažnyčia, armija, Prezidentu sudaro antrą faktorių.

Component Transformation Matrix

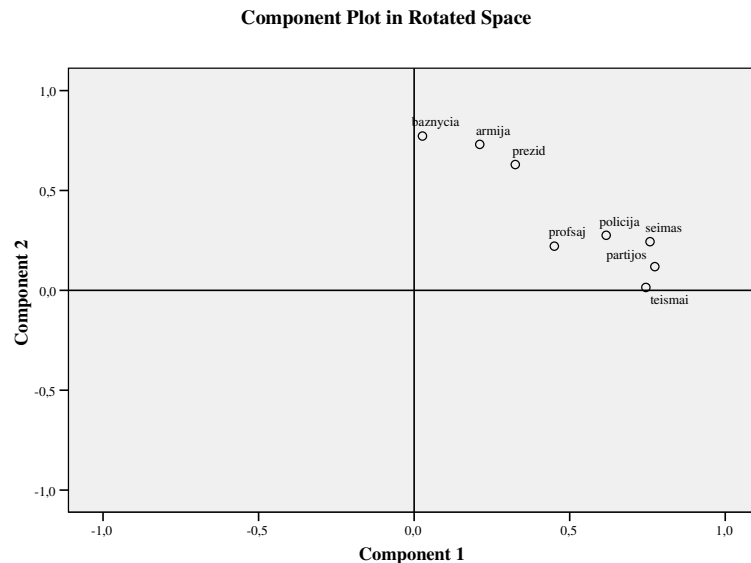
Component	1	2
1	,827	,562
2	-,562	,827

Extraction Method: Principal Component Analysis.

Rotation Method: Varimax with Kaiser Normalization.

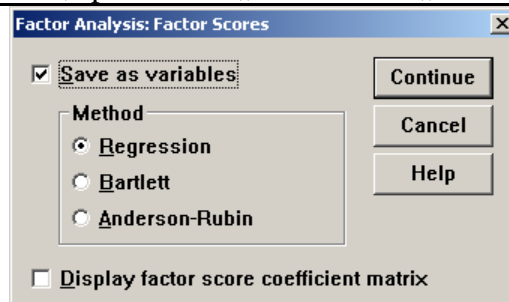
Faktorių transformacijos matrica parodo, kokius svorius pritaikius buvo pasukta pradinė, nepasuktų faktorių matrica ir gauta pasuktų faktorių svorių matrica. Ši informacija praktiškai nenaudojama rezultatų interpretacijoje, tačiau verta atkreipti dėmesį, kad ne diagonalėje esantys svoriai rodo atlikto pasukimo dydį (svoriai virš |0,5| parodo, kad buvo atliktas didelis pasukimas).

Sprendimas su SPSS 15.0



Pastebėtina, kad pasitikėjimo profesinėmis sąjungomis krūvis pirmajame faktoriuje nesiekia 0,5, be to, kaip matėme bendrumo lentelėje, gauti faktoriai paaiškina išskirtinai mažą šio kintamojo sklaidos dalį (šio kintamojo bendrumo įvertis yra 0,252. Visa tai rodo, kad šis kintamasis turi mažai bendro su šiuo faktoriumi ir galbūt jo reiktų atsisakyti analizėje. Papildoma šio kintamojo analizė rodo, kad net 225 atvejais buvo neatsakyta į šį klausimą, todėl greičiausiai šis klausimas buvo sunkiau suprantamas respondentams ir tai yra papildomas įrodymas, kad analizėje jo reiktų atsisakyti. Atsisakius šio kintamojo, dviejų faktorių paaiškinta sklaida kintamuosiuose pakyla iki 56,701 proc. Atliekant faktorinę analizę be **profsaj** kintamojo (pasitikėjimas profesinėmis sąjungomis), papildomai pasirenkame tokius nustatymus:

Meniu punktai: Analize → Data Reduction → Factor...: „Scores“ kortelėje „Save as variables“ skiltyje pažymime „Method“ – „Regression“, spaudžiame „Continue“ ir „OK“.



Sintaksės komandos

FACTOR

```
/VARIABLES teismai partijos armija seimas policija prezid baznycia  
/MISSING LISTWISE /ANALYSIS teismai partijos armija seimas policija prezid  
baznycia  
/PRINT INITIAL CORRELATION SIG KMO AIC EXTRACTION ROTATION  
/FORMAT SORT  
/PLOT EIGEN ROTATION  
/CRITERIA MINEIGEN(1) ITERATE(25)  
/EXTRACTION PC  
/CRITERIA ITERATE(25)  
/ROTATION VARIMAX  
/SAVE REG(ALL)
```

Sprendimas su SPSS 15.0

/METHOD=CORRELATION .

Tokiu būdu duomenų faile išsaugojame du faktorius kaip duomenų stulpelius tolesnei analizei. Pirmas iš jų atspindi standartizuotą pasitikėjimo demokratinės valstybės institucijomis laipsnį, o antras – standartizuotą pasitikėjimo tradicinio autoriteto institucijomis laipsnį. Faktoriai turi vidurkį 0 ir standartinį nuokrypį 1. Šiuose pavyzdžiuose naudojami **pasitdem** ir **pasitrad** kintamieji buvo gauti būtent tokios faktorinės analizės būdu, trūkstamas reikšmes pakeičiant kintamųjų vidurkiais.

Galima rezultatų interpretacija:

Antrasis faktorius yra aiškiai susijęs su tradiciniu hierarchiniu autoritetu, pirmasis daugiau su politinėmis demokratinės valstybės institucijomis⁶⁶. Taip pat verta pastebėti, kad pirmojo faktoriaus institucijomis pasitikima mažiausia, o antro daugiausia, todėl pirmasis faktorius apjungia institucijas, kuriomis nepasitikima, o antras – kuriomis pasitikima. Taigi, matome, kad respondentams tradicinės institucijos sudaro daugiau vertų pasitikėjimo institucijų grupę, o demokratinių institucijų grupė – mažiau vertų pasitikėjimo institucijų grupę. Kadangi pirmąją institucijų grupę pasitikima vidutiniškai mažiau negu antrąją, kita faktorių interpretacija galėtų būti skirtis tarp „tradicinio“ (moraliniu autoritetu paremto) ir „uždirbamo“ (institucijų veiksmingumu paremto) gyventojų pasitikėjimo.

Iš metodologinės pusės, išsiaiškinome, kad pasitikėjimas įvairiomis institucijomis yra nevienalytis empirinis konstruktas, bet apima bent dvi skirtingas institucijų grupes.

9.1.2. Tiriamoji faktorinė analizė su Stata

Sprendimas su Stata SE 10.0

Parengiamieji veiksmai

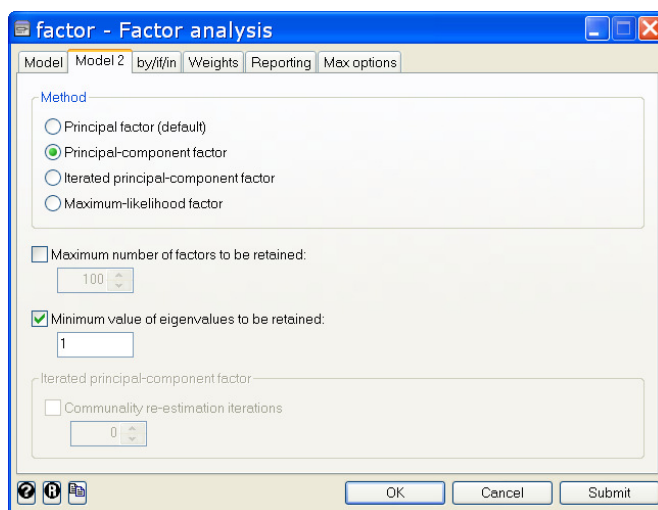
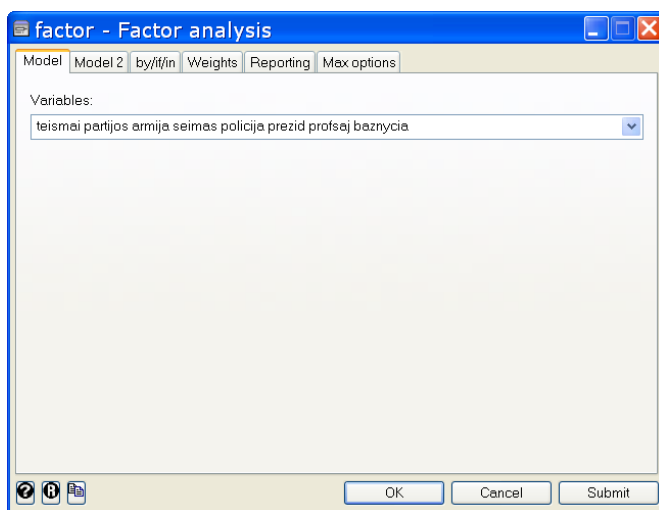
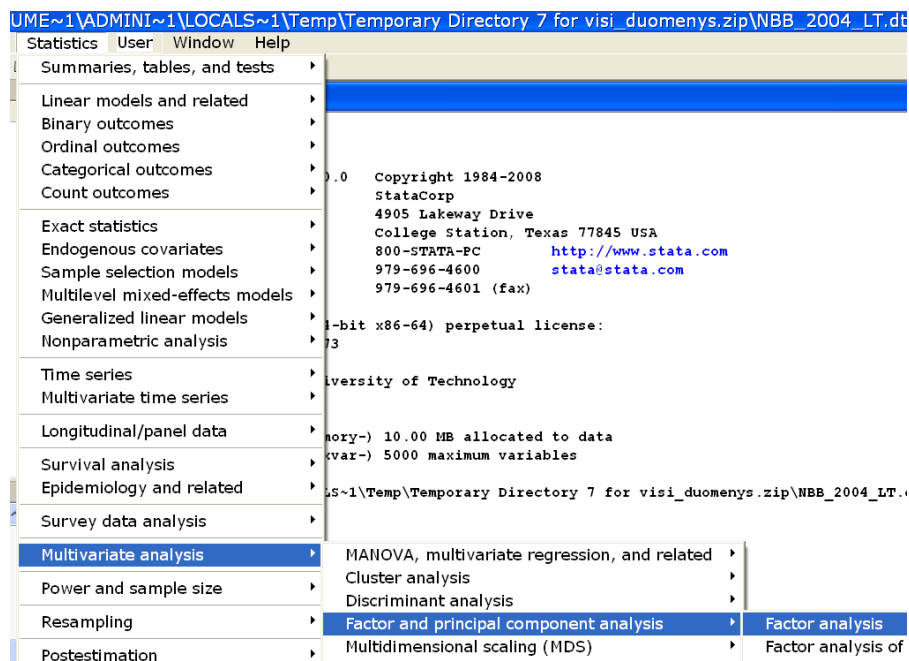
Nėra

Analizės veiksmai

Meniu punktai: Statistics → Multivariate analysis → Factor and principal component analysis → Factor analysis

⁶⁶ Tiesa, pasitikėjimas policija konceptualiai ne visiškai tinka. Kaip matėme bendrumų lentelėje, šio kintamojo bendrumas yra pats mažiausias, todėl jį galima būtų išmesti iš analizės. Tokia analizė būtų konceptualiai grynesnė.

Sprendimas su Stata SE 10.0



Sintaksės komandos

```
factor teismai partijos armija seimas policija prezid profsaj baznycia, pcf  
mineigen(1)
```

Gauname tokius rezultatus:

Sprendimas su Stata SE 10.0

```

Factor analysis/correlation
Method: principal-component factors
Rotation: (unrotated)
Number of obs   =    821
Retained factors =     2
Number of params =    15
  
```

Factor	Eigenvalue	Difference	Proportion	Cumulative
Factor1	3.09746	2.01120	0.3872	0.3872
Factor2	1.08627	0.25529	0.1358	0.5230
Factor3	0.83098	0.10766	0.1039	0.6268
Factor4	0.72332	0.04226	0.0904	0.7173
Factor5	0.68106	0.06731	0.0851	0.8024
Factor6	0.61376	0.03905	0.0767	0.8791
Factor7	0.57471	0.18226	0.0718	0.9509
Factor8	0.39245	.	0.0491	1.0000

LR test: independent vs. saturated: $\chi^2(28) = 1354.79$ Prob> $\chi^2 = 0.0000$

Factor loadings (pattern matrix) and unique variances

Variable	Factor1	Factor2	Uniqueness
teismai	0.6239	-0.4067	0.4454
partijos	0.7057	-0.3374	0.3882
armija	0.5850	0.4853	0.4222
seimas	0.7635		0.3665
policija	0.6652		0.5432
prezid	0.6226	0.3373	0.4986
profsaj	0.4967		0.7483
baznycia	0.4563	0.6228	0.4039

(blanks represent $\text{abs}(\text{loading}) < .3$)

Atliekame faktorių pasukimą: Statistics → Multivariate analysis → Factor and principal component analysis → Postestimation → Rotate loadings. Pasirenkame Varimax faktorių pasukimo metodą su Kaizerio normalizacija ir nurodome, kad būtų sukami 2 faktoriai.

Sprendimas su Stata SE 10.0

C:\DOCUME~1\ADMINI~1\LOCALS~1\Temp\Temporary Directory 7 for visi_duomenys.zip\NBB_2004_LT

Graphics Statistics User Window Help

Summaries, tables, and tests

Linear models and related

Binary outcomes

Ordinal outcomes

Categorical outcomes

Count outcomes

Exact statistics

Endogenous covariates

Sample selection models

Multilevel mixed-effects models

Generalized linear models

Nonparametric analysis

Time series

Multivariate time series

Longitudinal/panel data

Survival analysis

Epidemiology and related

Survey data analysis

Multivariate analysis

Power and sample size

Resampling

Postestimation

Other

Factors

Retained factors = 2

Number of params = 15

Difference	Proportion	Cumulative
2.01120	0.3872	0.3872
0.25529	0.1358	0.5230
0.10766	0.1039	0.6268
0.04226	0.0904	0.7173
0.06731	0.0851	0.8024
0.03905	0.0767	0.8791
0.18226	0.0718	0.9509
.	0.0491	1.0000

variated: chi2(28) = 1354.79 Prob>chi2 = 0.0000

and unique variances

Factor2	Uniqueness
1	0.4067
2	0.4454

MANOVA, multivariate regression, and related

Cluster analysis

Discriminant analysis

Factor and principal component analysis

Multidimensional scaling (MDS)

Correspondence analysis

Biplot

Procrustes transformation

Procrustes overlay

Factor analysis

Factor analysis

Principal component analysis

PCA of a correlation matrix

Rotate loadings

Postestimation

rotate - Orthogonal and oblique rotations after fac...

Main Reporting Opt options

Rotation method

☒ Orthogonal rotation

Varimax (default)

☐ Oblique rotation

Promax

Option

☐ Remove rotation results from the estimation command (i.e. clear)

☐ Apply the Kaiser normalization

☒ Specify the number of factors or components to be rotated

2 #

OK Cancel Submit

rotate - Orthogonal and oblique rotations after fac...

Main Reporting Opt options

☒ Display numbers as blanks when (|number| < #)

0.3 #

☒ Show rotatemat output

☐ Override display format for matrices

Create...

☐ Suppress display of rotated loadings

☐ Suppress display of rotation matrix

OK Cancel Submit

Sintaksės komandos

rotate, factors(2) detail blanks(0.3)

Gauname tokius rezultatus :

Sprendimas su Stata SE 10.0

Rotation of Factor loadings[8,2]

```

Criterion          varimax
Rotation class     orthogonal
Kaiser normalization off
  
```

Rotated factors (blanks represent abs(<.3)

	Factor1	Factor2
teismai	0.74444	
partijos	0.77228	
armiija		0.73196
seimas	0.75582	
policija	0.61504	
prezid	0.32010	0.63160
profsaj	0.44867	
baznycia		0.77182

Orthogonal rotation

	Factor1	Factor2
Factor1	0.82236	0.56897
Factor2	-0.56897	0.82236

```

Factor analysis/correlation          Number of obs   =    821
Method: principal-component factors   Retained factors =     2
Rotation: orthogonal varimax (Kaiser off) Number of params =    15
  
```

Factor	Variance	Difference	Proportion	Cumulative
Factor1	2.44638	0.70903	0.3058	0.3058
Factor2	1.73735	.	0.2172	0.5230

LR test: independent vs. saturated: $\chi^2(28) = 1354.79$ Prob> $\chi^2 = 0.0000$

Rotated factor loadings (pattern matrix) and unique variances

Variable	Factor1	Factor2	Uniqueness
teismai	0.7444		0.4454
partijos	0.7723		0.3882
armiija		0.7320	0.4222
seimas	0.7558		0.3665
policija	0.6150		0.5432
prezid	0.3201	0.6316	0.4986
profsaj	0.4487		0.7483
baznycia		0.7718	0.4039

(blanks represent abs(loading)<.3)

Factor rotation matrix

	Factor1	Factor2
Factor1	0.8224	0.5690
Factor2	-0.5690	0.8224

Nubrėžiame faktorių grafiką: Statistics → Multivariate analysis → Factor and principal component analysis → Postestimation → Loading plot

Sprendimas su Stata SE 10.0

\\DOCUMENT~1\ADMINI~1\LOCALS~1\Temp\Temporary Directory 7 for visi_duomenys.zip\NBB_2004_LT

Statistics User Window Help

Summaries, tables, and tests

Linear models and related

Binary outcomes

Ordinal outcomes

Categorical outcomes

Count outcomes

Exact statistics

Endogenous covariates

Sample selection models

Multilevel mixed-effects models

Generalized linear models

Nonparametric analysis

Time series

Multivariate time series

Longitudinal/panel data

Survival analysis

Epidemiology and related

Survey data analysis

Multivariate analysis

Power and sample size

Resampling

Postestimation

Other

	Difference	Proportion	Cumulative
	0.70903	0.3058	0.3058
	.	0.2172	0.5230

urated: $\chi^2(28) = 1354.79$ Prob> $\chi^2 = 0.0000$

matrix) and unique variances

Factor2	Uniqueness
	0.4454
	0.3882
7320	0.4222
	0.3665
	0.5432
6316	0.4986
	0.7483
7718	0.4039

g) < .3)

MANOVA, multivariate regression, and related

Cluster analysis

Discriminant analysis

Factor and principal component analysis

Multidimensional scaling (MDS)

Correspondence analysis

Biplot

Procrustes transfo

Procrustes overla

Cronbach's alpha

Factor analysis

Factor analysis

Principal compc

PCA of a correl

Rotate loadings

Postestimation

Loading plot

loadingplot - Loading plots

Main Y axis X axis Titles Overall

Plot all combinations of the following: 2 Factors

Abbreviate variable names to: 12 Characters

☐ Use unrotated results

☐ Graph as a matrix plot; default is a scatterplot

Marker properties

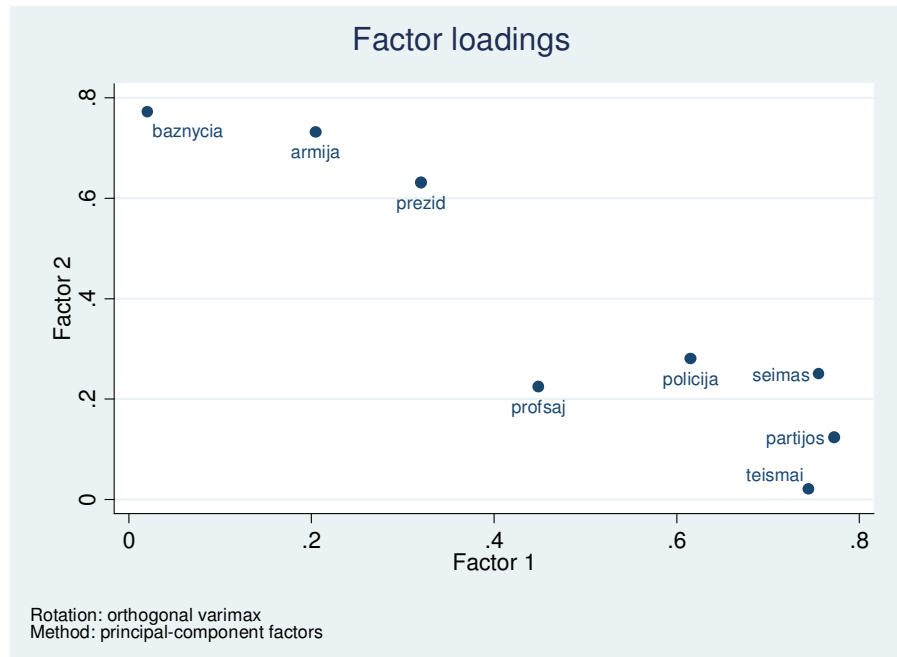
Label properties

OK Cancel Submit

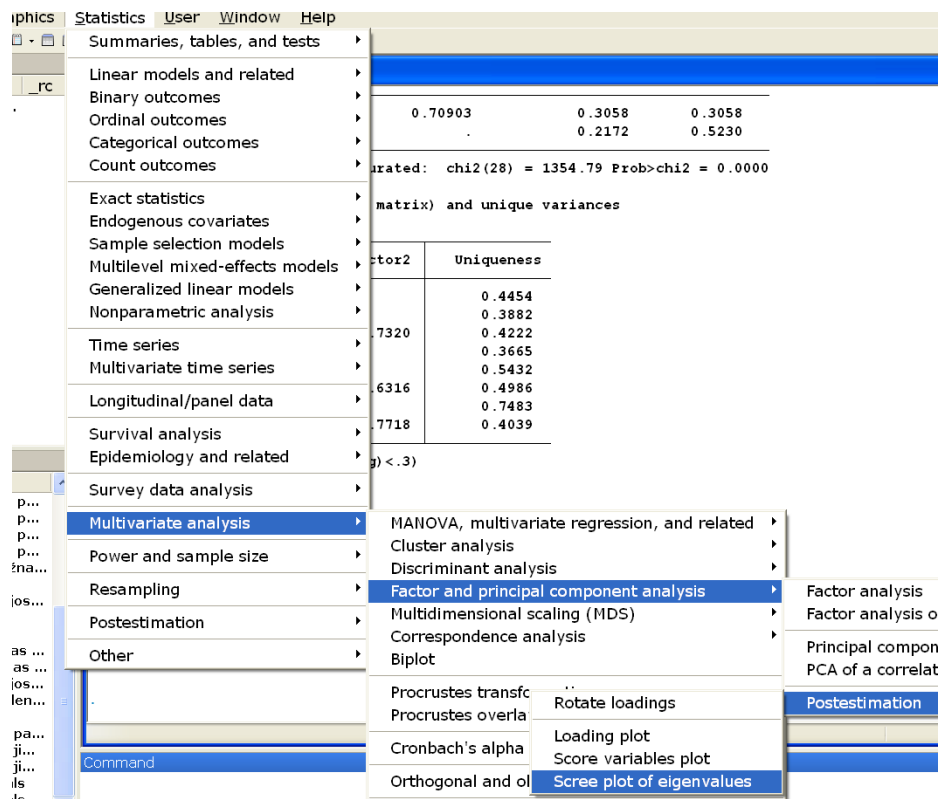
Sintaksės komandos

loadingplot

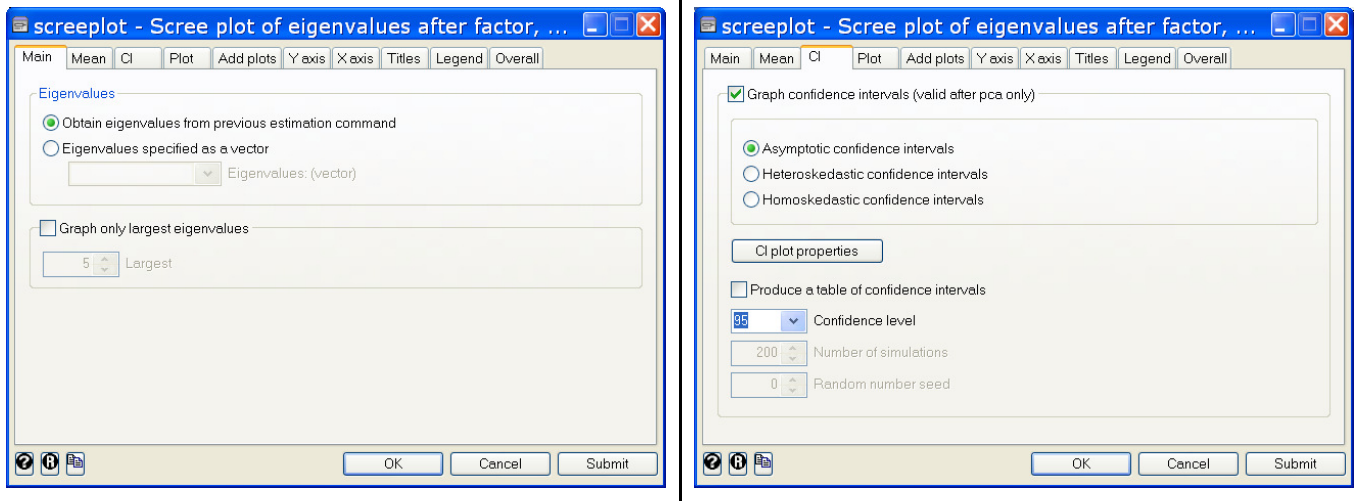
Sprendimas su Stata SE 10.0



Nubraižome faktorių tikrinių reikšmių grafiką: Statistics → Multivariate analysis → Factor and principal component analysis → Postestimation → Scree plot of eigenvalues

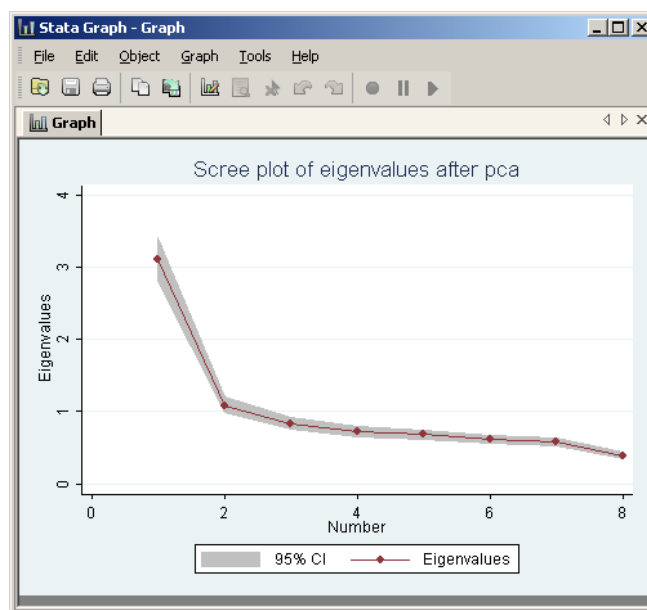


Sprendimas su Stata SE 10.0



Sintaksės komandos

screepplot, ci



Du faktorius tolimesnei analizei išsaugojame duomenų faile kaip kintamuosius šia komanda:

```
predict f1 f2
```

Galima rezultatų interpretacija nesiskiria nuo tos, kuri buvo pateikta aptariant SPSS faktorinės analizės atvejį, todėl čia nebesikartosime.

10. DAUGIAMAČIŲ SKALIŲ ANALIZĖ (MDS)

Daugiamačių skalių analizė (angl. *Multidimensional scaling*, toliau – MDS) yra dar viena **duomenų struktūrizavimo metodikų grupė** skirta atskleisti **latentinę arba „nematomą“ duomenų struktūrą** (Kruskal ir Wish 1978). Savo esme ji panaši į daugelį kitų matematinių duomenų struktūrizavimo metodikų grupių, dvi iš kurių – faktorinę ir klasterinę analizę⁶⁷ – jau aptarėme. **Duomenys** šiuo atveju yra identiški klasterinės analizės duomenims savo prasme, t. y. MDS struktūrizuoja artumo tarp objektų matricas. **Artumai** (išreikšti skaičiumi), kaip žinia, gali būti arba **objektyviai** išmatuoti (pavyzdžiui, atstumai tarp miestų, žodžio pasikartojimo dažnis pastraipoje ar pan.), arba atspindėti **subjektyvius** vertinimus (objektų panašumo vertinimas „žmogaus akimis“). Be to, skaitiniai artumai gali būti suskaičiuoti duomenims, kurie savo esme nėra artumai, o tai – gana dažna situacija socialinių mokslų duomenų atveju⁶⁸. **Objektai**, tarp kurių skaičiuojami atstumai, gali būti bet kokie reiškiniai ir jų charakteristikos, asmenys ir jų bruožai, fiziniai objektai ir pan., t. y. visa tai, tarp ko prasminga įvertinti panašumą. Pagrindinis **MDS uždavinys – atvaizduoti objektus kuo mažesnio rango mažamatėje erdvėje** taip, kad **artumai** tarp objektų kuo labiau **atitiktų atstumus** tarp jų toje erdvėje⁶⁹. Kaip tuos objektus atvaizduoti mažamatėje erdvėje **skaičiavimo algoritmai skiriasi**. Tie skirtumai priklauso nuo to, kokius duomenis tyrinėtojas turi, nes MDS pagalba galima analizuoti **simetrines ir asimetrines, vieną ir kelias, metrinių ir nemetrinių artumų matricas**.

Visų pirma, galima išskirti **klasikinį modelį**, kuriame sprendinys sudaromas matricų algebros pagalba, o duomenys yra **viena simetrinė metrinių artumų matrica**. Vėliau buvo sukurti **modernieji modeliai**⁷⁰, kurių pagalba galima analizuoti ne tik klasikinio modelio, bet ir kitokio pobūdžio artumų matricas. Svarbiausias iš moderniųjų modelių socialiniams mokslams yra **nemetrinių artumų matricų** (nepaisant jų kiekio ir simetriškumo) analizė. Būtent, šiuos du modelius ir aptarsime pateikdami MDS pavyzdžius šiame studijų pakete⁷¹. Prieš tai dar reikia priminti pačią **daugiamačių skalių analizės (MDS) logiką**:

Pirmiausiai, sudaroma objektų **skaitinių artumų matrica: objektyviai, subjektyviai ar apskaičiuojant pagal objektų savybes ar vertinimus**. Čia svarbu atkreipti dėmesį į tai, koks matas yra parinktas – metrinis ar nemetrinis. Pirmu atveju (ir esant vienai simetriškai matricai) galima taikyti klasikinį MDS modelį, o antru – nemetrinį modernųjį.

Pagal turimus duomenis (artumų matricą) **parenkamas duomenų struktūrizavimo – MDS – modelis: klasikinis metrinis, modernusis metrinis arba modernusis nemetrinis**. Jei turime vieną

⁶⁷ Dar paminėtina korespondencinė, latentinių klasių, latentinių profilių ir latentinių charakteristikų analizė.

⁶⁸ Čia egzistuoja begalė artumo matų, kurie gali būti skirstomi į matuojančius objektų panašumą (pavyzdžiui, Pirsono koreliacijos koeficientas) ir nepanašumą (pavyzdžiui, Euklidinis atstumas) bei tinkamus matuoti artumą tarp kintamųjų išmatuotų dvireikšmėse (pavyzdžiui, Žakardo panašumo matas), tvarkos (pavyzdžiui, Spirmeno koreliacijos koeficientas) ir kiekybinėse (pavyzdžiui, Minkovskio atstumas) skalėse. Pastaruoju atveju pirmi du matai (Žakardo ir Spirmeno) yra nemetriniai, o paskutinis (Minkovskio) – metrinis. Be to, artumo matas galima skirstyti į matuojančius objektų vertinimo profilių (profile) panašumą (pavyzdžiui, Pirsono koreliacijos koeficientas) ir objektų vertinimo balų (verčių, scores) panašumą (pavyzdžiui, Manhattan atstumas).

⁶⁹ Taigi paprastai gaunamos kelios (2-3) dimensijos, kurios „gerai“ paaiškina (atspindi) duomenų (artumų) sklaidą, t. y. šios dimensijos charakterizuoja duomenis pagal tam tikrus latentinius („nematomus“) požymius. Įvardijus tuos latentinius požymius, paaiškinama duomenų struktūra – atskleidžiami keli objektų panašumą (objektyviai ar subjektyviai išmatuota) charakterizuojantys požymiai (MDS atveju – dimensijos).

⁷⁰ Paprastai tariant, jų sukūrimo pagrindas yra artumo tarp objektų matų metrinis ar monotoninis transformavimas į skirtumus tarp objektų (*disparities*), kuriuos atitinkančių atstumų tarp objektų mažamatėje erdvėje po to ieškoma. Ši transformacija yra vienas iš iteracinio optimizavimo uždavinio, skirto surasti geriausią atitikimą tarp objektų artumo matų ir jų atstumų mažamatėje erdvėje, žingsnių.

⁷¹ Būtent šiuo du modelius – klasikinį metrinių ir modernųjį nemetrinių – kaip prasmingiausių socialinių mokslų atstovams rekomenduoja statistikai (žr., pavyzdžiui, Bartholomew ir kt. 2002, p. 61-62).

simetrinę metrinių artumų matricą geriausia naudoti klasikinį modelį, jei turime asimetrinę ar kelias metrinių artumų matricas, tada reikia naudoti metrinį modernųjį modelį, o jei artumai simetrinėje arba asimetrinėje matricoje (ar keliose) teatspindi ranginę tvarką tarp objektų – reikia naudoti nemetrinį modernųjį modelį.

Taikant moderniuosius modelius (pavyzdžiui, nemetrinį modernųjį) parenkama **pradinė objektų konfigūracija** erdvėje, nuo kurios pradedama iteracinė optimizavimo procedūra. Čia egzistuoja įvairių galimybių: pradinė konfigūracija sudaroma **pagal klasikinį modelį, simplekso ar atsitiktiniu būdu**. Šis žingsnis svarbus tuo, kad būtų galima išvengti iteracinės optimizavimo procedūros konvergavimo vietiniame minimume (o ne globaliame minimume). Patartina, sudaryti bent keletą skirtingų pradinių konfigūracijų ir po to palyginti jų charakteristikas (STRESS reikšmių didumą), tačiau net ir tai negarantuoja visiškos „sėkmės“.

Sudaromi keli skirtingą skaičių dimensių turintys sprendiniai, palyginamos jų **kiekybinės ir kokybinės**⁷² **tinkamumo charakteristikos** ir priimamas sprendimas, **kelių dimensių sprendinys** geriausiai atspindi originalius objektų artumus.

Kokybinės charakteristikos visiems modeliams yra **bendros**⁷³:

Sprendinio paprastumas ir naudingumas interpretuojant latentinę duomenų struktūrą;

Sprendinio erdvinio žemėlapių interpretavimo paprastumas;

Sprendinio erdvinio žemėlapių stabilumas, kuriam nedaro įtakos nedideli originalių duomenų skirtumai ar pakeitimai;

Liekamųjų paklaidų prasmingos struktūros nebuvimas.

Kiekybinės charakteristikos skiriasi priklausomai nuo modelio:

Klasikinio modelio atveju atsižvelgiama į sprendinio tinkamumo (*goodness-of-fit*) charakteristiką, atspindinčią tai, kiek originalių duomenų sklaidos paaiškina sprendinys su tam tikru dimensių skaičiumi. Optimalus sprendinys yra toks, kai pakankamai mažas dimensių kiekis (2-3) paaiškina pakankamai daug (daugiau nei 80 proc.) duomenų sklaidos. Priemonė, kaip geriausiai patikrinti klasikinio modelio sprendinio tinkamumą yra **tikrinių reikšmių ir dimensių skaičiaus grafikas** (*scree plot*) bei **paašikintos sklaidos didumų analizė**.

Moderniųjų modelių atveju taip pat pagrindinis kriterijus yra sprendinio tinkamumo (*badness-of-fit*) charakteristika (dažniausiai vadinama angliškos santrumpos STRESS vadinimu). Tačiau čia ji yra kitaip apskaičiuojama⁷⁴ nei klasikiniame modelyje. Be to, kadangi originalūs artumai tarp objektų yra transformuojami, tai būtina atsižvelgti į tų transformacijų kokybę: atitikimą tarp pradinių artumų ir transformuotų artumų (atstumų euklidinėje erdvėje). Taigi sprendinio tinkamumui įvertinti naudotinos tokios priemonės⁷⁵:

⁷² Čia, kaip ir kitų duomenų struktūrizavimo metodikų atvejais, reikia atkreipti dėmesį į tai, kad vien kiekybinių sprendinio tinkamumo charakteristikų priimant sprendimą, kiek dimensių palikti interpretavimo etapui, nepakanka, nes MDS pasitelkiama kaip tiriamoji metodika turimų duomenų apibūdinimui ir suvokimui. Taigi jei sprendinys yra geras savo kiekybinėmis tinkamumo charakteristikomis, bet menkai padeda aiškiai ir lengvai suprantamai struktūrizuoti turimus duomenis, tai jis nėra geras kokybiniu – interpretavimo – požiūriu.

⁷³ Reikia pabrėžti, kad kokybinės tinkamumo charakteristikos taikomos **tik tiems** sprendiniams, kurie atitinka minimalius kiekybinius tinkamumo kriterijus!

⁷⁴ Egzistuoja ne mažiau nei 7 tokios skirtingai apskaičiuojamos charakteristikos (populiariausia – Kraskelo STRESS I), kurių didumai dar yra ir iteracinio optimizavimo procedūros konvergavimo kriterijumi.

⁷⁵ Visais atvejais papildomai reikėtų atsižvelgti į objektų kiekį, artumų matricos formą ir kitus dalykus (žr. Kruskal ir Wish 1978, p. 52-53).

Tinkamumo charakteristikų (*badness-of-fit*) didumų analizė: $> 0,2$ – blogas sprendinys, $0,2-0,1$ – patenkinamas sprendinys, $< 0,1$ – geras sprendinys⁷⁶.

Tinkamumo charakteristikų ir dimensių skaičiaus brėžinys (*scree plot*)⁷⁷: optimalus sprendinys yra ties šio brėžinio „alkūne“.

Šepardo sklaidos diagramos (*Shepard diagram*): tarp atstumų ir transformuotų artumų (optimalaus sprendinio atveju taškai brėžinyje turėtų būti išsidėstę arti 45° kampu pasvirusios tiesės), tarp atstumų ir artumų (optimalaus sprendinio atveju taškai brėžinyje turėtų išsidėstyti nuolat (monotoniškai) didėjančia arba mažėjančia tvarka) bei tarp transformuotų artumų ir artumų (optimalaus sprendinio atveju taškai brėžinyje turėtų būti išsidėstę taip, kad nebūtų (arba būtų labai mažai) horizontalių pakopų, t. y. taškų vienodame lygyje). Beje, paskutinės dvi diagramos yra perteklinės metrinų modernių modelių atveju, nes transformuoti artumai yra proporcingi pradiniais artumams.

Sprendinio grafinio vaizdavimo – erdvinio žemėlapio (*spatial map*) – interpretavimas:

pagal taškų artumą erdviniam žemėlapyje;

„nubraižant ir įvardijant“ prasmingas dimensijas;

statistinių technikų, pavyzdžiui, regresinės analizės, pagalba.

Pažymėtina tai, kad sprendinio dimensijos nėra fiksuotos, t. y. erdvinį žemėlapi galima **pasukti, apversti ar pakeisti jo koordinačių ašių susikirtimo taško vietą nekeičiant atstumų tarp objektų**. Taigi sprendinio interpretacija yra gana „laisvas“ tyrinėtojo veiksmas.

Pateikdami daugiamačių skalių analizės (MDS) pavyzdžius naudosime kelis duomenų rinkinius. Klasikiniam modeliui iliustruoti naudosime atstumų tarp Lietuvos miestų matricą. Pateikdami nemetrinio moderniojo modelio pavyzdžius naudosime Lietuvos gyventojų geokultūrinių nuostatų tyrimo atlikto 2006 m. spalio mėn. duomenis (SPSS laikmena – „*geokult_2006.sav*“, Stata laikmena – „*geokult_2006.dta*“). Šio tyrimo ir jame tyrinėjamų reiškinių bei kintamųjų aprašas yra pateiktas pirmojoje studijų paketo dalyje.

10.1. Klasikinio MDS modelio pavyzdys su Stata

Kaip minėta pirmiausia aptarsime atvejį, kai turime vieną simetrišką metrinų artumų tarp objektų matricą, ir norime atvaizduoti objektus mažamateje erdvėje. Šiuo atveju svarbu atkreipti dėmesį į tai, kad duomenys tikrai turi būti metriniai, kitaip atvaizduoti atstumai neturės prasmės. Taigi mūsų **tikslas** – atskleisti metrinų artumų tarp nagrinėjamų objektų latentinę struktūrą, išskiriant dimensijas, kurių pagrindu suformuoti tie artumai tarp objektų. Kad iliustruotume, kaip sprendžiamas toks uždavinys, pasitelksime atstumų tarp Lietuvos miestų duomenis (žr. 10.1 lentelę) ir nubraižysime dviejų dimensių „Lietuvos žemėlapi“.

⁷⁶ Tiesa, labai mažos tinkamumo charakteristikos reikšmės gali signalizuoti ne apie gerą tinkamumą, o apie degeneravusį sprendinį.

⁷⁷ Jei tinkamumo charakteristikos ne nuolat nuosekliai mažėja, o pasitaiko „šuoliukų aukštyn“, tai gali reikšti, kad sprendinys konverguoja link lokalaus minimumo, o ne globalaus.

10.1 lentelė. Atstumai tarp Lietuvos miestų.

Miestas	Alytus	Biržai	Druskininkai	Joniškis	Jurbarkas	Kaunas	Kėdainiai	Klaipėda	Marijampolė	Mažeikiai	Panevėžys	Rokiškis	Skuodas	Šiauliai	Šilutė	Švenčionys	Varėna	Vilkaviškis	Vilnius	Visaginas
Alytus	0	243	57	248	144	66	120	281	61	294	171	238	307	211	242	186	49	84	101	254
Biržai	243	0	296	94	208	178	125	261	234	183	66	64	236	101	265	174	287	250	202	161
Druskininkai	57	296	0	300	171	119	172	320	88	343	224	288	349	264	269	213	58	111	128	281
Joniškis	248	94	300	0	161	182	134	196	238	98	97	158	151	39	200	243	297	234	231	250
Jurbarkas	144	208	171	161	0	86	114	149	83	172	158	246	178	118	98	261	193	72	186	291
Kaunas	66	178	119	182	86	0	54	215	56	228	106	171	241	146	184	175	115	73	101	205
Kėdainiai	120	125	172	134	114	54	0	208	110	183	59	147	234	97	193	162	168	127	130	193
Klaipėda	281	261	320	196	149	215	208	0	233	120	239	325	75	159	53	370	329	222	310	401
Marijampolė	61	234	88	238	83	56	110	233	0	255	162	232	262	201	181	222	110	23	137	267
Mažeikiai	294	183	343	98	172	228	183	120	255	0	165	247	55	80	153	327	342	244	299	335
Panevėžys	171	66	224	97	158	106	59	239	162	165	0	88	219	79	224	162	220	178	136	170
Rokiškis	238	64	288	158	246	171	147	325	232	247	88	0	300	165	312	130	243	249	159	97
Skuodas	307	236	349	151	178	241	234	75	262	55	219	300	0	134	124	381	355	251	336	388
Šiauliai	211	101	264	39	118	146	97	159	201	80	79	165	134	0	157	241	260	190	213	249
Šilutė	242	265	269	200	98	184	193	53	181	153	224	312	124	157	0	355	291	170	284	385
Švenčionys	186	174	213	243	261	175	162	370	222	327	162	130	381	241	355	0	169	245	85	68
Varėna	49	287	58	297	193	115	168	329	110	342	220	243	355	260	291	169	0	133	84	237
Vilkaviškis	84	250	111	234	72	73	127	222	23	244	178	249	251	190	170	245	133	0	161	284
Vilnius	101	202	128	231	186	101	130	310	137	299	136	159	336	213	284	85	84	161	0	152
Visaginas	254	161	281	250	291	205	193	401	267	335	170	97	388	249	385	68	237	284	152	0

Šaltinis: Lietuvos automobilių kelių direkcija prie Susisiekimo ministerijos (<http://www.lra.lt/lt.php/atstumai/526>).

Pastaba: artumai čia yra atstumai nuo vieno miesto iki kito kilometrais. Tokia matrica dar vadinama pilnąja simetrine nepanašumų matrica, nes diagonalėje (papildintuose langeliuose) yra mažiausią nepanašumą atspindintis skaičius 0, o virš diagonalės įtraukti tie patys atstumai, kaip ir po diagonale.

Akivaizdu, kad šie atstumai yra iš tikrųjų metriniai, taigi artumams tarp miestų atvaizduoti mažamataje erdvėje galime naudoti **klasikinį MDS modelį**. Pavyzdys, kaip atlikti klasikinę daugiamačių skalių analizę su Stata⁷⁸, yra pateiktas 10.2 lentelėje.

10.2 lentelė. Klasikinė metrinė vienos simetrinės artumų matricos daugiamačių skalių analizė (MDS). Sprendimas su Stata SE 10.0

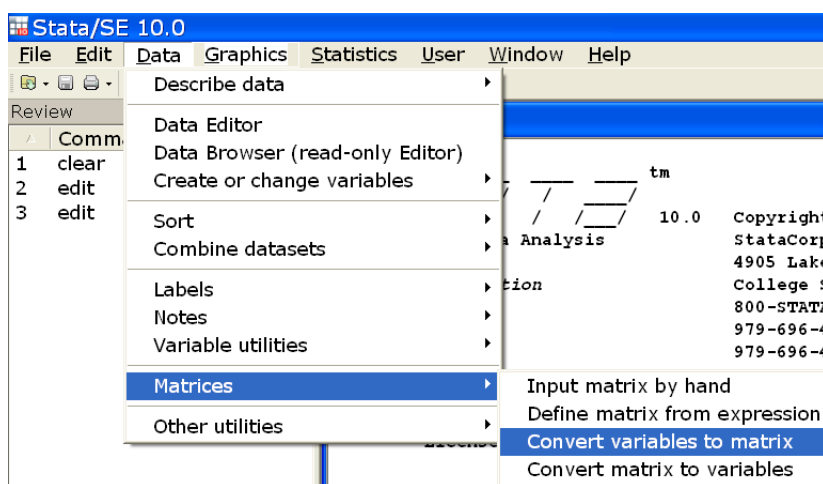
Parengiamieji veiksmai

Nukopijuojame 10.1. lentelę į tuščią Stata duomenų redagavimo langą (Copy → Paste)

Data Editor																
miestas[1] = Alytus																
	miestas	alytus	birai	druskininkai	jonikis	jurbarkas	kaunas	kdainiai	klaipda	marijampol	maeikiai	paneuvys	rokikis	skuodas	iauliai	
1	Alytus	0	243	57	248	144	66	120	281	61	294	171	238	307	211	
2	Birpai	243	0	296	94	208	178	125	261	234	183	66	64	236	101	
3	Druskininkai	57	296	0	300	171	119	172	320	88	343	224	288	349	264	
4	Joniškis	248	94	300	0	161	182	134	196	238	98	97	158	151	99	
5	Jurbarkas	144	208	171	161	0	86	114	149	83	172	158	246	178	118	
6	Kaunas	66	178	119	182	86	0	54	215	56	228	106	171	241	146	
7	Kėdainiai	120	125	172	134	114	54	0	208	110	183	59	147	234	97	
8	Klaipėda	281	261	320	196	149	215	208	0	233	120	239	325	75	159	
9	Marijampolė	61	234	88	238	83	56	110	233	0	255	162	232	262	201	
10	Mažeikiai	294	183	343	98	172	228	183	120	255	0	165	247	55	80	
11	Panevėžys	171	66	224	97	158	106	59	239	162	165	0	88	219	79	
12	Rokiškis	238	64	288	158	246	171	147	325	232	247	88	0	300	165	
13	Skuodas	307	236	349	151	178	241	234	75	262	55	219	300	0	134	
14	Šiauliai	211	101	264	39	118	146	97	159	201	80	79	165	134	0	
15	Šilutė	242	265	269	200	98	184	193	53	181	153	224	312	124	157	
16	Švenčionys	186	174	213	243	261	175	162	370	222	327	162	130	381	241	
17	Varena	49	287	58	297	193	115	168	329	110	342	220	243	355	260	
18	Vilkaviškis	84	250	111	234	72	73	127	222	23	244	178	249	251	190	
19	Vilnius	101	202	128	231	186	101	130	310	137	299	136	159	336	213	
20	Visaginas	254	161	281	250	291	205	193	401	267	335	170	97	388	249	

Tada reikia paversti mūsų duomenų matricą tokia, kad su ja būtų galima atlikti mus dominančią analizę. Tam reikia nurodyti Stata paketui, kad turimi duomenys yra matrica (konvertuoti kintamuosius į Stata matricą). Su šia konvertuota matrica ir bus atliekama MDS analizė.

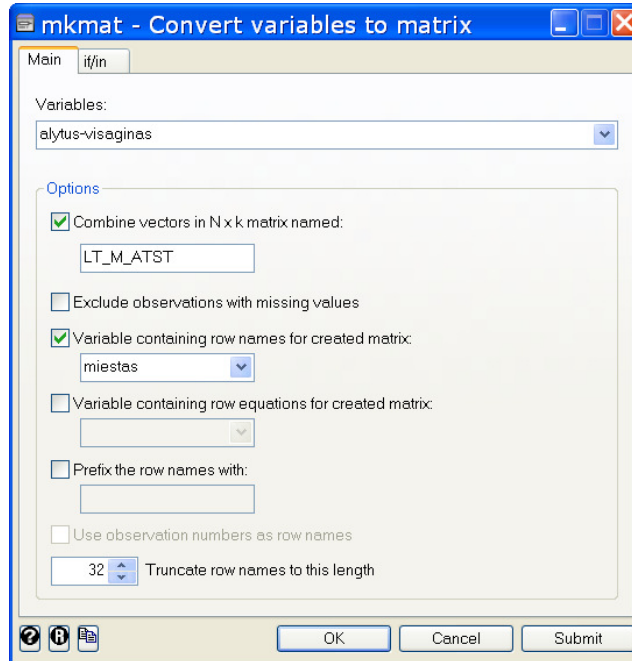
Meniu punktai: Data → Matrices → Convert variables to matrix



Atsidariusiame lange „*mkmat - Convert variables to matrix*“ nurodome, kurie kintamieji turi būti paversti Stata matrica, parinkdami **alytus-visaginas** ties „*Variabales*“. Po to „*Options*“ skiltyje nurodome matricos pavadinimą, ties „*Combine vectors in N x k matrix named*“ įvesdami **LT_M_ATST**, o ties „*Variable containing row names for created matrix*“ nurodome, kad Stata matricos eilučių pavadinimai būtų kuriami pagal kintamojo **miestas** reikšmes, ir spaudžiame „**OK**“.

⁷⁸ SPSS pakete atlikti tokią analizę galimybės nėra.

Sprendimas su Stata SE 10.0

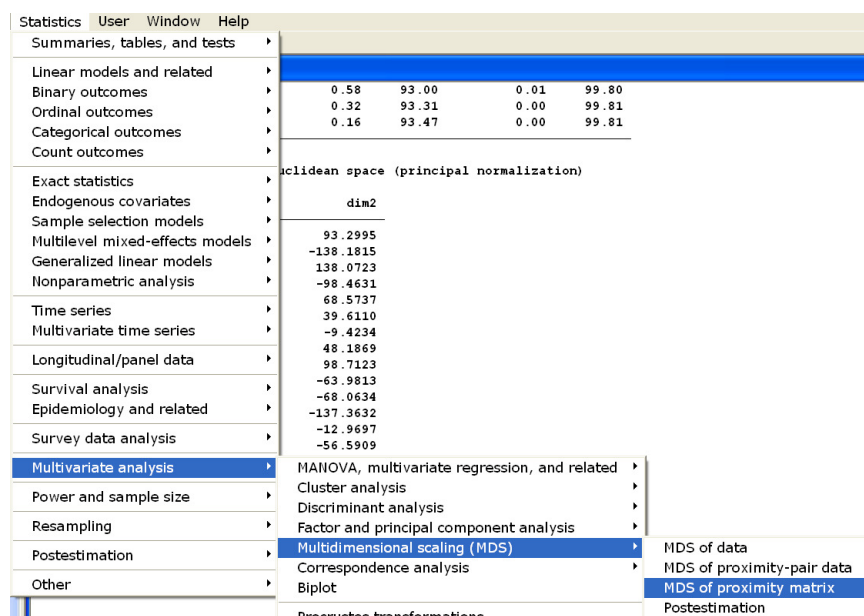


Sintaksės komandos

```
mkmat alytus-visaginas, matrix(LT_M_ATST) rownames(miestas)
```

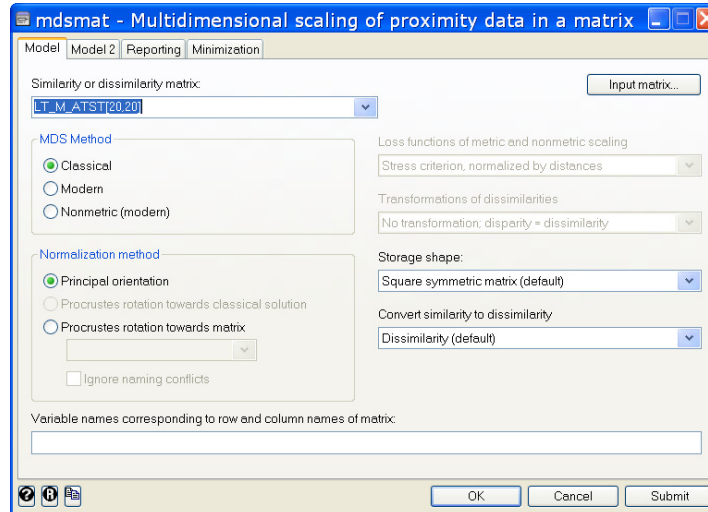
Analizės veiksmi

Meniu punktai: Multivariate analysis → Multidimensional scaling (MDS) → MDS of proximity matrix

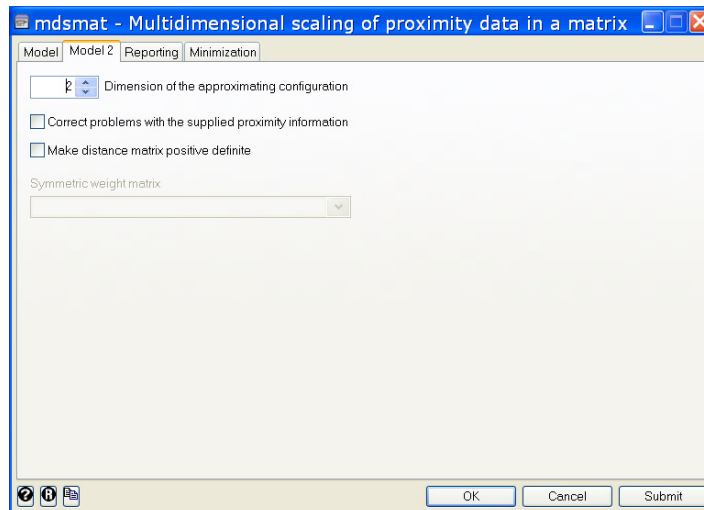


Atsidariusiame lange „*mdsmat – Multidimensional scaling of proximity data in a matrix*“ ties „*Similarity or dissimilarity matrix*“ parenkame sudarytą Stata matricą LT_M_ATST[20, 20].

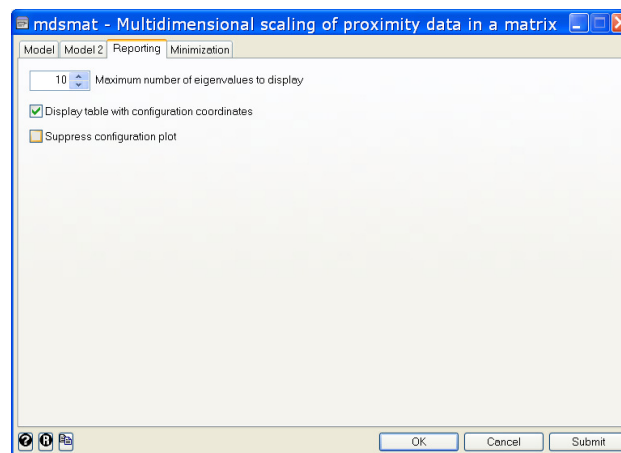
Sprendimas su Stata SE 10.0



Kortelėje „*Model 2*“ nurodome, kad galutinis sprendinys turėtų būti dviejų dimensijų, ties „*Dimension of the approximating configuration*“ nurodydami 2.



Kortelėje „*Reporting*“ nurodome, kad būtų pateiktos objektų koordinatės pažymėdami „*Display table with configuration coordinates*“ ir spaudžiame „*OK*“.



Sintaksės komandos

Sprendimas su Stata SE 10.0

```
mdsmat LT_M_ATST, method(classical) config dimensijon(2)
```

Gauname tokius rezultatus **rezultatų išvesties lange**:

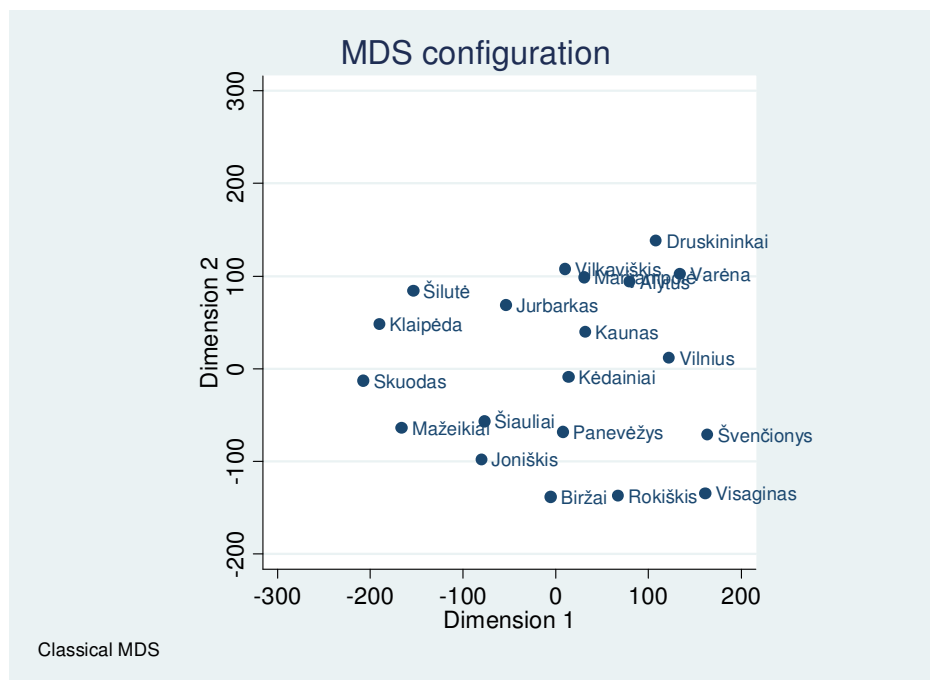
Classical metric multidimensional scaling
dissimilarity matrix: LT_M_ATST

Eigenvalues > 0	=	11	Number of obs	=	20
Retained dimensions	=	2	Mardia fit measure 1	=	0.8620
			Mardia fit measure 2	=	0.9958

Dimension	Eigenvalue	abs(eigenvalue)		(eigenvalue)^2	
		Percent	Cumul.	Percent	Cumul.
1	256537.77	53.30	53.30	72.09	72.09
2	158394.96	32.91	86.20	27.48	99.58
3	8885.6373	1.85	88.05	0.09	99.66
4	7329.2314	1.52	89.57	0.06	99.72
5	5870.3594	1.22	90.79	0.04	99.76
6	4577.1771	0.95	91.74	0.02	99.78
7	3243.0477	0.67	92.42	0.01	99.80
8	2793.8683	0.58	93.00	0.01	99.80
9	1520.3402	0.32	93.31	0.00	99.81
10	766.49288	0.16	93.47	0.00	99.81

Configuration in 2-dimensional Euclidean space (principal normalization)

Category	dim1	dim2
Alytus	79.8222	93.2995
Biržai	-5.3695	-138.1815
Druskininkai	108.5879	138.0723
Joniskis	-79.9019	-98.4631
Jurbarkas	-53.4778	68.5737
Kaunas	32.0204	39.6110
Kėdainiai	13.7006	-9.4234
Klaipėda	-190.0884	48.1869
Marijampolė	30.7848	98.7123
Mažeikiai	-166.0247	-63.9813
Panevėžys	7.6670	-68.0634
Rokiškis	67.5052	-137.3632
Skuodas	-207.4134	-12.9697
Šiauliai	-76.8211	-56.5909
Šilutė	-153.5302	84.0855
Švenčionys	164.0039	-71.4894
Varėna	134.0884	101.8556
Vilkaviškis	10.2191	107.5038
Vilnius	122.6172	11.5308
Visaginas	161.6101	-134.9056

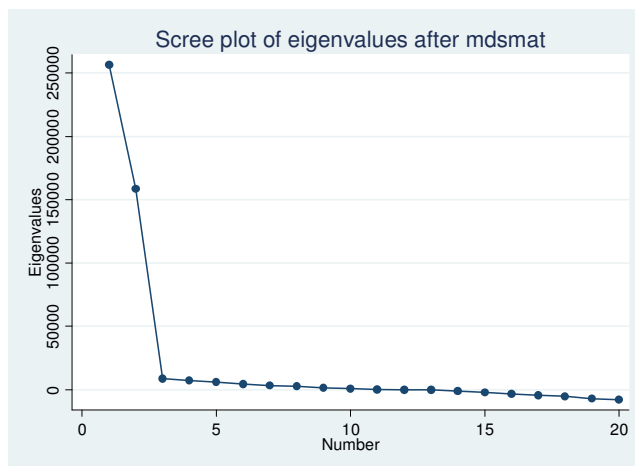


Išvados ir interpretacija

Pirmiausia žiūrime į sprendinių tinkamumo charakteristikų reikšmes (*Mardia fit measure 1 ir 2*). Abi jos viršija 80 proc., todėl galime sakyti, kad dviejų dimensijų sprendinys gerai atspindi duomenų struktūrą. Tą patvirtina ir tikrinių verčių ir dimensijų skaičiaus grafikas (*scree plot*).

Meniu punktai: Graphics → Multivariate analysis graphs → Scree plot → OK;

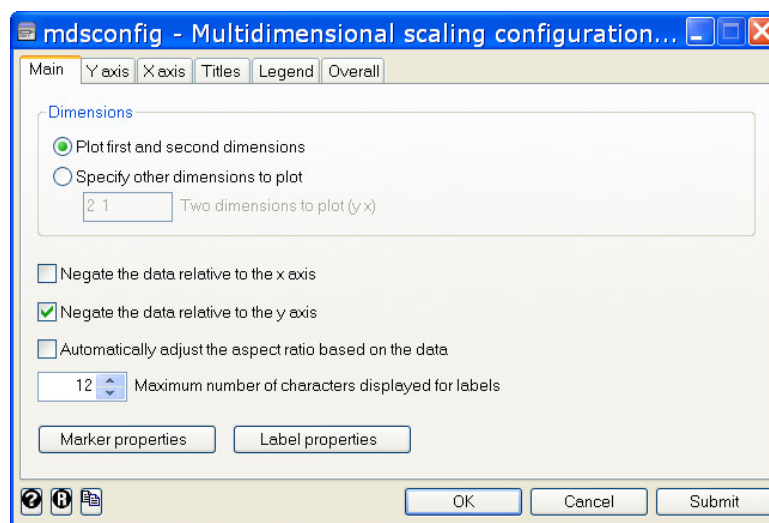
Sintaksė: screeplot



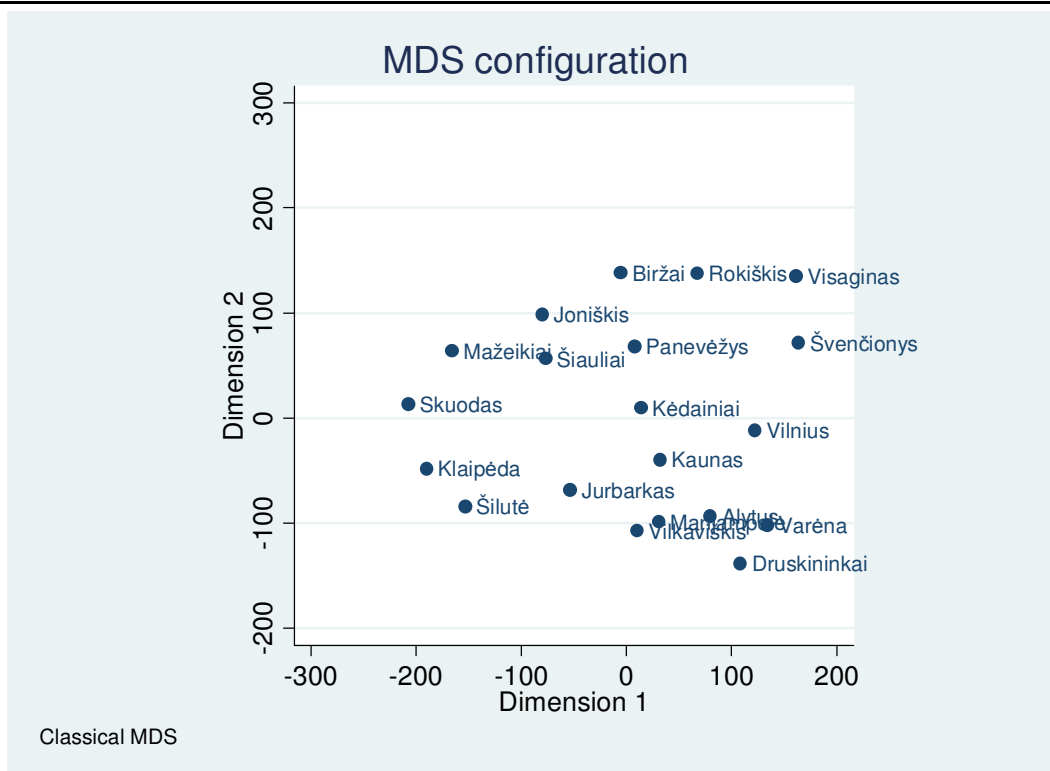
Tada žiūrime į dviejų dimensijų erdvinį žemėlapi (*MDS configuration*) ir bandome interpretuoti dimensijas. Aišku, šiuo atveju žinome, kad dimensijos turėtų atspindėti pasaulio šalis „Šiaurė-Pietūs“ ir „Rytai-Vakarai“. Tačiau pažūrėję į erdvinį brėžinį matome, kad „Šiaurė-Pietūs“ dimensija yra apversta aukštyn kojomis (Rokiškis yra pietuose, o Druskininkai šiaurėje, nors turėtų būti atvirkščiai). Kadangi dimensijų orientacija erdviniam žemėlapyje nėra fiksuota, tai galime apversti Y ašį ir turėtume gauti miestų išdėstymą, kuris mažai skirsis nuo tikrojo Lietuvos žemėlapi.

Meniu punktai: Graphics → Multivariate analysis graphs → After multidimensional scaling → Aproximating configuration plot → atsidariusiame lange pažymime „Negate the data relative to the y axis“ → OK;

Sintaksė: mdsconfig, ynegate



Gautas rezultatas jau beveik visai panašus į Lietuvos žemėlapi. Dar šiek tiek pasukus (30-45°) žemėlapi pagal laikrodžio rodyklę reprezentacija būtų beveik ideali. Pažymėtina, kad atstumai tarp miestų yra „tikri“, t. y. klasikinis modelis daugiau ar mažiau išlaiko tikrąsias artumų vertes tarp objektų ir erdvinėje reprezentacijoje.



Atlikę klasikinę daugiamačių skalių analizę su atstumų tarp Lietuvos miestais duomenimis pamatėme, kad pagal informaciją apie metrinčius artumus tarp objektų MDS pagalba iš tikrųjų galima nubraižyti tikrąjį Lietuvos žemėlapi. Žinoma, turėti metrinčių artumų matricą socialinių mokslų atstovui yra gana retas reiškinys. Paprastai jie susiduria su artumais, kurie nurodo ranginę objektų tvarką, todėl klasikinis modelis netinka ir dažniausiai tenka naudoti modernųjį nemetrinį modelį. Apie jį kaip tik ir kalbėsime sekančiame poskyryje.

10.2. Moderniojo nemetrinio MDS modelio pavyzdys su SPSS ir Stata

Klasikinio MDS modelio **išplėtimas nemetrinių, asimetrinių ir daugelio artumų matricų analizės kryptimi** įvyko septintajame ir aštuntajame praėjusio amžiaus dešimtmetyje. Taigi dabar šių metodikų pagalba galima nagrinėti įvairiausių duomenis ir spręsti įvairiausių uždavinius. Žinoma, kad jų visų pristatyti šiame studijų pakete neįmanoma. Todėl pateikdami pavyzdžius apie moderniųjį MDS modelių taikymą pasirinkome **socialiniams ir humanitariniams mokslams bene artimiausią atvejį – vienos nemetrinės simetrinės objektų artumo matricos analizę**. Nuo klasikinio modelio analizuotos ji skiriasi tik vienu požymiu – **artumų matricoje objektų panašumo ar nepanašumo įvertinimai atspindi tik jų ranginę tvarką** (pavyzdžiui, konservatorių partija panašesnė į krikščionių demokratų partiją nei į socialdemokratų), o ne konkrečią metriką (pavyzdžiui, atstumą kilometrais).

Taigi moderniąją nemetrinę MDS taikančio tyrinėtojo **tikslas** – atskleisti artumų (ranginės tvarkos prasme) tarp nagrinėjamų objektų latentinę struktūrą, išskiriant dimensijas, kurių pagrindu suformuoti tie artumai tarp objektų. Moderniosios nemetrinės MDS procedūra beveik niekuo nesiskiria nuo klasikinės, ir kadangi bendroji logika buvo pateikta pristatant pačią metodų grupę, tai iškart galime pereiti prie pavyzdžių pateikimo. Kad iliustruotume, kaip taikoma modernioji nemetrinė MDS, panagrinėsime 2006 m. Lietuvos gyventojų požiūrius į įvairius skirtingų šalių santvarkų aspektus. Respondentų buvo klausta, kaip jie vertina 14 buvusios TSRS, JAV, Vakarų Europos šalių, dabartinės Rusijos, Baltarusijos ir Lietuvos santvarkos aspektų (atsakymų variantai: nuo 1 – labai blogai iki 5 – labai gerai). Šių vertinimų pagrindu ir tiriamosios faktorinės analizės

pagalba⁷⁹ buvo nustatyta, kad žmonių vertinimuose išsiskiria du požūriai (visų santvarkų atvejais) – politinės gerovės ir socialinės bei kultūrinės gerovės vertinimas. Taigi atitinkamai kiekvienos santvarkos atskiri vertinimai buvo sugrupuoti į dvi „didžiąsias“ skales.

Turėdami šiuos duomenis pabandysime išsiaiškinti, ar egzistuoja kokia nors šalių vertinimo pagal tuos du aspektus latentinė struktūra, t. y. kurios šalys ir kurie jų santvarkų aspektai vertinami panašiai. Kadangi mūsų turimi duomenys nėra artumo matrica, ją reikės sudaryti panaudojant tinkamą artumo matą. Kokį matą benaudotume, jis vis tiek objektų artumą atspindės tik ranginės tvarkos prasme (pavyzdžiui, JAV politinė gerovė gali būti vertinama panašiau su Vakarų Europos nei su Baltarusijos, tačiau įvertinti to panašumo kokiu nors objektyviu matu – neįmanoma). Taigi šiems artumams atvaizduoti mažamateje erdvėje naudosime modernųjį nemetrinį MDS modelį. Pavyzdys, kaip atlikti moderniąją nemetrinę daugiamačių skalių analizę su SPSS (ir ALSCAL, ir PROXSCAL procedūromis⁸⁰) ir Stata, yra pateiktas 10.3-10.5 lentelėse.

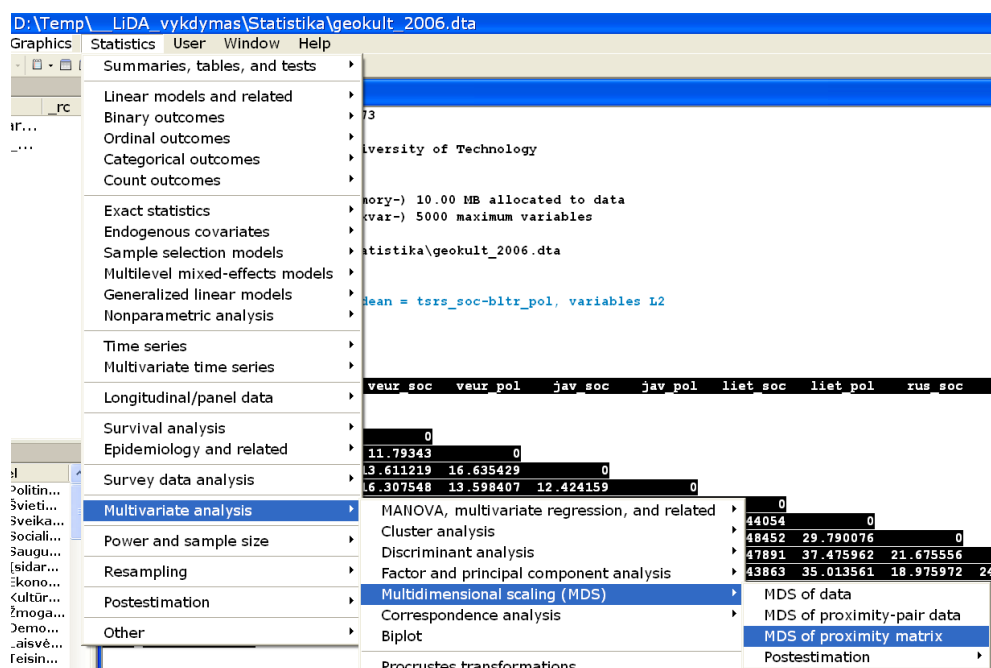
10.3 lentelė. Modernioji nemetrinė vienos simetrinės artumų matricos daugiamačių skalių analizė (MDS) su Stata.

Sprendimas su Stata SE 10.0											
Parengiamieji veiksmai											
Kadangi turimi duomenys nėra artumų matrica, tai visų pirma reikia sudaryti Stata artumų matricą . Kaip artumo matą pasirinkime tradicinį Euklidinį atstumą . Beje, sudaryti Stata artumų matricą galima tik sintaksės komandų pagalba. Su šia sudaryta Stata artumų matrica ir bus atliekama MDS analizė.											
Sintaksės komandos											
matrix dissimilarity GNT_Euclidean = tsrs_soc-bltr_pol, variables L2											
Taigi bus sudaryta „ GNT_Euclidean “ artumų matrica, kurioje artumai suskaičiuoti tarp kintamųjų (variables), o kaip artumo matas naudojamas Euklidinis atstumas (L2). Pasižiūrėti, kaip atrodo artumų matrica galima įvedus sintaksės komandą:											
matrix list GNT_Euclidean.											
Toliau pateiktas šios komandos rezultatas – simetrinė artumų matrica. Su šia artumų matrica bus atliekama tolesnė MDS analizė.											
<pre> symmetric GNT_Euclidean[12,12] tsrs_soc tsrs_pol veur_soc veur_pol jav_soc jav_pol liet_soc liet_pol rus_soc rus_pol bltr_soc tsrs_soc 0 tsrs_pol 40.699573 0 veur_soc 33.378324 55.526475 0 veur_pol 33.821209 55.44149 11.79343 0 jav_soc 34.650317 55.702002 13.611219 16.635429 0 jav_pol 35.679527 56.471664 16.307548 13.598407 12.424159 0 liet_soc 39.086698 40.26738 37.988452 37.85645 38.698291 39.128233 0 liet_pol 40.930412 42.707779 38.252444 37.356808 39.003657 38.480428 17.144054 0 rus_soc 34.228058 31.386068 40.924656 41.150708 41.814504 42.62177 26.648452 29.790076 0 rus_pol 46.189425 27.572772 54.221618 54.191604 54.637571 55.380381 35.47891 37.475962 21.675556 0 bltr_soc 40.233412 32.484347 48.469733 48.589728 49.433375 50.204908 31.43863 35.013561 18.975972 24.731435 0 bltr_pol 55.012986 30.904495 64.383249 64.449412 64.74924 65.559837 43.223935 45.978884 32.906919 21.376493 26.528088 bltr_pol bltr_pol 0 </pre>											

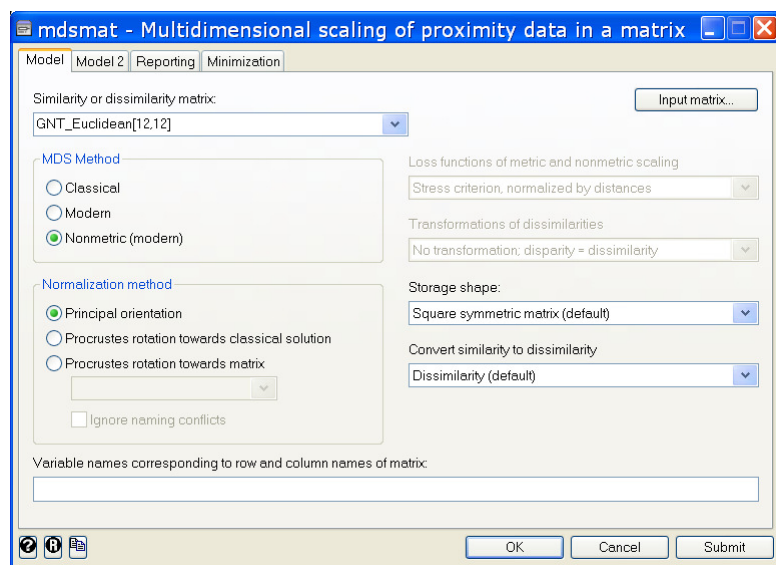
⁷⁹ Ji čia nepateikiama, o sudarytosios skalės jau yra šiame skyriuje naudojamose duomenų laikmenose (atitinkamai SPSS – „*geokult_2006.sav*“, Stata – „*geokult_2006.dta*“).

⁸⁰ Įvairūs autoriai pateikia argumentų, kad ALSCAL algoritmas linkęs konverguoti ties lokaliais minimumais, todėl SPSS pakete buvo įgyvendintas už jį „geresnis“ PROXSCAL algoritmas. Tačiau kadangi ALSCAL nebuvo visiškai atsisakyta, tai aptarsime jį dėl dviejų priežasčių. Jis buvo vienas populiariausių MDS algoritmų pastarąjį dvidešimtmetį, todėl žinoti, kaip jis atliekamas, vis dar turėtų būti pravartu. Be to, daugelio sprendinių sudarymas ir palyginimas yra geriausias vaistas nuo lokalių minimumų moderniojoje MDS, taigi turėsime progą palyginti dviejų skirtingų algoritmų sprendinius.

Meniu punktai: Multivariate analysis → Multidimensional scaling (MDS) → MDS of proximity matrix



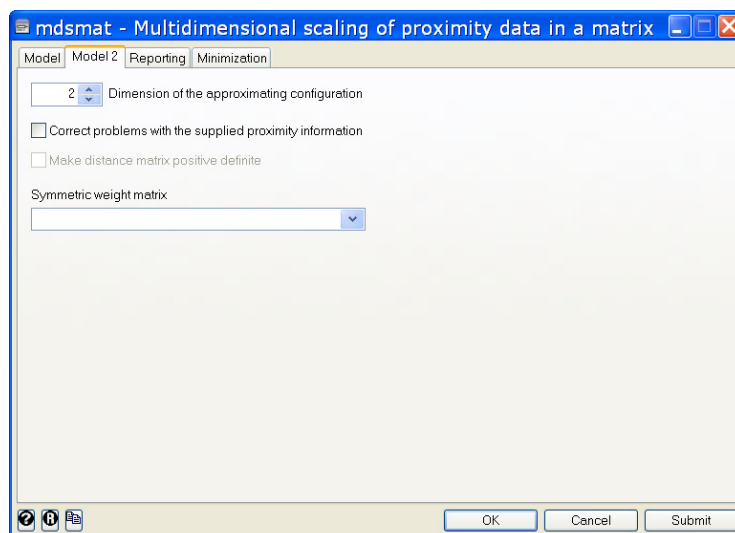
Atsidariusiame lange „*mdsmat – Multidimensional scaling of proximity data in a matrix*“ ties „*Similarity or dissimilarity matrix*“ parenkame sudarytą Stata artumų matricą GNT_Euclidean[12,12], ir nurodome, kad būtų skaičiuojamas nemetrinis modelis, „*MDS Method*“ skiltyje pasirinkdami „*Nonmetric (modern)*“⁸¹.



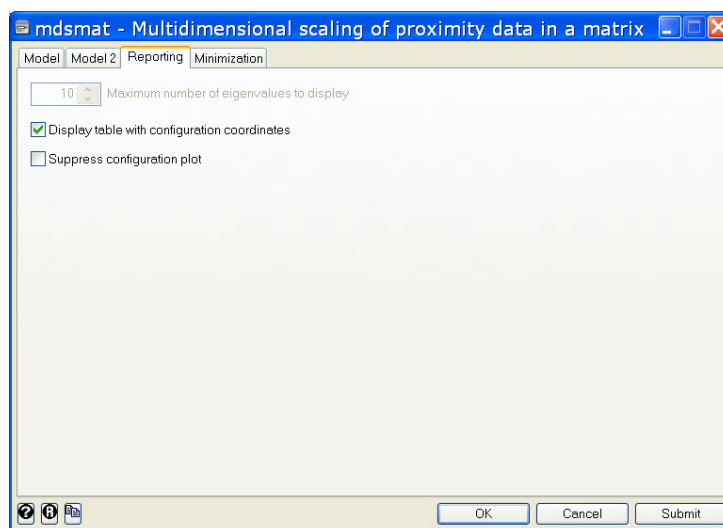
Kortelėje „*Model 2*“ nurodome, kad galutinis sprendinys turėtų būti dviejų dimensijų, ties „*Dimension of the approximating configuration*“ nurodydami 2. Pradžiai geriausia pasirinkti įprasčiausią – 2 dimensijų – sprendinį ir įvertinti jo kiekybines tinkamumo charakteristikas. Tik jei jos netenkina, patartina sudarinėti didesnio dimensijų skaičiaus sprendinius.

⁸¹ Čia svarbu pažymėti, kad Stata pakete nemetrinės daugiamačių skalių analizės optimizavimo funkcija (*loss function*) yra Kraskelo STRESS I

Sprendimas su Stata SE 10.0



Kortelėje „**Reporting**“ nurodome, kad būtų pateiktos objektų koordinatės mažamatėje erdvėje pažymėdami „**Display table with configuration coordinates**“.



Paskutinėje kortelėje „**Minimization**“ yra galimybė nurodyti pradinę objektų konfigūraciją. Kaip žinia, ji labai svarbus dalykas atliekamos iteracinės sprendinio optimizavimo procedūros prasme. Taigi čia būtina išbandyti įvairius variantus ir po to palyginti galutinius sprendinius. Jei jie esmingai nesiskiria, tai geriausias tinkamumo charakteristikas turintį sprendinį galima laikyti „galutiniu“. Stata programiniame pakete toks sprendinių tinkamumo charakteristikų palyginimas atliekamas automatiškai⁸² ties „**Starting configuration**“ pasirinkus „**Random configuration**“ ir įrašius atsitiktinį skaičių (mes pasirinkome 12344321) ties „**Seed**“ bei ties „**Protect optimizations**“ nurodžius skaičių, kuris reiškia, kelių atsitiktinių pradinių konfigūracijų galutinių sprendinių tinkamumo charakteristikos bus lyginamos (mes pasirinkome 10). Be to, patartina pažymėti varnelę ties „**Suppress the iteration log**“, kad nebūtų rodomi tarpiniai iteracijų rezultatai, kurie nėra labai prasmingi savaime. Atlikę visus šiuos veiksmus spaudžiame „**OK**“.

⁸² Dar yra galimybė papildomai palyginti pagal klasikinį MDS modelį sudarytų (ir atsitiktiniu būdu bei pagal klasikinį modelį sudarytų) pradinių konfigūracijų galutinių sprendinių rezultatus, tačiau čia to nedarysime paprastumo dėlei.

Sprendimas su Stata SE 10.0

Sintaksės komandos

```
mdsmat GNT_Euclidean, method(nonmetric) dimension(2) config
initialize(random(12344321)) protect(10) nolog
```

Gauname tokius rezultatus rezultatų išvesties lange:

(loss(stress) assumed)
(transform(monotonic) assumed)

run	mrc	#iter	lossval	seed random configuration
1	0	114	.02131761	X2ab77915d056f86b2e4fa5d3310e96f72255
2	0	121	.02131764	Xa748d985c43f462544a474abacbdd93d34a1
3	0	142	.022577	X4060f8cd8fb4929f88bffc459ed444bd242a
4	0	79	.02257702	Xb77376456501dabd7c2da3f83b1018de2091
5	0	78	.02257704	X01bf959dc6200fb2c6b0288371b83a2e44fb
6	0	87	.02257705	X806d2ab50787bb00950bfb2d2c0856350db3
7	0	119	.04948527	X652deb2deba11618b2169652cd28c9a61066
8	7	22	.05181836	Xaeeebffdd7e318c607efae98e2cb4a342e76
9	7	16	.1814126	X17ae5fe54edb76745bf2de77bb1379623027
10	7	15	.32380599	X532f7a5dfdcf65258728375fb64ff8ca2de7

Modern multidimensional scaling
dissimilarity matrix: GNT_Euclidean

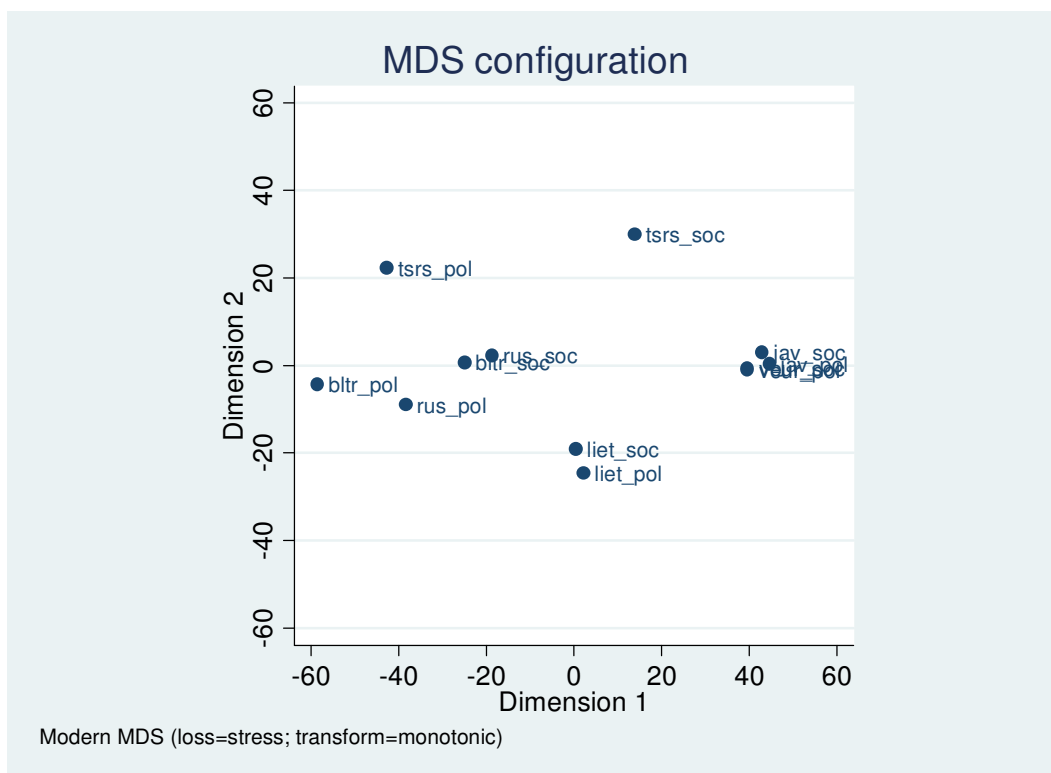
Loss criterion: stress = raw_stress/norm(distances)
Transformation: monotonic (nonmetric)

Normalization: principal

Number of obs	=	12
Dimensions	=	2
Loss criterion	=	0.0213

Configuration in 2-dimensional Euclidean space (principal normalization)

Category	dim1	dim2
tsrs_soc	13.8529	29.8336
tsrs_pol	-42.7728	22.3484
veur_soc	39.6266	-0.6969
veur_pol	39.6606	-1.0293
jav_soc	43.0338	2.9588
jav_pol	44.7166	0.3445
liet_soc	0.4543	-19.1176
liet_pol	2.1606	-24.5559
rus_soc	-18.6939	2.3434
rus_pol	-38.4480	-8.9014
bltr_soc	-24.9978	0.7424
bltr_pol	-58.5930	-4.2700



Išvados ir interpretacija

Pirmiausia žiūrime į skirtingų atsitiktinių būdu sudarytų pradinių konfigūracijų galutinių sprendinių tinkamumo charakteristikų palyginimą. Matome, kad paskutinių trijų sprendinių tinkamumo charakteristikos (Kraskelo STRESS I) yra gerokai blogesnės (atitinkamai 0,0518, 0,1814 ir 0,3238). Tačiau apskritai, dviejų dimensijų sprendinys yra labai gerai tinkamas duomenims struktūrizuoti (Kraskelo STRESS I = 0,0213)⁸³.

Toliau⁸⁴ svarbu atsižvelgti į optimizavimo procedūros metu atliktų artumų transformacijų kokybę: atitikimą tarp artumų, transformuotų artumų ir atstumų. Tam kaip minėta sudaromos Šepardo diagramos. Stata programiniame pakete standartiškai⁸⁵ galima sudaryti tik dvi iš trijų galimų Šepardo diagramų: tarp atstumų ir transformuotų artumų bei tarp atstumų ir artumų. Tiesa, transformuotų artumų ir artumų atitikimą Stata pakete galima patikrinti skaičiuojant Pirsono ir Spirmeno koreliacijas.

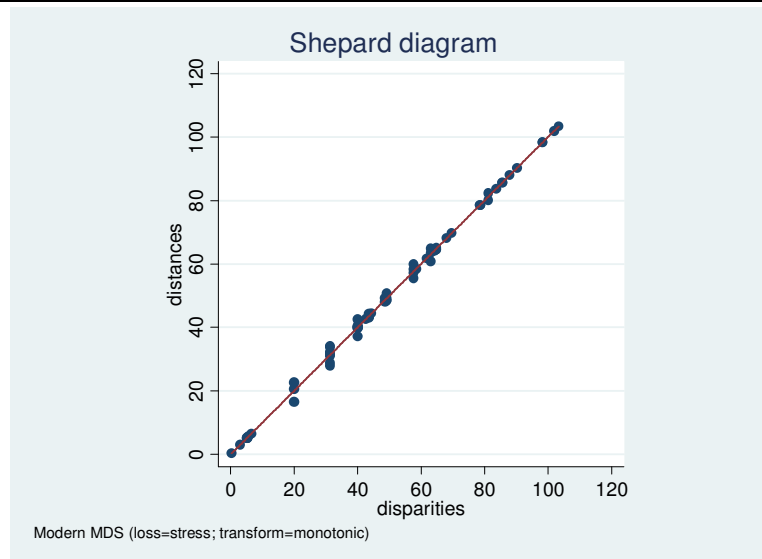
Meniu punktai (Šepardo diagrama tarp atstumų transformuotų artumų): Graphics → Multivariate analysis graphs → After multidimensional scaling → Shepard diagram;

Sintaksė: mdsshepard

⁸³ Aišku, tokia maža STRESS reikšmė gauta iš dalies ir dėl to, kad objektų kiekis labai nedidelis – tik 12.

⁸⁴ Žinoma, dar reiktų patyrinėti tinkamumo charakteristikų ir dimensijų skaičiaus brėžinį (*scree plot*), tačiau Stata pakete modernių MDS modelių atvejais jis nėra braižomas.

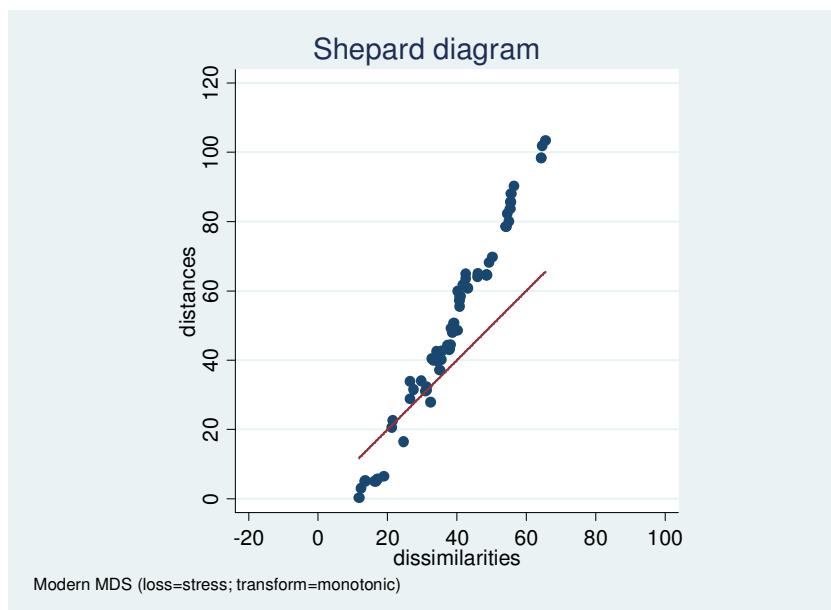
⁸⁵ Aišku, „rankomis“ galima nubraižyti visas ir netgi keletą papildomų, tačiau tai – ne šio studijų paketo apimties medžiaga.



Kaip matome, taškai brėžinyje išsidėstę labai netoli 45° kampu brėžinį kertančios tiesės, taigi sprendinys pagal šį brėžinį yra optimalus.

Meniu punktai (Šepardo diagrama tarp atstumų ir artumų): Graphics → Multivariate analysis graphs → After multidimensional scaling → Shepard diagram → atsidariusiame lange pažymime „*Suppress the linear transformation of the dissimilarities*“ → OK;

Sintaksė: `mdsshepard, notransform`



Kaip matome, taškai brėžinyje išsidėstę nuolat (monotoniškai) didėjančia tvarka, taigi sprendinys ir pagal šį brėžinį optimalus.

Meniu punktai (koreliacijos tarp transformuotų artumų ir artumų): tik sintaksės komanda;

Sintaksė: `estat correlations`

Sprendimas su Stata SE 10.0

Correlations of disparities and Euclidean distances

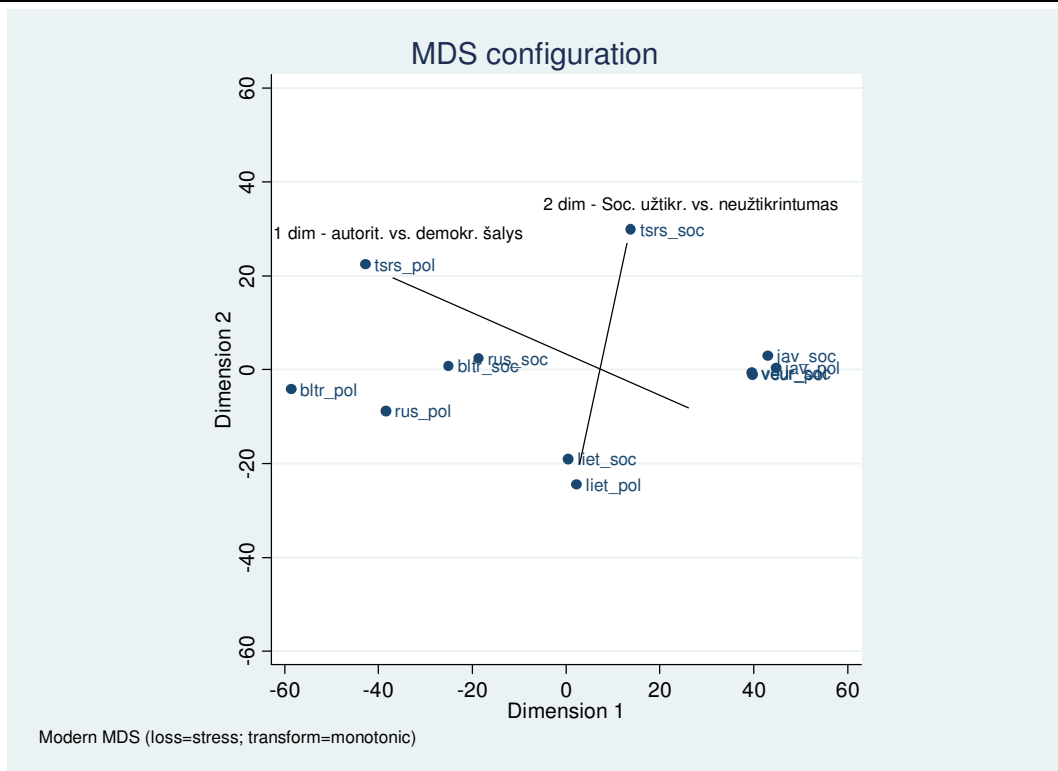
Category	N	Pearson	Spearman
tsrs_soc	11	0.9961	0.9723
tsrs_pol	11	0.9986	0.9723
veur_soc	11	1.0000	1.0000
veur_pol	11	1.0000	0.9886
jav_soc	11	0.9999	1.0000
jav_pol	11	0.9999	0.9818
liet_soc	11	0.9958	0.9932
liet_pol	11	0.9954	0.9795
rus_soc	11	0.9964	0.9886
rus_pol	11	0.9986	0.9909
bltr_soc	11	0.9965	0.9886
bltr_pol	11	0.9994	0.9977
Total	132	0.9990	0.9967

Kaip matome, visos koreliacijos artimos arba lygios vienetui, taigi sprendinys ir pagal šį kriterijų optimalus.

Bendra tinkamumo charakteristikų išvada – didesnio dimensijų skaičiaus sprendinio skaičiuoti neverta, nes dviejų dimensijų sprendinys optimalus.

Taigi jau galime žiūrėti į **dviejų dimensijų erdvinį žemėlapi** (*MDS configuration*) ir bandyti jį **interpretuoti**. Pirmiausiai, galime nagrinėti **objektų artimumą erdviniam žemėlapyje**. Matome, kad labai aiškų klasterį sudaro JAV ir Vakarų Europos politinės bei socialinės ir kultūrinės gerovės vertinimai, t. y. šių šalių santvarkų aspektai vertinami labai panašiai. Kitas klasteris yra Lietuvos santvarkos vertinimas, t. y. Lietuvos politinė ir socialinė bei kultūrinė gerovė vertinama panašiai. Kiti du klasteriai susiję su Rusijos ir Baltarusijos santvarkų vertinimų panašumais. Įdomu pastebėti, kad panašiai vertinama Rusijos ir Baltarusijos socialinė ir kultūrinė gerovė bei Rusijos ir Baltarusijos politinė gerovė, bet šie panašumai kažkiek nutolę vieni nuo kitų, t. y. lietuviai skirtingai vertina socialinę bei kultūrinę ir politinę gerovę šiose šalyse. Tuo tarpu TSRS politinės ir socialinės bei kultūrinės gerovės vertinimai galima sakyti mažai kuo panašūs. Kodėl taip yra iš dalies gali paaiškinti **sprendinio dimensijų įvardijimas**. Kadangi dimensijas galima **išbrėžti interpretuotojo nuožiūra**, tai pirmoji mūsų išbrėžta akivaizdžiai atspindi autoritarinių ir demokratinių šalių santvarkų vertinimų panašumo dimensiją. Viena jos gale yra JAV ir Vakarų Europos šalių politinių gerovės vertinimai, priešingame gale – TSRS politinės gerovės vertinimai. Antroji mūsų išbrėžta dimensija atspindi socialinio užtikrintumo ir gerovės, o tam tikra prasme ir praeities socialinio saugumo nostalgijos dimensiją. Čia dabartinė Lietuvos socialinė ir kultūrinė gerovė priešpastatyta (vertinimų prasme) buvusiai socialinei ir kultūrinei gerovei TSRS.

Sprendimas su Stata SE 10.0



10.4 lentelė. Modernioji nemetrinė vienos simetrinės artumų matricos daugiamačių skalių analizė (MDS) su SPSS: ALSCAL algoritmas.

Sprendimas su SPSS 15.0

Parengiamieji veiksmai

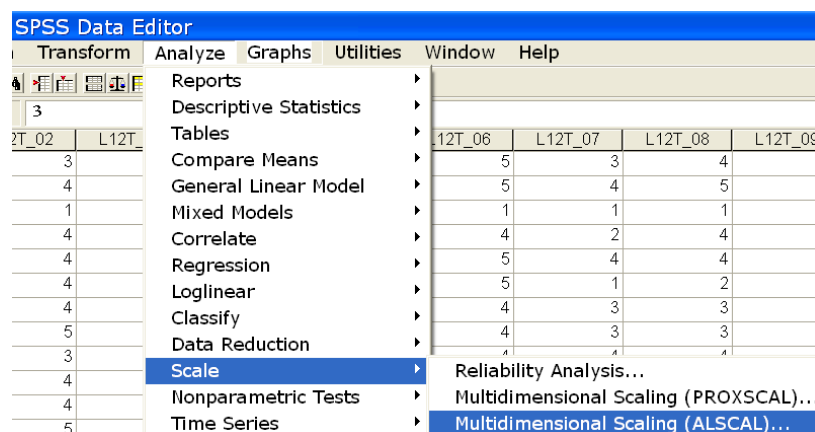
Nėra.

Sintaksės komandos

Nėra.

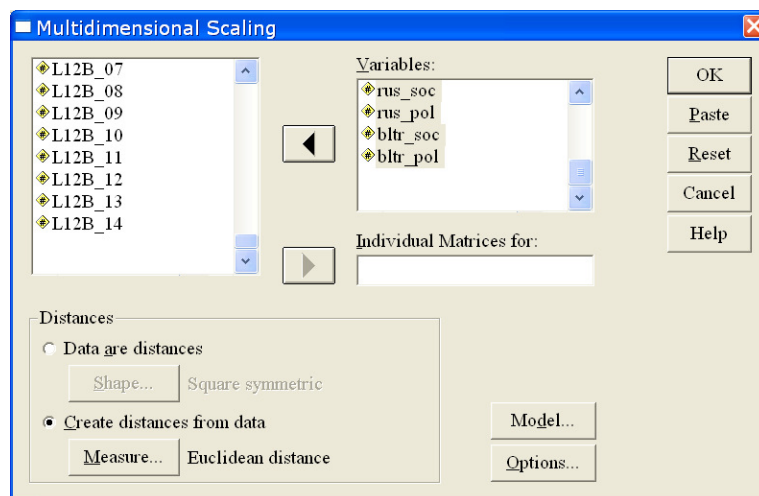
Analizės veiksmai

Meniu punktai: Analyze → Scale → Multidimensional scaling (ALSCAL)

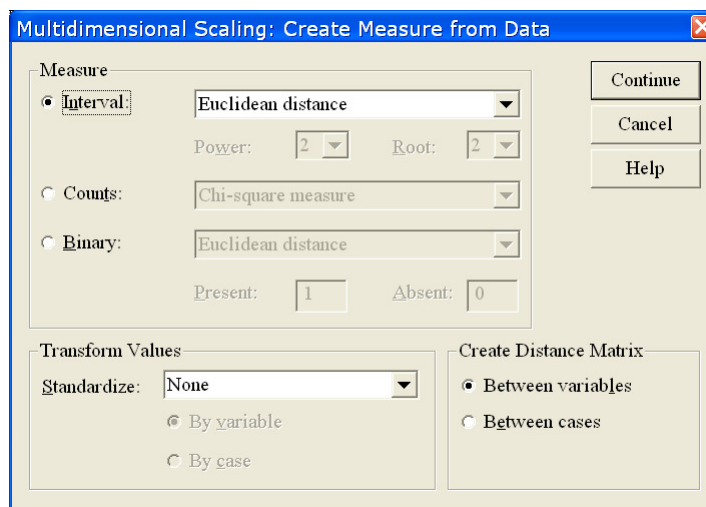


Sprendimas su SPSS 15.0

Atsidariusiame lange „*Multidimensional Scaling*“ į dešinėje pusėje esantį langelį „*Variables*“ įkeliam kintamuosius **tsrs_soc – bltr_pol**, ir nurodome, kad artumų matrica būtų sudaroma iš turimų duomenų, ties „*Distances*“ pažymėdami „*Create distances from data*“. Tada spaudžiame „*Measure...*“ mygtuką.



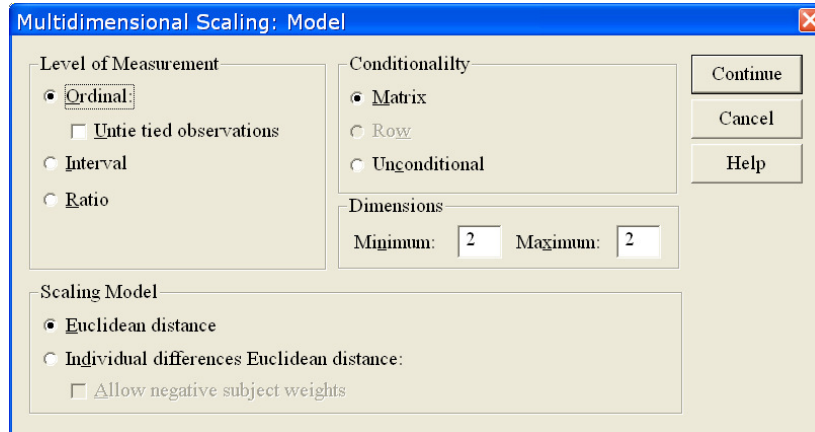
Atsidariusiame lange „*Measure*“ skiltyje nurodome skaičiuoti Euklidinį artumo matą, ties „*Interval*“ išrinkdami „*Euclidean distance*“. Be to, nurodome artumus skaičiuoti tarp kintamųjų, o ne atvejų, ties „*Create Distance Matrix*“ parinkdami „*Between variables*“ ir spaudžiame „*Continue*“.



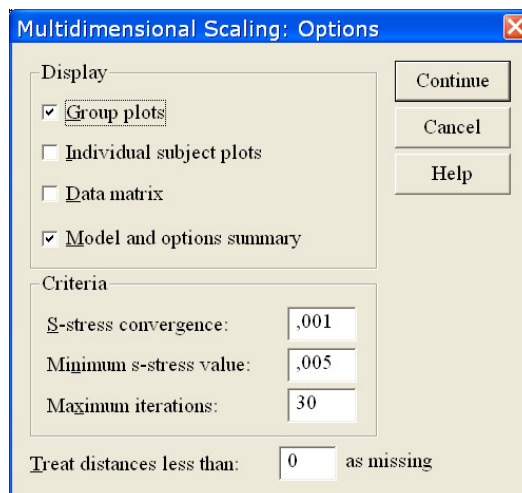
Grįžę į „*Multidimensional Scaling*“ langą spaudžiame „*Model...*“ ir atsidariusiame lange parenkame MDS modelio parametrus. Nurodome, kad būtų skaičiuojamas nemetrinis MDS modelis, ties „*Level of Measurement*“ pažymėdami „*Ordinal*“⁸⁶. Po to, nurodome, kad būtų atliekama vienos artumų matricos analizė, ties „*Scaling Model*“ parinkdami „*Euclidean distance*“. Galiausiai, parenkame 2 dimensijų galutinį sprendinį ties „*Dimensions*“ į abu laukelius įrašydami 2, ir spaudžiame „*Continue*“.

⁸⁶ Čia svarbu pažymėti, kad ALSCAL algoritme optimizavimo funkcija yra Jango S-STRESS I.

Sprendimas su SPSS 15.0



Grįžę į „*Multidimensional Scaling*“ langą spaudžiame „*Options...*“ ir atsidariusiame lange parenkame rezultatų išvesties bei optimizavimo procedūros parametrus. „*Display*“ skiltyje nurodome, kad būtų nubraižytas erdvinis brėžinys bei pateikta įvairi papildoma informacija apie sprendinį, atitinkamai pažymėdami „*Group plots*“ ir „*Model and options summary*“. Po to, spaudžiame „*Continue*“, o grįžę į „*Multidimensional Scaling*“ langą – „*OK*“.



Sintaksės komandos

```
PROXIMITIES tsrs_soc tsrs_pol veur_soc veur_pol jav_soc jav_pol liet_soc liet_pol
rus_soc rus_pol bltr_soc bltr_pol /PRINT NONE /MATRIX OUT('C:\spssalsc.tmp')
/MEASURE=EUCLID /STANDARDIZE=NONE /VIEW=VARIABLE .
ALSCAL
/MATRIX= IN('C:\spssalsc.tmp')
/LEVEL=ORDINAL
/CONDITION=MATRIX
/MODEL=EUCLID
/CRITERIA=CONVERGE(.001) STRESSMIN(.005) ITER(30) CUTOFF(0) DIMENS(2,2)
/PLOT=DEFAULT
/PRINT=HEADER .
ERASE FILE='C:\spssalsc.tmp'.
```

Gauname tokius rezultatus **rezultatų išvesties lange:**

Iteration history for the 2 dimensional solution (in squared distances)

Young's S-stress formula 1 is used.

Sprendimas su SPSS 15.0

Iteration	S-stress	Improvement
1	,02431	
2	,02112	,00319
3	,02032	,00080

Iterations stopped because
S-stress improvement is less than ,001000

Stress and squared correlation (RSQ) in distances

RSQ values are the proportion of variance of the scaled data (disparities)
in the partition (row, matrix, or entire data) which
is accounted for by their corresponding distances.
Stress values are Kruskal's stress formula 1.

For matrix
Stress = ,02832 RSQ = ,99649

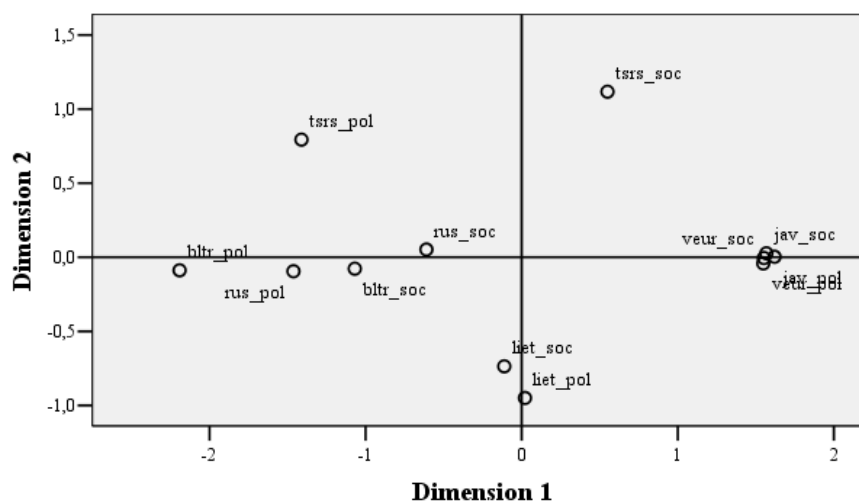
Configuration derived in 2 dimensions

Stimulus Coordinates

Stimulus Number	Stimulus Name	Dimension	
		1	2
1	tsrs_soc	,5497	1,1173
2	tsrs_pol	-1,4113	,7941
3	veur_soc	1,5518	-,0052
4	veur_pol	1,5476	-,0420
5	jav_soc	1,5673	,0271
6	jav_pol	1,6197	,0029
7	liet_soc	-,1116	-,7361
8	liet_pol	,0211	-,9500
9	rus_soc	-,6106	,0521
10	rus_pol	-1,4626	-,0948
11	bltr_soc	-1,0695	-,0773
12	bltr_pol	-2,1917	-,0881

Derived Stimulus Configuration

Euclidean distance model

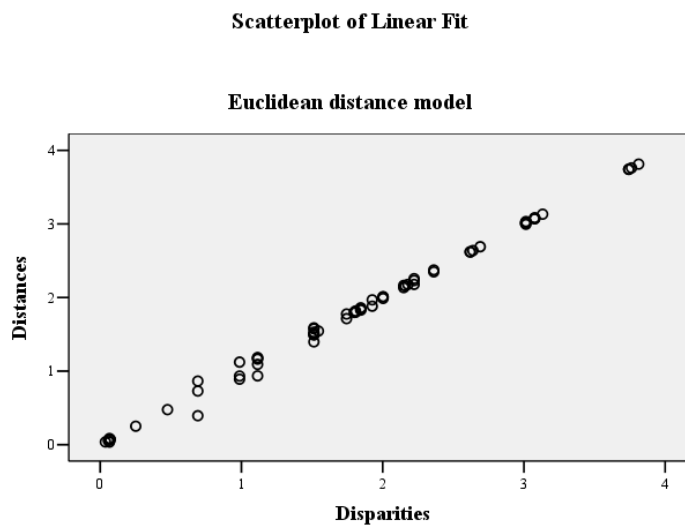


Išvados ir interpretacija

Pirmiausia žiūrime į sprendinio tinkamumo charakteristikas. Matome, kad dviejų dimensijų sprendinys yra labai gerai tinkamas duomenims struktūrizuoti (Jango S-STRESS I = 0,02032, Kraskelo STRESS I = 0,02832, $R^2 = 0,99649$)⁸⁷.

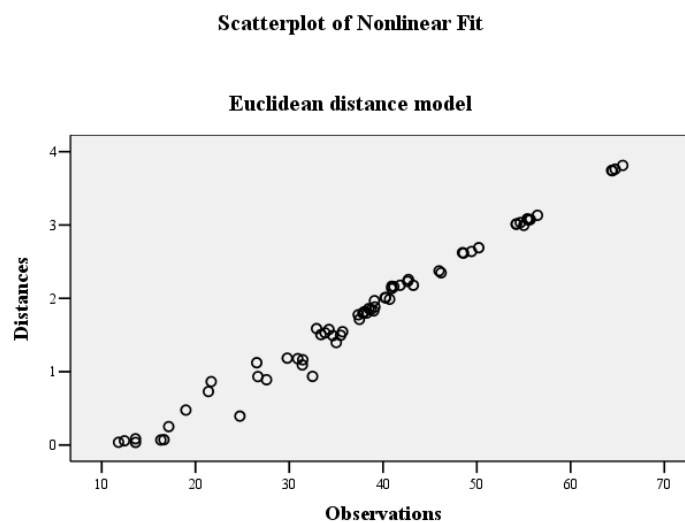
Toliau⁸⁸ nagrinėjame optimizavimo procedūros metu atliktų artumų transformacijų kokybę: atitikimą tarp artumų, transformuotų artumų ir atstumų. Tam, kaip minėta, sudaromos Šepardo diagramos. ALSCAL procedūros metu jos sudaromos automatiškai.

Šepardo diagrama tarp atstumų ir transformuotų artumų:



Kaip matome, taškai brėžinyje išsidėstę labai netoli 45° kampu brėžinį kertančios tiesės, taigi sprendinys pagal šį brėžinį optimalus.

Šepardo diagrama tarp atstumų ir artumų:



Kaip matome, taškai brėžinyje išsidėstę nuolat (monotoniškai) didėjančia tvarka, taigi sprendinys ir

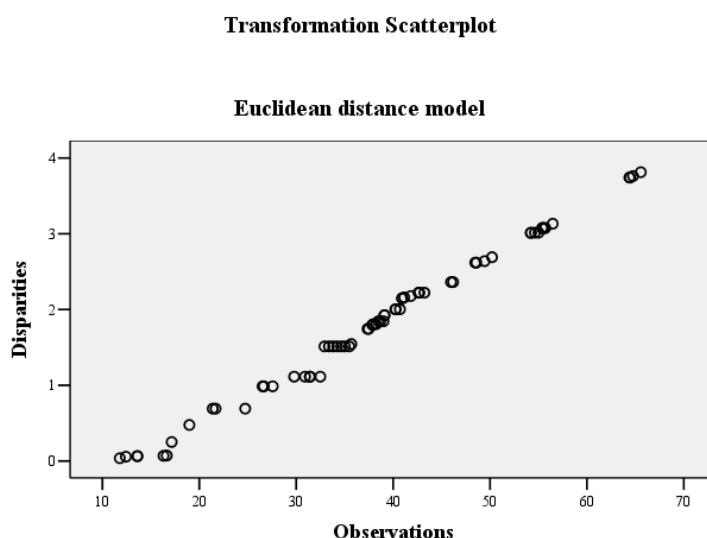
⁸⁷ Primename, kad tokios mažos STRESS reikšmės gautos iš dalies ir dėl to, kad objektų kiekis labai nedidelis – tik 12.

⁸⁸ Žinoma, dar reiktų patyrinėti tinkamumo charakteristikų ir dimensijų skaičiaus brėžinį (*scree plot*), tačiau pasirinkus ALSCAL algoritmą, jis nėra braižomas.

Sprendimas su SPSS 15.0

pagal šį brėžinį optimalus.

Šepardo diagrama tarp transformuotų artumų ir artumų:



Kaip matome, horizontalių pakopų, t. y. taškų vienodame lygyje, kiekis yra labai nedidelis, taigi sprendinys ir pagal šį brėžinį optimalus.

Bendra tinkamumo charakteristikų išvada – didesnio dimensijų skaičiaus sprendinio skaičiuoti neverta, nes dviejų dimensijų sprendinys optimalus.

Taigi dabar jau galime žiūrėti į **dviejų dimensijų erdvinį žemėlapi** (*Derived stimulus configuration*) ir bandyti jį **interpretuoti**. Tiesa, sprendinys iš esmės nesiskiria nuo to, kuris buvo gautas su Stata paketu, todėl interpretacijos nebekartosime. Pažymėsime tik tai, kad, viena vertus, kadangi sprendiniai beveik nesiskiria, tai galime būti dar labiau užtikrinti savo ankstesnių išvadų pagrįstumu. Kita vertus, parodėme, kad ALSCAL algoritmas „nėra jau toks blogas“: jis iš karto konvergavo link globalaus minimumo, o ne lokalaus. Tiesa, SPSS vartotojams nerekomenduojame jo naudoti, nes yra geresnė ir stabilesnė alternatyva – PROXSCAL algoritmas. Jo aptarimas kaip tik patiktas sekančioje lentelėje.

10.5 lentelė. Modernioji nemetrinė vienos simetrinės artumų matricos daugiamačių skalių analizė (MDS) su SPSS: PROXSCAL algoritmas.

Sprendimas su SPSS 15.0

Parengiamieji veiksmai

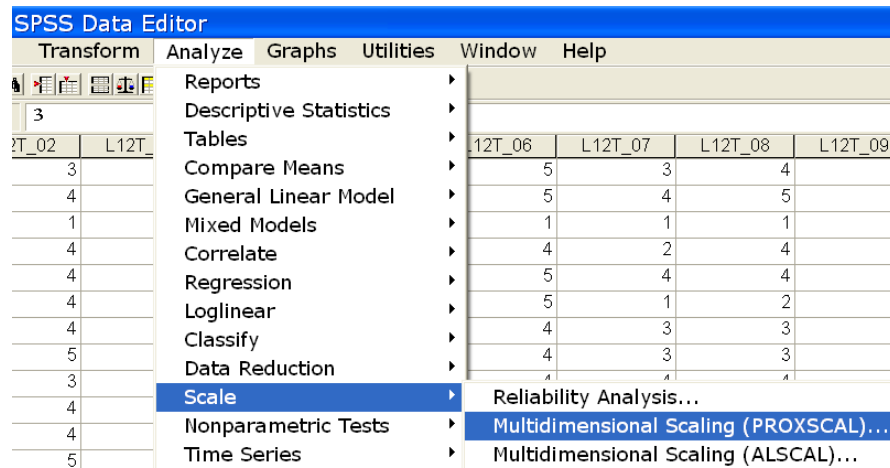
Nėra.

Sintaksės komandos

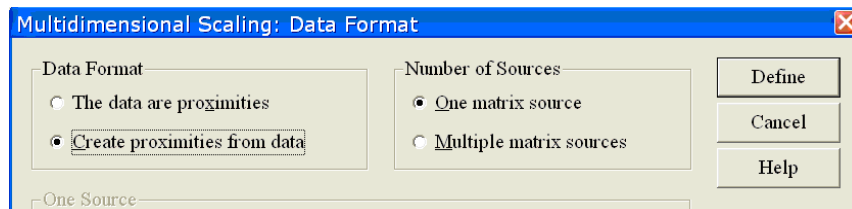
Nėra.

Analizės veiksmai

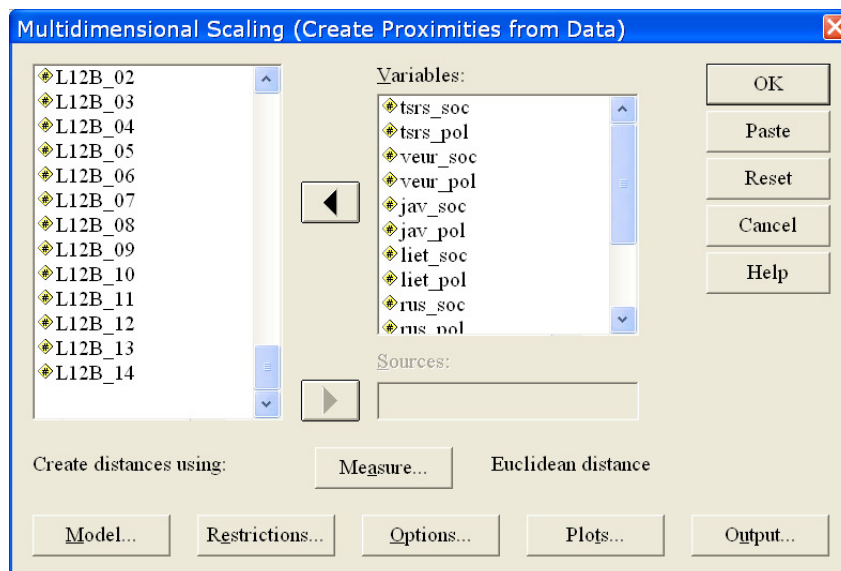
Meniu punktai: Analyze → Scale → Multidimensional scaling (PROXSCAL)



Atsidariusiame lange „**Multidimensional Scaling: Data Format**“ nurodome, kad artumų matrica būtų sudaroma iš turimų duomenų, ties „**Data Format**“ pažymėdami „**Create proximities from data**“, ir kad būtų atliekama vienos artumų matricos analizė, ties „**Number of Sources**“ parinkdami „**One matrix source**“. Tada spaudžiame „**Define**“.



Atsidariusiame lange „**Multidimensional Scaling (Create Proximities form Data)**“ į dešinėje pusėje esantį langelį „**Variables**“ įkeliam kintamuosius **tsrs_soc – bltr_pol**. Tada spaudžiame „**Measure...**“ mygtuką.



Atsidariusiame lange skiltyje „**Measure**“ nurodome skaičiuoti Euklidinį artumo matą, ties „**Interval**“ išrinkdami „**Euclidean distance**“. Be to, nurodome artumus skaičiuoti tarp kintamųjų, o ne atvejų, ties „**Create Distance Matrix**“ parinkdami „**Between variables**“ ir spaudžiame „**Continue**“.

Grįžę į „*Multidimensional Scaling (Create Proximities form Data)*“ langą spaudžiame „*Model...*“ ir atsidariusiame lange parenkame MDS modelio parametrus. Nurodome, kad būtų skaičiuojamas nemetrinis MDS modelis, skiltyje „*Proximity Transformations*“ pažymėdami „*Ordinal*“⁸⁹ bei parenkame 2 dimensijų galutinį sprendinį ties „*Dimensions*“ į abu laukelius įrašydami 2. Po to, spaudžiame „*Continue*“.

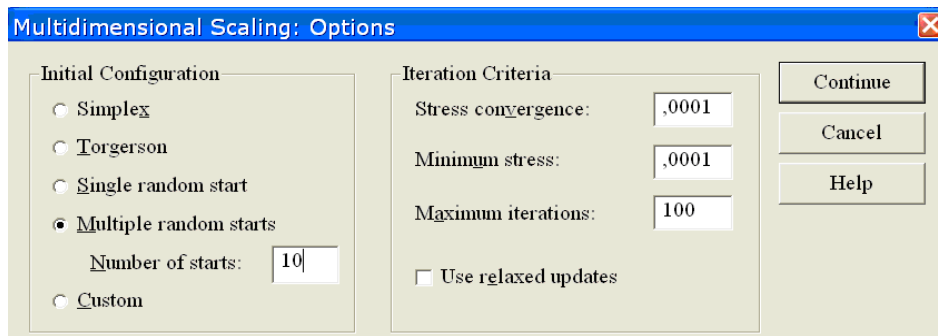
Grįžę į „*Multidimensional Scaling (Create Proximities form Data)*“ spaudžiame „*Options...*“ ir nurodome, kaip bus sudaroma pradinė objektų konfigūracija. SPSS PROXSCAL algoritme yra galimybė pradinę konfigūraciją sudaryti atsitiktiniu būdu, pagal klasikinį modelį ir simplekso būdu. Tačiau, kaip žinia, geriausių rezultatų galima pasiekti tik sudarius daug sprendinių ir palyginus jų kiekybines tinkamumo charakteristikas. Toks sprendinių tinkamumo charakteristikų palyginimas atliekamas automatiškai⁹⁰, skiltyje „*Initial Configuration*“ pasirinkus „*Multiple random starts*“ ir ties „*Number of starts*“ nurodžius skaičių, kuris reiškia, kelių atsitiktinių pradinių konfigūracijų galutinių sprendinių tinkamumo charakteristikos bus lyginamos (mes pasirinkome 10). Atlikę visus šiuos veiksmus

⁸⁹ Čia svarbu pažymėti, kad PROXSCAL algoritme optimizavimo funkcija yra normuotasis STRESS.

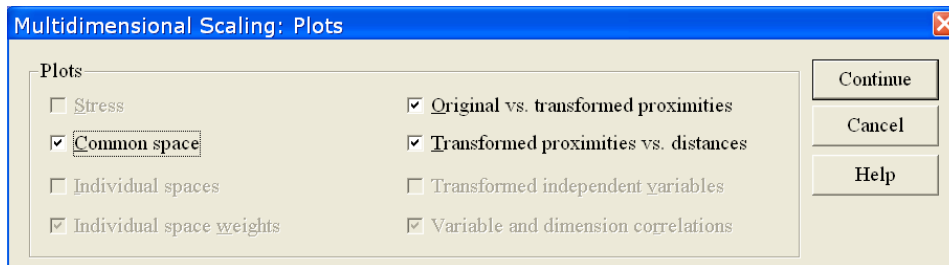
⁹⁰ Dar yra galimybė papildomai palyginti atsitiktiniu būdu bei pagal klasikinį modelį ir simplekso būdu sudarytų pradinių konfigūracijų galutinių sprendinių rezultatus, tačiau čia to nedarysime paprastumo dėlei.

Sprendimas su SPSS 15.0

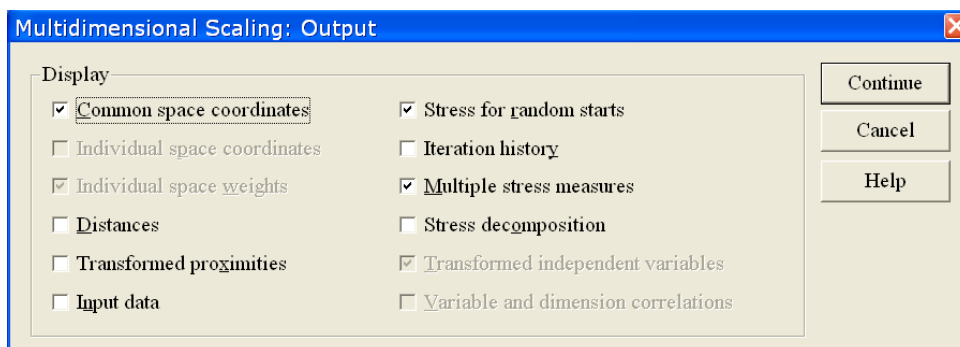
spaudžiame „Continue“.



Grįžę į „*Multidimensional Scaling (Create Proximities form Data)*“ langą spaudžiame „Plots...“ ir atsidariusiame lange nurodome, kokie grafikai turi būti braizomi: erdvinis žemėlapis (pažymime „*Common space*“), artumų ir transformuotų artumų Šepardo diagrama (pažymime „*Original vs. transformed proximities*“) bei transformuotų artumų ir atstumų Šepardo diagrama (pažymime „*Transformed proximities vs. distances*“). Atlikę visus šiuos veiksmus spaudžiame „Continue“.



Grįžę į „*Multidimensional Scaling (Create Proximities form Data)*“ langą spaudžiame „Output...“ ir atsidariusiame lange parenkame rezultatų išvesties turinį. „Display“ skiltyje nurodome, kad būtų pateiktos objektų koordinatės mažamateje erdvėje pažymėdami „*Common space coordinates*“. Taip pat nurodome pateikti kiekybines galutinio sprendinio tinkamumo charakteristikas pažymėdami „*Multiple stress measures*“ bei skirtingų atsitiktinių pradinių konfigūracijų galutinių sprendinių tinkamumo charakteristikas pažymėdami „*Stress for random starts*“. Po to, spaudžiame „Continue“, o grįžę į „*Multidimensional Scaling (Create Proximities form Data)*“ langą – „OK“.



Sintaksės komandos

```
PROXIMITIES tsrs_soc tsrs_pol veur_soc veur_pol jav_soc jav_pol liet_soc liet_pol
rus_soc rus_pol bltr_soc bltr_pol /PRINT NONE /MATRIX OUT( 'C:\spssprxs.tmp' )
/MEASURE=EUCLID /STANDARDIZE=NONE /VIEW=VARIABLE .
PROXSCAL
/MATRIX=IN( 'C:\spssprxs.tmp' )
/INITIAL=RANDOM(10)
/TRANSFORMATION=ORDINAL (KEEPTIES)
```

Sprendimas su SPSS 15.0

```

/ACCELERATION=NONE
/CRITERIA=DIMENSIONS(2,2) MAXITER(100) DIFFSTRESS(.0001) MINSTRESS(.0001)
/PRINT=COMMON RANDOM STRESS
/PLOT=COMMON TRANSFORMATIONS ( ALL ) RESIDUALS ( ALL ) .
ERASE FILE='C:\spssprxs.tmp' .

```

Gauname tokius rezultatus **rezultatų išvesties lange**:

Multiple Random Starts

Random Start	Seed	Normalized Raw Stress
1	1E+009	,0013
2	189061494	,0012
3	163113147	,0008 ^a
4	562548763	,0008
5	15053862	,0011
6	2E+009	,0095
7	2E+009	,0033
8	389105623	,1116
9	423042446	,0033
10	2E+009	,0015

a. Lowest Stress was found at random start 3 with seed 163113147.

Stress and Fit Measures

Normalized Raw Stress	,00077
Stress-I	,02782 ^a
Stress-II	,05880 ^a
S-Stress	,00080 ^b
Dispersion Accounted For (D.A.F.)	,99923
Tucker's Coefficient of Congruence	,99961

PROXSCAL minimizes Normalized Raw Stress.

a. Optimal scaling factor = 1,001.

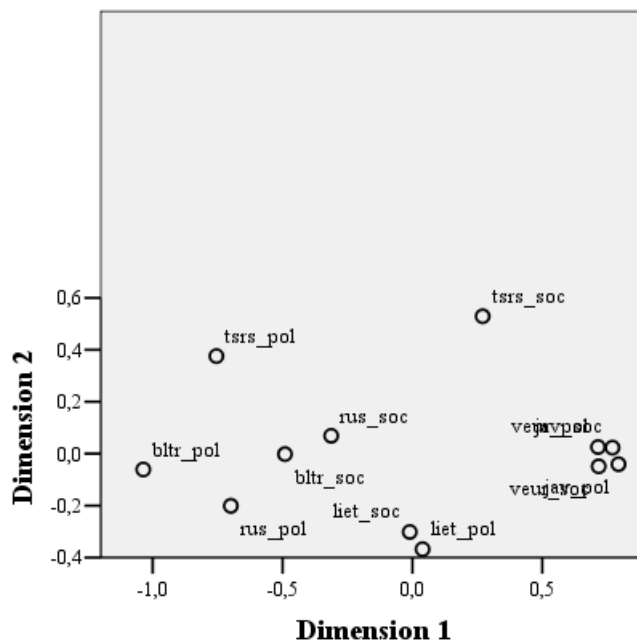
b. Optimal scaling factor = 1,000.

Final Coordinates

	Dimension	
	1	2
Socialines ir kulturines geroves vertinimas: TSRS	,270	,529
Politines geroves vertinimas: TSRS	-,755	,376
Socialines ir kulturines geroves vertinimas: Vakarų Europa	,717	-,049
Politines geroves vertinimas: Vakarų Europa	,715	,026
Socialines ir kulturines geroves vertinimas: JAV	,770	,023
Politines geroves vertinimas: JAV	,793	-,041
Socialines ir kulturines geroves vertinimas: Lietuva	-,010	-,301
Politines geroves vertinimas: Lietuva	,039	-,368
Socialines ir kulturines geroves vertinimas: Rusija	-,313	,069
Politines geroves vertinimas: Rusija	-,699	-,201
Socialines ir kulturines geroves vertinimas: Baltarusija	-,491	-,002
Politines geroves vertinimas: Baltarusija	-1,036	-,061

Object Points

Common Space



Išvados ir interpretacija

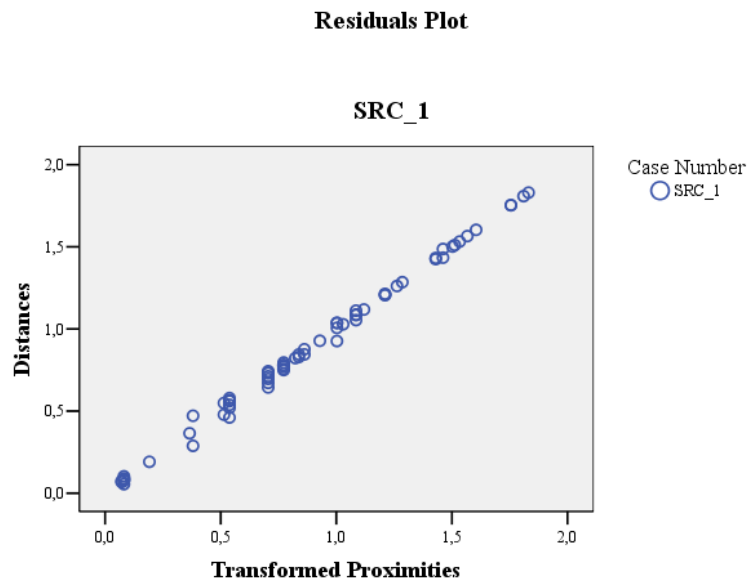
Pirmiausia žiūrime į skirtingų atsitiktinių būdu sudarytų pradinių konfigūracijų galutinių sprendinių tinkamumo charakteristikų palyginimą. Matome, kad kelių sprendinių tinkamumo charakteristikos (normuotojo STRESS) yra gerokai blogesnės (tarkim, 8 sprendinio normuotasis STRESS = 0,1116).

Sprendimas su SPSS 15.0

Tačiau apskritai, dviejų dimensijų sprendinys yra labai gerai tinkamas duomenims struktūrizuoti (pavyzdžiui, nuormuotas STRESS = 0,00077, Kraskelo STRESS I = 0,02782, Jango S-STRESS I = 0,00080, D.A.F. = 0,99923)⁹¹.

Toliau⁹² reikia išnagrinėti optimizavimo procedūros metu atliktų artumų transformacijų kokybę: atitikimą tarp artumų, transformuotų artumų ir atstumų. Tam, kaip minėta, sudaromos Šepardo diagramos. PROXSCAL procedūros metu jos sudaromos automatiškai, tačiau standartiškai⁹³ – tik dvi iš trijų galimų: tarp atstumų transformuotų artumų bei tarp transformuotų artumų ir artumų.

Šepardo diagrama tarp atstumų ir transformuotų artumų:



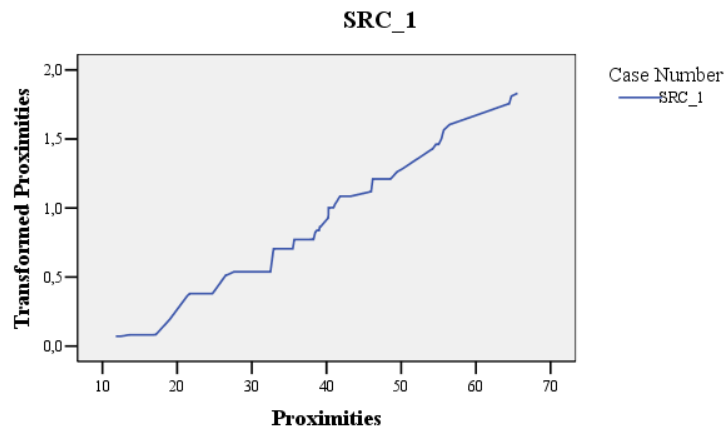
Kaip matome, taškai brėžinyje išsidėstę labai netoli 45° kampu brėžinį kertančios tiesės, taigi sprendinys pagal šį brėžinį optimalus.

Šepardo diagrama tarp transformuotų artumų ir artumų:

⁹¹ Vėlgi primename, kad tokios mažos STRESS reikšmės gautos iš dalies ir dėl to, kad objektų kiekis labai nedidelis – tik 12.

⁹² Žinoma, dar reikėtų patyrinti tinkamumo charakteristikų ir dimensijų skaičiaus brėžinį (*scree plot*), tačiau pasirinkus ALSCAL algoritimą, jis nėra braižomas.

⁹³ Aišku, „rankomis“ galima nubraižyti visas ir netgi keletą papildomų, tačiau tai – ne šio studijų paketo apimties medžiaga.

Transformation Plot

Kaip matome, horizontalių pakopų, t. y. taškų vienodame lygyje, kiekis yra labai nedidelis, taigi sprendinys ir pagal šį brėžinį optimalus.

Bendra tinkamumo charakteristikų išvada – didesnio dimensijų skaičiaus sprendinio skaičiuoti neverta, nes dviejų dimensijų sprendinys optimalus.

Taigi dabar jau galime žiūrėti į **dviejų dimensijų erdvinį žemėlapi** (*Object Points, Common Space*) ir bandyti jį **interpretuoti**. Tiesa, sprendinys iš esmės nesiskiria nuo to, kuris buvo gautas su Stata paketu ir ALSCAL algoritmu, todėl interpretacijos nebekartosime. Vėl pažymėsime tai, kad kadangi sprendiniai beveik nesiskiria, tai galime būti dar labiau užtikrinti savo ankstesnių išvadų pagrįstumu.

Taigi atlikę Lietuvos gyventojų požiūrių į du skirtingų šalių santvarkų aspektus – politinę ir socialinę bei kultūrinę gerovę – nemetrinę daugiamačių skalių analizę **išsiaiškinome, kad egzistuoja šalių vertinimo pagal tuos du aspektus latentinė struktūra (klasteriai)**. Taip pat **nustatėme, kad lietuviai šalis pateiktas vertino pagal dvi latentines dimensijas**: autoritarizmo ir demokratijos bei socialinio saugumo (gerovės) ir neužtikrintumo.

11. LITERATŪRA

Bartholomew, David J., Fiona Steele, Irini Moustaki ir Jane I. Galbraith. 2002. *The Analysis and Interpretation of Multivariate Data for Social Scientists*. Boca Raton, London, New York, Washington, D.C.: Chapman & Hall/CRC Press.

Čekanavičius, Vydas ir Gediminas Murauskas. 2000. *Statistika ir jos taikymai. I*. Vilnius: TEV.

Kruskal, Joseph B. ir Myron Wish. 1978. *Multidimensional Scaling*. Beverly Hills, CA: Sage Publications.

Vaitkevičius, Raimundas ir Aušra Saudargienė. 2006. *Statistika su SPSS psichologiniuose tyrimuose*. Kaunas: Vytauto Didžiojo universiteto leidykla.